

Applications of Machine Learning Methods in Complex Economics and Financial Networks

Lead Guest Editor: Benjamin M. Tabak

Guest Editors: Thiago C. Silva, Liang Zhao, and Ahmet Sensoy





Applications of Machine Learning Methods in Complex Economics and Financial Networks

Complexity

Applications of Machine Learning Methods in Complex Economics and Financial Networks

Lead Guest Editor: Benjamin M. Tabak

Guest Editors: Thiago C. Silva, Liang Zhao, and
Ahmet Sensoy



Copyright © 2020 Hindawi Limited. All rights reserved.

This is a special issue published in "Complexity." All articles are open access articles distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Chief Editor

Hiroki Sayama, USA

Editorial Board

Oveis Abedinia, Kazakhstan
José Ángel Acosta, Spain
Carlos Aguilar-Ibanez, Mexico
Mojtaba Ahmadiéh Khanesar, United Kingdom
Tarek Ahmed-Ali, France
Alex Alexandridis, Greece
Basil M. Al-Hadithi, Spain
Juan A. Almendral, Spain
Diego R. Amancio, Brazil
David Arroyo, Spain
Mohamed Boutayeb, France
Átila Bueno, Brazil
Arturo Buscarino, Italy
Ning Cai, China
Eric Campos, Mexico
Émile J. L. Chappin, The Netherlands
Yu-Wang Chen, United Kingdom
Diyi Chen, China
Giulio Cimini, Italy
Danilo Comminiello, Italy
Sergey Dashkovskiy, Germany
Manlio De Domenico, Italy
Pietro De Lellis, Italy
Albert Diaz-Guilera, Spain
Thach Ngoc Dinh, France
Jordi Duch, Spain
Marcio Eisenkraft, Brazil
Joshua Epstein, USA
Mondher Farza, France
Thierry Floquet, France
José Manuel Galán, Spain
Lucia Valentina Gambuzza, Italy
Harish Garg, India
Bernhard C. Geiger, Austria
Carlos Gershenson, Mexico
Peter Giesl, United Kingdom
Sergio Gómez, Spain
Lingzhong Guo, United Kingdom
Xianggui Guo, China
Sigurdur F. Hafstein, Iceland
Chittaranjan Hens, India
Giacomo Innocenti, Italy
Sarangapani Jagannathan, USA



Mahdi Jalili, Australia
Peng Ji, China
Jeffrey H. Johnson, United Kingdom
Mohammad Hassan Khooban, Denmark
Abbas Khosravi, Australia
Toshikazu Kuniya, Japan
Vincent Labatut, France
Lucas Lacasa, United Kingdom
Guang Li, United Kingdom
Qingdu Li, China
Chongyang Liu, China
Xinzhi Liu, Canada
Xiaoping Liu, Canada
Rosa M. Lopez Gutierrez, Mexico
Vittorio Loreto, Italy
Noureddine Manamanni, France
Didier Maquin, France
Eulalia Martínez, Spain
Marcelo Messias, Brazil
Ana Meštrović, Croatia
Ludovico Minati, Japan
Saleh Mobayen, Iran
Christopher P. Monterola, Philippines
Marcin Mrugalski, Poland
Roberto Natella, Italy
Sing Kiong Nguang, New Zealand
Nam-Phong Nguyen, USA
Irene Otero-Muras, Spain
Yongping Pan, Singapore
Daniela Paolotti, Italy
Cornelio Posadas-Castillo, Mexico
Mahardhika Pratama, Singapore
Luis M. Rocha, USA
Miguel Romance, Spain
Avimanyu Sahoo, USA
Matilde Santos, Spain
Ramaswamy Savitha, Singapore
Michele Scarpiniti, Italy
Enzo Pasquale Scilingo, Italy
Dan Selișteanu, Romania
Dehua Shen, China
Dimitrios Stamovlasis, Greece
Samuel Stanton, USA
Roberto Tonelli, Italy



Shahadat Uddin, Australia
Gaetano Valenza, Italy
Jose C. Valverde, Spain
Alejandro F. Villaverde, Spain
Dimitri Volchenkov, USA
Christos Volos, Greece
Zidong Wang, United Kingdom
Qingling Wang, China
Wenqin Wang, China
Yan-Ling Wei, Singapore
Honglei Xu, Australia
Yong Xu, China
Xinggang Yan, United Kingdom
Zhile Yang, China
Baris Yuce, United Kingdom
Massimiliano Zanin, Spain
Hassan Zargarzadeh, USA
Rongqing Zhang, China
Xianming Zhang, Australia
Xiaopeng Zhao, USA
Quanmin Zhu, United Kingdom

Contents

Applications of Machine Learning Methods in Complex Economics and Financial Networks

Benjamin M. Tabak , Thiago C. Silva , Liang Zhao, and Ahmet Sensoy


Editorial (2 pages), Article ID 4247587, Volume 2020 (2020)

Portfolio Optimization with Asset-Liability Ratio Regulation Constraints

De-Lei Sheng  and Peilong Shen



Research Article (13 pages), Article ID 1435356, Volume 2020 (2020)

Measure User Intimacy by Mining Maximum Information Transmission Paths

Lin Guo  and Dongliang Zhang

Research Article (9 pages), Article ID 2376451, Volume 2020 (2020)

Modeling Repayment Behavior of Consumer Loan in Portfolio across Business Cycle: A Triplet Markov Model Approach

Shou Chen  and Xiangqian Jiang 



Research Article (11 pages), Article ID 5458941, Volume 2020 (2020)


Modeling Investor Behavior Using Machine Learning: Mean-Reversion and Momentum Trading Strategies

Thiago Christiano Silva , Benjamin Miranda Tabak , and Idamar Magalhães Ferreira

Research Article (14 pages), Article ID 4325125, Volume 2019 (2019)





Public Procurement Announcements in Spain: Regulations, Data Analysis, and Award Price Estimator Using Machine Learning

Manuel J. García Rodríguez , Vicente Rodríguez Montequín , Francisco Ortega Fernández, and

Joaquín M. Villanueva Balsera 

Research Article (20 pages), Article ID 2360610, Volume 2019 (2019)

Chinese Currency Exchange Rates Forecasting with EMD-Based Neural Network

Jying-Nan Wang , Jiangze Du , Chonghui Jiang , and Kin-Keung Lai 

Research Article (15 pages), Article ID 7458961, Volume 2019 (2019)

A Study of RMB Internationalization Path Based on Border Area Perspective

Po Sheng Ko, Cheng Chung Wu , Ying Shih Mai, and Zhongrong Xu




Research Article (10 pages), Article ID 2834894, Volume 2019 (2019)

A Hybrid Approach Using Oversampling Technique and Cost-Sensitive Learning for Bankruptcy Prediction

Tuong Le , Minh Thanh Vo, Bay Vo , Mi Young Lee , and Sung Wook Baik 



Research Article (12 pages), Article ID 8460934, Volume 2019 (2019)

A Differential Evolution-Oriented Pruning Neural Network Model for Bankruptcy Prediction

Yajiao Tang, Junkai Ji , Yulin Zhu, Shangce Gao , Zheng Tang, and Yuki Todo 

Research Article (21 pages), Article ID 8682124, Volume 2019 (2019)

Analysis of Financing Efficiency of Chinese Agricultural Listed Companies Based on Machine Learning

Lixia Liu  and Xueli Zhan 



Research Article (11 pages), Article ID 9190273, Volume 2019 (2019)

Application of BP Neural Network Model in Risk Evaluation of Railway Construction

Yang Changwei , Li Zonghao , Guo Xueyan , Yu Wenying , Jin Jing , and Zhu Liang 


Research Article (12 pages), Article ID 2946158, Volume 2019 (2019)

Stock Price Pattern Prediction Based on Complex Network and Machine Learning

Hongduo Cao , Tiantian Lin, Ying Li , and Hanyu Zhang


Research Article (12 pages), Article ID 4132485, Volume 2019 (2019)

Big Data Market Optimization Pricing Model Based on Data Quality

Jian Yang , Chongchong Zhao, and Chunxiao Xing





Research Article (10 pages), Article ID 5964068, Volume 2019 (2019)

Pricing Strategies in Dual-Channel Supply Chain with a Fair Caring Retailer

Lufeng Dai , Xifu Wang, Xiaoguang Liu, and Lai Wei



Research Article (23 pages), Article ID 1484372, Volume 2019 (2019)

An Improved Demand Forecasting Model Using Deep Learning Approach and Proposed Decision Integration Strategy for Supply Chain

Zeynep Hilal Kilimci , A. Okay Akyuz , Mitat Uysal, Selim Akyokus , M. Ozan Uysal, Berna Atak Bulbul , and Mehmet Ali Ekmis

Research Article (15 pages), Article ID 9067367, Volume 2019 (2019)

Is Deep Learning for Image Recognition Applicable to Stock Market Prediction?

Hyun Sik Sim , Hae In Kim, and Jae Joon Ahn 

Research Article (10 pages), Article ID 4324878, Volume 2019 (2019)

Editorial

Applications of Machine Learning Methods in Complex Economics and Financial Networks

Benjamin M. Tabak ¹, **Thiago C. Silva** ², **Liang Zhao**,³ and **Ahmet Sensoy**⁴

¹Fundação Getúlio Vargas, Escola de Políticas Públicas e Governo (FGV/EPPG), Brasília, Brazil

²Universidade Católica de Brasília and Universidade de São Paulo, Brasília, Brazil

³Universidade de São Paulo, Ribeirão Preto, Brazil

⁴Bilkent University, Ankara, Turkey

Correspondence should be addressed to Benjamin M. Tabak; benjaminm.tabak@gmail.com

Received 20 February 2020; Accepted 21 February 2020; Published 25 April 2020

Copyright © 2020 Benjamin M. Tabak et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The availability of large databases and significant improvements in computational power has been key determinants in the explosive increase of interest in machine learning. In this sense, machine-learning methods, such as neural networks and genetic algorithms, have been used as methodological tools to understand how complex adaptive systems behave and to integrate many streams of unstructured and structured data. Economics and finance, on the flipside, have experienced an increasing interest in micro-level analysis, but the empirical methodologies are restricted to mostly linear methods brought by traditional econometric methods.

This cross-discipline special issue aims at integrating conceptual methodologies of the machine-learning domain with empirical issues that we find in economics and finance. There is a large room for exploration at the intersection of these two areas. Machine learning goes beyond regression methods, and we can use them in a variety of ways. Thus, it can give new insights on how economics and finance data are organized. The application of these methods may contribute to the debate on assessing, monitoring, and forecasting economic and financial variables which is quite relevant.

In this special issue, we welcome new insights, models, and applications in a wide variety of topics that bridge topics in machine learning to complex economics and finance networks. The application and adaptation of re-supervised learning methods, such as data and community clustering, ranking, anomaly detection, and semisupervised

and supervised learning techniques, such as classification and regression, applied to finance and economics, are of great interest.

There are many gaps in the literature, and we address some of them within this special issue. We provide a variety of papers that contribute to the debate on the use of machine learning in economics and finance.

In this special issue, we collect several contributions. We have papers that study consumer loans “Modeling Repayment Behavior of Consumer Loan in Portfolio across Business Cycle: A Triplet Markov Model Approach,” trading strategies “Modeling Investor Behavior Using Machine Learning: Mean-Reversion and Momentum Trading Strategies,” public procurement announcements “Public Procurement Announcements in Spain: Regulations, Data Analysis, and Award Price Estimator Using Machine Learning,” exchange rate forecasts “Chinese Currency Exchange Rates Forecasting with EMD-Based Neural Network,” internalization of RMB “A Study of RMB Internationalization Path Based on Border Area Perspective,” and bankruptcy prediction “A Hybrid Approach Using Oversampling Technique and Cost-Sensitive Learning for Bankruptcy Prediction” and “A Differential Evolution-Oriented Pruning Neural Network Model for Bankruptcy Prediction.”

Few papers also discuss efficiency “Analysis of Financing Efficiency of Chinese Agricultural Listed Companies Based on Machine Learning,” risk evaluation “Application of BP Neural Network Model in Risk Evaluation of Railway

Construction,” stock price prediction “Stock Price Pattern Prediction Based on Complex Network and Machine Learning” and “Is Deep Learning for Image Recognition Applicable to Stock Market Prediction?” pricing models and strategies “Big Data Market Optimization Pricing Model Based on Data Quality” and “Pricing Strategies in Dual-Channel Supply Chain with a Fair Caring Retailer,” demand forecasting “An Improved Demand Forecasting Model Using Deep Learning Approach and Proposed Decision Integration Strategy for Supply Chain,” portfolio optimization “Portfolio Optimization with Asset-Liability Ratio Regulation Constraints,” and measure intimacy “Measure User Intimacy by Mining Maximum Information Transmission Paths.”

Further research could also employ novel methods to exploit prediction of crashes [1], evaluate bank system supervision [2], evaluate dynamic trees for financial data [3], and study efficiency of institutions [4–6].

Conflicts of Interest

The editors declare that they have no conflicts of interest regarding the publication of this special issue.

Acknowledgments

This work was supported in part by the São Paulo State Research Foundation (FAPESP) under grant numbers 2015/50122-0 and 2013/07375-0, Pró-Reitoria de Pesquisa of University of São Paulo (PRP-USP) under grant number 2018.1.1702.59.8, and the Brazilian National Council for Scientific and Technological Development (CNPq) under grant number 303199/2019-9. Thiago C. Silva (Grant nos. 308171/2019-5 and 408546/2018-2) and Benjamin M. Tabak (Grant nos. 310541/2018-2 and 425123/2018-9) gratefully acknowledge financial support from the CNPq foundation.

*Benjamin M. Tabak
Thiago C. Silva
Liang Zhao
Ahmet Sensoy*

References

- [1] D. O. Cajueiro, B. M. Tabak, and F. K. Werneck, “Can we predict crashes? The case of the Brazilian stock market,” *Physica A: Statistical Mechanics and Its Applications*, vol. 388, no. 8, pp. 1603–1609, 2009.
- [2] T. Papadimitriou, P. Gogas, and B. M. Tabak, “Complex networks and banking systems supervision,” *Physica A: Statistical Mechanics and Its Applications*, vol. 392, no. 19, pp. 4429–4434, 2013.
- [3] A. Sensoy and B. M. Tabak, “Dynamic spanning trees in stock market networks: the case of Asia-Pacific,” *Physica A: Statistical Mechanics and Its Applications*, vol. 414, pp. 387–402, 2014.
- [4] T. C. Silva, S. M. Guerra, B. M. Tabak, and R. C. C. Miranda, “Financial networks, bank efficiency and risk-taking,” *Journal of Financial Stability*, vol. 25, pp. 247–257, 2016.
- [5] T. C. Silva, B. M. Tabak, D. O. Cajueiro, and M. V. B. Dias, “A comparison of DEA and SFA using micro- and macro-level perspectives: efficiency of Chinese local banks,” *Physica A: Statistical Mechanics and Its Applications*, vol. 469, pp. 216–223, 2017.
- [6] T. C. Silva, B. M. Tabak, D. O. Cajueiro, and M. V. B. Dias, “Adequacy of deterministic and parametric frontiers to analyze the efficiency of Indian commercial banks,” *Physica A: Statistical Mechanics and Its Applications*, vol. 506, pp. 1016–1025, 2018.

Research Article

Portfolio Optimization with Asset-Liability Ratio Regulation Constraints

De-Lei Sheng ¹ and Peilong Shen²

¹*School of Applied Mathematics, Shanxi University of Finance and Economics, Taiyuan 030006, China*

²*School of Finance, Shanxi University of Finance and Economics, Taiyuan 030006, China*

Correspondence should be addressed to De-Lei Sheng; tjhsdl@126.com

Received 12 June 2019; Accepted 8 February 2020; Published 23 March 2020

Guest Editor: Thiago Christiano Silva

Copyright © 2020 De-Lei Sheng and Peilong Shen. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

This paper considers both a top regulation bound and a bottom regulation bound imposed on the asset-liability ratio at the regulatory time T to reduce risks of abnormal high-speed growth of asset price within a short period of time (or high investment leverage), and to mitigate risks of low assets' return (or a sharp fall). Applying the stochastic optimal control technique, a Hamilton–Jacobi–Bellman (HJB) equation is derived. Then, the effective investment strategy and the minimum variance are obtained explicitly by using the Lagrange duality method. Moreover, some numerical examples are provided to verify the effectiveness of our results.

1. Introduction

Taking liabilities into the traditional portfolio models, Sharpe and Tint [1] put forward the asset-liability problem for pension fund management under the mean-variance framework. Kell and Müller [2] point out that liabilities affect the efficient frontier of this asset-liability problem. But early research on asset-liability problems was limited to the standard single-period mean-variance criterion; Leippold et al. [3] obtain an analytical optimal strategy and efficient frontier for asset-liability problems by using embedding technique proposed by Li and Ng [4] under a multiperiod mean-variance framework. Leippold et al. [3] also consider the multiperiod mean-variance asset-liability problem and show that the optimal strategy can be decomposed into orthogonal sets and the efficient frontier spanned by an orthogonal basis of dynamic returns. For the continuous-time asset-liability problem, Chiu and Li [5] derive the optimal policy and the mean-variance efficient frontier by applying the technique of stochastic linear-quadratic control. In comparison with the embedding technique and the stochastic linear-quadratic control method, the Lagrange duality method is more convenient to solve the mean-

variance problem. Fu et al. [6] consider the continuous-time mean-variance portfolio selection under different borrowing interest rates and lending interest rates with the Lagrange duality method. Pan and Xiao [7] also investigate the continuous-time asset-liability management problem by considering the stochastic interest rates and inflation risks. Considering a financial market consists of a risk-free bond, a stock, and a derivative, Li et al. [8] give the optimal investment strategies of a continuous-time mean-variance asset-liability management in presence of stochastic volatility. For other related literature studies, you can refer to Wei and Wang [9]; Zhang et al. [10]; Duarte et al. [11]; and so on.

As everyone knows, the assets scale pursued by investment growing too fast over a relatively short period of time, which indicates that the growth rate increasing faster than the normal rate of growth, is itself accompanied by high potential risks. The most typical example in practice is the stock market crash which has happened many times almost in all countries with stock markets. The earliest stock market crash occurred in 1720, the Mississippi Stock Disaster in France and the South Sea Stock Disaster in Britain. The share price of the Mississippi company rose from around 500 livres

in May 1719, to nearly 10,000 livres in February 1720, but it declined to 500 livres in September 1721. The stock prices of South Sea Company soar from £128 in January, £175 in February, £330 in March, and £550 in May, and then, the shares leaped to £1000 per share by August 1720 and finally peaked at this level but soon it plunged and triggered an avalanche of selling. As for the most devastating stock market crashes of the United States, such as the American stock market crash in 1929 and the American stock market crash in 1987, all ended in a catastrophic decline after a period of drastic growth. When the China stock market crash broke out in 2007, the Shanghai stock exchange composite index surged from around 2700 points to the peak value 6124 points, which took only a few months. But then it went down sharply to the lowest value 1644 points. Another China stock market crash in 2015, the Shanghai stock exchange composite index rise from around 3000 points to 5178 points on June 12, but only in the next two months, it plunged to around 2800 points.

The sustained high-speed growth of stock prices within a short period of time is actually accumulating enough destructive energy which may erupt at any time. Sophisticated investors will actively sell their risky stocks in time when the stock returns reach a certain high level, instead of blindly pursuing much higher returns, so as to avoid the sharp drop before these assets are successfully sold. This high return level targeted by these traders to sell assets actively is a practical top bound used in investment management practice and determined by investment managers initiatively, which is actually the concrete example for the application of top investment regulation bound in business. Thus, it confirms the real existence of the investment regulation top bound in real investment practice. The requirement of investment regulation bottom bound is necessary for the pursuit of much higher investment returns. Only if the assets scale pursued by investment is larger than liability, then it can ensure that the investment activities are really valuable and vigorous.

These kinds of investment risks of asset price collapse resulting from the abnormally rapid rise within a limited time period have extremely destructive, which may even be the real manifestation of systemic financial risks. As for systemic financial risks, Guerra et al. [12] and Souza et al. [13] conduct in-depth and innovative research studies and propose a novel methodology to measure systemic risk in networks composed of financial institutions. They define the bank's probability of default and calculate this probability using Merton's method inspired by Black and Scholes [14] and Merton [15]. Interestingly, the probability of default defined by Guerra et al. [12] and Souza et al. [13] can also be used to research the asset-liability management problem with unbearable collapse risk resulted from the abnormally rapid rise of asset price within a limited time period. For instance, you can take the horizon on which the entity (firm) may default to be short enough and then large declines in assets or large increases in liabilities can be researched similar to the probability of default calculated by Merton's method. However, based on different considerations, we realize our ideas about the asset-liability management problem with unbearable investment risk using a widely used mathematical method in this paper.

Financial leverage needs to be considered because it is ubiquitous in practice, and the excessive leverage is an issue tackled in the Basel III requirements. In the narrow sense, financial leverage refers to a measure of operating large-scale business with less money, which is widely used in various economic activities. For example, the transaction relying on futures margin is commonly used in futures markets and the margin rate usually varies between 5% and 15% of the total contract value. Only a down payment of 5% of the full contract value is required, and an investor can carry out the financial transaction equivalent to 20 times the amount of the margin. Another example is the way to invest in real estate with installment repayment, and investors only need to pay the down payment, usually 10% to 30%, but they can leverage the investment scales of the total market value. Quite evidently, investors generally use smaller self-owned funds to operate businesses of great market value with financial leverage. If these assets appreciate, investors can immediately obtain considerable investment returns. However, once the value of these assets shrinks, the losses are magnified and the corresponding leverage multiples are catastrophic and unbearable. Therefore, the essence of financial leverage lies in magnifying the gains and losses to meet the needs of investors who have insufficient funds but want to do large-scale businesses. There is a large body in the literature of financial stability which estimates systemic risk and explicitly takes into account excessive leverage. Related works can be referred to the study by Silva et al. [16] and many others.

However, financial markets themselves also have the function of enlarging profits and losses. As long as investors invest in risky assets in financial markets, they are actually using the financial market to enlarge their scale of assets, which is essentially consistent with the core meaning of financial leverage in narrow sense. Therefore, in a broad sense, a financial market itself can be regarded as a financial leverage. In other words, the financial leverage has already been used by investors as long as they invest in financial markets in pursuit of higher investment returns, so it is also necessary to consider leverage regulation even if you only invest in ordinary financial markets.

Different from the previous research on the modeling method of the asset-liability problem, both a bottom regulation bound and a top regulation bound are imposed on the asset-liability ratio to control the variation range so as to prevent the risks of asset price collapse resulted from an abnormal high-speed growth (such as stock market crash), asset shrink, and the collapse of investment leverage. The dynamic process of asset-liability ratio is defined as an asset process being divided by a liability process after eliminating the influence of inflation. A stochastic optimal control model is formulated under the framework of mean variance with the variance being minimized, given a determined expectation of asset-liability ratio at regulatory time T . Using the Lagrange multiplier method, the original problem is transformed into an unconstrained optimization problem, and then a Hamilton–Jacobi–Bellman (HJB) equation is established by adopting technique of stochastic control. At last, using Lagrange duality between the original problem

and the unconstrained problem, the minimum variance and the effective investment strategy are obtained.

The remainder of this paper is organized as follows. Section 2 describes the models of asset-liability ratio and the constrained control problem. In Section 3, a Lagrange unconstrained problem is solved by technique of stochastic optimal control, and the results of the original problem are obtained according to Lagrange duality. Meanwhile, a special case is also solved at the end of this section. Some interpretations of main results are presented in Section 4, and numerical examples are illustrated in Section 5. At last, Section 6 gives a conclusion of the research work.

2. Formulation of the Model

Throughout this paper, $(\Omega, \mathcal{F}, P, \{F_t\}_{0 \leq t \leq T})$ denotes a complete probability space satisfying the usual condition. A finite constant $T > 0$ represents the preselected investment regulatory time; \mathcal{F}_t is the smallest σ -field generated by all random information available until time t , and all random variables and stochastic processes involved in this article are \mathcal{F}_t measurable for every $t \in [0, T]$.

2.1. Financial Market. Similar to the previous research work, this paper considers an inflation-affected financial market in continuous time, which consists of one risk-free bond, one inflation-linked index bond, and one risky stock. The inflation rate $P(t)$ can be regarded as the Consumer Price Index, which is described by a price level process as follows:

$$P(t) = \exp \left\{ \int_0^t \left(\tilde{\mu}_s - \frac{1}{2} \tilde{\sigma}_s^2 \right) ds + \int_0^t \tilde{\sigma}_s d\tilde{W}(s) \right\}, \quad (1)$$

To avoid the influence of control variable on the liability process $L(t)$, we also assume the liability process is governed by a geometric Brownian motion as follows:

$$L(t) = L_0 \exp \left\{ \int_0^t \left(\mu_s - \frac{1}{2} \sigma_s^2 \right) ds + \int_0^t \sigma_s dW(s) \right\}, \quad L_0 > 0. \quad (5)$$

The real liability process $\hat{L}(t)$ after eliminating inflation is defined as $\hat{L}(t) = L(t)/P(t)$. It is easy to get the following form of expression using Itô's formula:

where $\tilde{\mu}_t$ is the expected inflation rate at time t ; $\tilde{\sigma}_t > 0$ is the volatility of inflation rate; and $\tilde{W}(t)$ is a standard Brownian motion on the probability space (Ω, \mathcal{F}, P) .

The price process of the risk-free bond is modeled as $B(t) = e^{\tilde{r}t}$, where the constant $\tilde{r} \geq 0$ represents the nominal interest rate.

The inflation-linked index bond $\{I(t), t \geq 0\}$ has the same risk source with the price level process, and thus, it can be expressed as

$$I(t) = \exp \left\{ \int_0^t \left((r + \tilde{\mu}_s) - \frac{1}{2} \tilde{\sigma}_s^2 \right) ds + \int_0^t \tilde{\sigma}_s d\tilde{W}(s) \right\}, \quad (2)$$

where r is the real interest rate at time t .

The price process of risky stock is formulated as

$$S(t) = S_0 \exp \left\{ \int_0^t \left(\hat{\mu}_s - \frac{1}{2} \hat{\sigma}_s^2 \right) ds + \int_0^t \hat{\sigma}_s d\hat{W}(s) \right\}, \quad (3)$$

where $\hat{\mu}_t$ is the investment return rate satisfying $\hat{\mu}_t > \tilde{r}$ and $\hat{\sigma}_t$ is the volatility of the risky asset.

2.2. Asset Process and Liability Process. A control vector $(\pi_0(t), \pi_1(t), \pi_2(t))$ represents the investment strategy need to be found, where $\pi_0(t)$ denotes the investment share of risk-free bond, $\pi_1(t)$ represents the investment share of inflation-linked index bond, and $\pi_2(t)$ signifies the investment share of risky stock, and they always satisfy the equation $\pi_0(t) + \pi_1(t) + \pi_2(t) = 1$.

The company's accumulated wealth is allowed to invest in the financial market; thus, the asset process can be described as an investment process. Assuming that $\pi := \{(\pi_0(t), \pi_1(t), \pi_2(t))\}_{0 \leq t \leq T}$ is an adaptive control variable, the dynamics of asset process is governed by the following equation:

$$\begin{cases} \frac{dX^\pi(t)}{X^\pi(t)} = \pi_0(t) \frac{dB(t)}{B(t)} + \pi_1(t) \frac{dI(t)}{I(t)} + \pi_2(t) \frac{dS(t)}{S(t)} \\ = [\tilde{r} + \pi_1(t)(\tilde{\mu}_t + r - \tilde{r}) + \pi_2(t)(\hat{\mu}_t - \tilde{r})] dt + \pi_1(t) \tilde{\sigma}_t d\tilde{W}(t) + \pi_2(t) \hat{\sigma}_t d\hat{W}(t), \\ X^\pi(0) = x_0. \end{cases} \quad (4)$$

$$\frac{d\hat{L}(t)}{\hat{L}(t)} = (\mu_t - \tilde{\mu}_t + \tilde{\sigma}_t^2) dt - \tilde{\sigma}_t d\tilde{W}(t) + \sigma_t dW(t),$$

$$\begin{aligned} \hat{L}(t) = L_0 \exp \left\{ \int_0^t \left(\mu_s - \tilde{\mu}_s + \frac{1}{2} \tilde{\sigma}_s^2 - \frac{1}{2} \sigma_s^2 \right) ds - \int_0^t \tilde{\sigma}_s d\tilde{W}(s) \right. \\ \left. + \int_0^t \sigma_s dW(s) \right\}. \end{aligned} \quad (6)$$

The asset process after eliminating the influence of inflation is defined as $\hat{X}^\pi(t) = X^\pi(t)/P(t)$; then, we get the following asset process according to Itô's formula:

$$\begin{aligned} \frac{d\widehat{X}^\pi(t)}{\widehat{X}^\pi(t)} &= [\tilde{r} - \tilde{\mu}_t + \tilde{\sigma}_t^2 + \pi_1(t)(\tilde{\mu}_t + r - \tilde{r} - \tilde{\sigma}_t^2) \\ &\quad + \pi_2(t)(\tilde{\mu}_t - \tilde{r})]dt + (\pi_1(t) - 1)\tilde{\sigma}_t d\widehat{W}(t) \\ &\quad + \pi_2(t)\tilde{\sigma}_t d\widehat{W}(t). \end{aligned} \quad (7)$$

2.3. *Asset-Liability Problem with Regulation Constraints.* The *asset-liability ratio* is defined as $Z^\pi(t) = \widehat{X}^\pi(t)/\widehat{L}(t)$, which describes the ratio between the size of asset and the size of liability at time t . According to Itô's formula, the dynamic process of the *asset-liability ratio* can be given by

$$\begin{aligned} \frac{dZ^\pi(t)}{Z^\pi(t)} &= (\tilde{r} - \mu_t - \sigma_t^2)dt - \sigma_t dW(t) + \pi_1(t)((\tilde{\mu}_t + r - \tilde{r})dt \\ &\quad + \tilde{\sigma}_t d\widehat{W}(t)) + \pi_2(t)((\tilde{\mu}_t - \tilde{r})dt + \tilde{\sigma}_t d\widehat{W}(t)). \end{aligned} \quad (8)$$

In practice, the regulation only occurs at certain fixed time which is called regulatory time. The regulatory time is usually selected in advance, which provides a time period of appropriate length, but the length may decrease if the regulatory frequency increased. In accordance with the need of practice, this paper considers the asset-liability problem with the constraints imposed at the regulatory time T to find the optimal investment strategy, in order to reduce the risk of abnormal high-speed growth of asset price within a short period of time or high investment leverage and also to lessen the risk of too low return rate or a sharp fall. Their mathematical descriptions of the *regulation constraints* at the regulatory time T are formulated as the following inequality:

$$\alpha < Z^\pi(T) < \beta, \quad (9)$$

where the *bottom regulation bound* $\alpha \geq 1$ and the *top regulation bound* β are two appropriate constants satisfying the requirement $\beta > \alpha$. The bottom bound α should not be too small; otherwise, it cannot cover the liability which means the investment has failed. The top bound β should not be too large; or else, a much larger value of β may lead to excessive leverage multiples and too much investment risks being pulled in the company. In practice, it is highly technical to determine the values of α and β , which requires abundant

practical experience and rigorous mathematical calculation, and only can be completed by well-trained and experienced investment managers based on the real market quotation and their practical experiences.

Let Π denote the set of all admissible strategies π . The investor aims to find an optimal portfolio $\pi \in \Pi$ of the following *original problem*:

$$\begin{cases} \min_{\pi \in \Pi} & \mathbf{Var}[Z^\pi(T)] = \mathbf{E}[Z^\pi(T) - (\alpha + \rho(\beta - \alpha))]^2 \\ \text{s.t.} & \mathbf{E}[Z^\pi(T)] = \alpha + \rho(\beta - \alpha) > 0, \\ & Z^\pi(t) \text{ satisfy (8), } \rho \in (0, 1), \text{ for } \pi \in \Pi, \end{cases} \quad (10)$$

where ρ is a constant. Only if ρ takes an unequivocal value, then the expectation of $Z^\pi(T)$ can be fixed at a certain value between the top regulation bound β and the bottom regulation bound α .

For a determined value of expectation $\alpha + \rho(\beta - \alpha)$, if there exists at least one admissible pair $(Z(\cdot), \pi(\cdot))$ satisfying $\mathbf{E}[Z^\pi(T)] = \alpha + \rho(\beta - \alpha)$, then the problem (10) is called *feasible*. Given $\alpha + \rho(\beta - \alpha)$, the optimal strategy π^* of (10) is called an *efficient strategy*. The pair $(Z^{\pi^*}(T), \mathbf{Var}[Z^{\pi^*}(T)])$ under this optimal strategy π^* is called an *efficient point*. The set of all efficient points is called the *efficient frontier*. Obviously, problem (10) is a dynamic quadratic convex optimization problem, and thus, it has an unique solution.

3. Solutions of the Problem

In this section, an unconstrained optimization problem is derived by the Lagrange multiplier method and then, a Hamilton–Jacobi–Bellman equation is deduced to solve the unconstrained problem. Next, solution of the original problem is obtained according to Lagrange duality between the unconstrained problem and the original problem. At the end, solution of a special case is given for comparison.

3.1. *Solution of an Unconstrained Problem.* For this convex optimization problem (10), the equality constraint $\mathbf{E}[Z^\pi(T)] = \alpha + \rho(\beta - \alpha)$ can be eliminated by using the Lagrange multiplier method, so we get a Lagrange dual problem:

$$\begin{cases} \min_{\pi \in \Pi} & \mathbf{E}[Z^\pi(T) - (\alpha + \rho(\beta - \alpha))]^2 + 2\lambda \mathbf{E}[Z^\pi(T) - (\alpha + \rho(\beta - \alpha))], \\ \text{s.t.} & Z^\pi(T) \text{ satisfy (8), } \rho \in (0, 1), \text{ for } \pi \in \Pi, \end{cases} \quad (11)$$

where $\lambda \in \mathbf{R}$ is a Lagrange multiplier.

After a simple calculation, it yields

$$\begin{aligned} & \mathbf{E}[Z^\pi(T) - (\alpha + \rho(\beta - \alpha))]^2 + 2\lambda \mathbf{E}[Z^\pi(T) - (\alpha + \rho(\beta - \alpha))] \\ &= \mathbf{E}[Z^\pi(T) - ((\alpha + \rho(\beta - \alpha)) - \lambda)]^2 - \lambda^2. \end{aligned} \quad (12)$$

Remark 1. The link between problems (10) and (11) is provided by the Lagrange duality theorem, which can be referred in the study by Luenberger [17] that

$$\begin{aligned} & \min_{\pi \in \prod} \mathbf{Var}[Z^\pi(T)] \\ & = \max_{\lambda \in \mathbf{R}} \min_{\pi \in \prod} \left\{ \mathbf{E}[Z^\pi(T) - ((\alpha + \rho(\beta - \alpha)) - \lambda)]^2 - \lambda^2 \right\}. \end{aligned} \quad (13)$$

Thus, for the fixed constants λ and u , problem (11) is equivalent to

$$\begin{cases} \min_{\pi \in \prod} \mathbf{E}[Z^\pi(T) - ((\alpha + \rho(\beta - \alpha)) - \lambda)]^2 \\ \text{s.t. } Z^\pi(t) \text{ satisfy (8), } \rho \in (0, 1), \text{ for } \pi \in \prod. \end{cases} \quad (14)$$

Therefore, in order to solve problem (11), we only need to solve problem (14). Fortunately, problem (14) can be solved by using the stochastic optimal control technology.

To solve problem (14), a truncated form beginning at time t is considered here, and the corresponding value function is defined as

$$V(t, z) = \inf_{\pi \in \prod} \mathbf{E} \left[(Z^\pi(T) - ((\alpha + \rho(\beta - \alpha)) - \lambda))^2 \mid_{Z^\pi(t)=z} \right], \quad (15)$$

with the boundary condition $V(T, z) = (z - ((\alpha + \rho(\beta - \alpha)) - \lambda))^2$.

The controlled infinitesimal generator for any test function $\phi(t, x)$ and any admissible control π is deduced as follows:

$$\begin{aligned} \mathcal{A}^\pi \phi(t, x) &= \phi_t + x \phi_x [\bar{r} - \mu_t - \sigma_t^2 + \pi_1(t)(\bar{\mu}_t + r - \bar{r}) \\ &+ \pi_2(t)(\hat{\mu}_t - \bar{r})] + \frac{x^2}{2} \phi_{xx} [\sigma_t^2 + \pi_1^2(t)\bar{\sigma}_t^2 + \pi_2^2(t)\hat{\sigma}_t^2], \end{aligned} \quad (16)$$

for any real valued function $\phi(t, x) \in \mathbf{C}^{1,2}([0, T], \mathbf{R})$ and admissible strategy π , where

$$\begin{aligned} \mathbf{C}^{1,2}([0, T], \mathbf{R}) &= \{\psi(t, x) \mid \psi(t, \cdot) \text{ is once continuously differentiable on } [0, T], \text{ and } \psi \\ &(\cdot, x) \text{ is twice continuously differentiable almost surely on } \mathbf{R}\}. \end{aligned} \quad (17)$$

If $V(t, z) \in \mathbf{C}^{1,2}([0, T] \times \mathbf{R})$, then $V(t, z)$ satisfies the following HJB equation:

$$\inf_{\pi \in \prod} \mathcal{A}^\pi V(t, z) = 0, \quad (18)$$

which can be given more specifically as follows:

$$\inf_{\pi \in \prod} \left\{ V_t + zV_z(\bar{r} - \mu_t - \sigma_t^2) + \frac{z^2}{2}V_{zz}\sigma_t^2 + zV_z[\pi_1(t)(\bar{\mu}_t + r - \bar{r}) + \pi_2(t)(\hat{\mu}_t - \bar{r})] + \frac{z^2}{2}V_{zz}(\pi_1^2(t)\bar{\sigma}_t^2 + \pi_2^2(t)\hat{\sigma}_t^2) = 0 \right\}. \quad (19)$$

For convenience, some simple notations are introduced as follows:

$$\begin{aligned} \delta &\triangleq \frac{(\bar{\mu}_t + r - \bar{r})^2}{\bar{\sigma}_t^2} + \frac{(\hat{\mu}_t - \bar{r})^2}{\hat{\sigma}_t^2}, \\ \varrho(t) &\triangleq \frac{(e^{(t-T)(\delta + \sigma_t^2)}\delta + \sigma_t^2)}{\delta + \sigma_t^2}, \\ \omega &\triangleq \frac{(\alpha + \rho(\beta - \alpha))}{(1 - \hat{\varrho})}, \\ \hat{\varrho} &\triangleq \varrho(0) = \frac{(e^{-T(\delta + \sigma_t^2)}\delta + \sigma_t^2)}{\delta + \sigma_t^2}, \\ \varepsilon &\triangleq \frac{e^{-T(-\bar{r} + \delta + \mu_t + \sigma_t^2)}z_0}{(1 - \hat{\varrho})}. \end{aligned} \quad (20)$$

Solving equation (19) gives the following important theorem.

Theorem 1. For the asset-liability management with investment regulation and eliminating inflation, the efficient investment strategy $\pi^* = (\pi_0^*, \pi_1^*, \pi_2^*)$ corresponding to problem (11) and problem (14) is given by

$$\begin{cases} \pi_0^* = 1 + \delta \left(1 - \frac{1}{z} \xi_\alpha^\beta(t) \right), \\ \pi_1^* = \frac{(\bar{\mu}_t + r - \bar{r})}{\bar{\sigma}_t^2} \left(\frac{\xi_\alpha^\beta(t)}{z} - 1 \right), \\ \pi_2^* = \frac{(\hat{\mu}_t - \bar{r})}{\hat{\sigma}_t^2} \left(\frac{\xi_\alpha^\beta(t)}{z} - 1 \right), \end{cases} \quad (21)$$

and the value function is

$$\begin{aligned}
V(t, z) &= z^2 e^{(t-T)(-\tilde{r}+\delta+2\mu_t+\sigma_t^2)} \\
&\quad - 2ze^{(t-T)(-\tilde{r}+\delta+\mu_t+\sigma_t^2)} \\
&\quad + ((\alpha + \rho(\beta - \alpha)) - \lambda)^2 \varrho(t),
\end{aligned} \tag{22}$$

where $\xi_\alpha^\beta(t) = e^{(t-T)(\tilde{r}-\mu_t)} ((\alpha + \rho(\beta - \alpha)) - \lambda)$.

Proof. According to the first-order necessary condition of extremum, equation (19) yields

$$\begin{cases} \pi_1^*(t) = -\frac{(\tilde{\mu}_t + r - \tilde{r})}{z\tilde{\sigma}_t^2} \frac{V_z}{V_{zz}}, \\ \pi_2^*(t) = -\frac{(\hat{\mu}_t - \tilde{r})}{z\tilde{\sigma}_t^2} \frac{V_z}{V_{zz}}. \end{cases} \tag{23}$$

Plugging the above expressions of $\pi_1^*(t)$ and $\pi_2^*(t)$ into (19), the HJB equation becomes

$$V_t + zV_z(\tilde{r} - \mu_t - \sigma_t^2) + \frac{z^2}{2}V_{zz}\sigma_t^2 - \delta\frac{V_z}{2V_{zz}} = 0. \tag{24}$$

Based on the boundary condition $V(T, z) = (z - ((\alpha + \rho(\beta - \alpha)) - \lambda))^2$, we try a conjecture

$$W(t, z) = f(t)z^2 + g(t)z + h(t), \tag{25}$$

for equation (24), which satisfies

$$\begin{aligned}
f(T) &= 1, \\
g(T) &= -2((\alpha + \rho(\beta - \alpha)) - \lambda), \\
h(T) &= ((\alpha + \rho(\beta - \alpha)) - \lambda)^2.
\end{aligned} \tag{26}$$

The derivatives of this conjecture $W(t, z)$ are given as follows:

$$\begin{aligned}
W_t &= f'(t)z^2 + g'(t)z + h'(t), \\
W_z &= 2f(t)z + g(t), \\
W_{zz} &= 2f(t).
\end{aligned} \tag{27}$$

Plugging the expressions of W_z and W_{zz} into (23), the optimal strategy becomes

$$\begin{cases} \pi_1^*(t) = -\frac{(\tilde{\mu}_t + r - \tilde{r})}{z\tilde{\sigma}_t^2} \left(z + \frac{g(t)}{2f(t)} \right), \\ \pi_2^*(t) = -\frac{(\hat{\mu}_t - \tilde{r})}{z\tilde{\sigma}_t^2} \left(z + \frac{g(t)}{2f(t)} \right). \end{cases} \tag{28}$$

Substituting these derivatives of $W(t, z)$ into (24), the HJB equation becomes

$$\begin{aligned}
&f'(t)z^2 + g'(t)z + h'(t) + (2f(t)z^2 + g(t)z)(\tilde{r} - \mu_t - \sigma_t^2) \\
&+ z^2 f(t)\sigma_t^2 - \delta \left(f(t)z^2 + g(t)z + \frac{g^2(t)}{4f(t)} \right) = 0.
\end{aligned} \tag{29}$$

Equation (29) can be split into three ordinary differential equations as follows:

$$\begin{cases} f'(t) + 2f(t)(\tilde{r} - \mu_t) - f(t)\sigma_t^2 - \delta f(t) = 0, \\ f(T) = 1, \\ g'(t) + g(t)(\tilde{r} - \mu_t - \sigma_t^2) - \delta g(t) = 0, \\ g(T) = -2((\alpha + \rho(\beta - \alpha)) - \lambda), \\ h'(t) - \frac{\delta}{4} \frac{g(t)^2}{f(t)} = 0, \\ h(T) = ((\alpha + \rho(\beta - \alpha)) - \lambda)^2. \end{cases} \tag{30}$$

Solving these ordinary differential equations in (30) yields

$$\begin{cases} f(t) = e^{(t-T)(-\tilde{r}+\delta+2\mu_t+\sigma_t^2)}, \\ g(t) = -2e^{(t-T)(-\tilde{r}+\delta+\mu_t+\sigma_t^2)} ((\alpha + \rho(\beta - \alpha)) - \lambda), \\ h(t) = ((\alpha + \rho(\beta - \alpha)) - \lambda)^2 \varrho(t). \end{cases} \tag{31}$$

Plugging the explicit expression of $f(t)$ and $g(t)$ into (28), and using the equality $\pi_0^*(t) + \pi_1^*(t) + \pi_2^*(t) = 1$, the optimal strategy is obtained as follows:

$$\begin{cases} \pi_0^*(t) = 1 + \delta \left(1 - e^{(t-T)(\tilde{r}-\mu_t)} \frac{((\alpha + \rho(\beta - \alpha)) - \lambda)}{z} \right), \\ \pi_1^*(t) = -\frac{(\tilde{\mu}_t + r - \tilde{r})}{\tilde{\sigma}_t^2} + \frac{(\tilde{\mu}_t + r - \tilde{r})}{z\tilde{\sigma}_t^2} e^{(t-T)(\tilde{r}-\mu_t)} ((\alpha + \rho(\beta - \alpha)) - \lambda), \\ \pi_2^*(t) = -\frac{(\hat{\mu}_t - \tilde{r})}{\tilde{\sigma}_t^2} + \frac{(\hat{\mu}_t - \tilde{r})}{z\tilde{\sigma}_t^2} e^{(t-T)(\tilde{r}-\mu_t)} ((\alpha + \rho(\beta - \alpha)) - \lambda). \end{cases} \tag{32}$$

Denoting

$$\xi_\alpha^\beta(t) \triangleq e^{(t-T)(\tilde{r}-\mu_t)} ((\alpha + \rho(\beta - \alpha)) - \lambda), \quad (33)$$

the optimal strategy can be simplified into

$$\begin{cases} \pi_0^*(t) = 1 + \delta \left(1 - \frac{1}{z} \xi_\alpha^\beta(t) \right), \\ \pi_1^*(t) = -\frac{(\tilde{\mu}_t + r - \tilde{r})}{\tilde{\sigma}_t^2} + \frac{(\tilde{\mu}_t + r - \tilde{r})}{z\tilde{\sigma}_t^2} \xi_\alpha^\beta(t), \\ \pi_2^*(t) = -\frac{(\tilde{\mu}_t - \tilde{r})}{\tilde{\sigma}_t^2} + \frac{(\tilde{\mu}_t - \tilde{r})}{z\tilde{\sigma}_t^2} \xi_\alpha^\beta(t). \end{cases} \quad (34)$$

Meanwhile, substituting the expressions of $f(t)$, $g(t)$, and $h(t)$ into $W(t, z)$, the explicit solution of HJB equation is obtained as follows:

$$\begin{aligned} W(t, z) &= z^2 e^{(t-T)(-2\tilde{r} + \delta + 2\mu_t + \sigma_t^2)} + ((\alpha + \rho(\beta - \alpha)) - \lambda)^2 \\ &\quad \varrho(t) - 2ze^{(t-T)(-\tilde{r} + \delta + \mu_t + \sigma_t^2)} ((\alpha + \rho(\beta - \alpha)) - \lambda). \end{aligned} \quad (35)$$

□

The following verification theorem shows the obtained results are exactly the optimal strategy and the optimal value function as required.

Theorem 2 (Verification Theorem). *If $W(t, z)$ is a solution of HJB equation (19) and satisfies the boundary condition $W(T, z) = (z - ((\alpha + \rho(\beta - \alpha)) - \lambda))^2$, then for all admissible strategies $\pi \in \Pi$, $V(t, z) \geq W(t, z)$. If π^* satisfies*

$$\pi^* \in \arg \inf_{\pi \in \Pi} \mathbf{E} \left[(Z^\pi(T) - ((\alpha + \rho(\beta - \alpha)) - \lambda))^2 | Z^\pi(t) = z \right], \quad (36)$$

then $V(t, z) = W(t, z)$ and π^* is the optimal investment strategy of problem (14).

Proof. Proof is similar to that given by Pham [18]; Chang [19]; and so on, so the detail is omitted. □

3.2. Solution of the Original Problem. The optimal value function of problem (11) is defined as

$$\hat{V}(0, z_0) = \inf_{\pi \in \Pi} \mathbf{E} \left[(Z^\pi(T) - ((\alpha + \rho(\beta - \alpha)) - \lambda))^2 \right] - \lambda^2, \quad (37)$$

where $z_0 = Z(0)$.

In this section, the Lagrange duality method is used to find a solution of original optimization problem (10) based on the obtained results of the unconstrained problem.

Theorem 3. *For original problem (10), the efficient strategy $\pi^* = (\pi_0^*, \pi_1^*, \pi_2^*)$ is given by*

$$\begin{cases} \pi_0^* = 1 - \delta \left(\frac{\xi_\alpha^\beta(t)}{z} - 1 \right), \\ \pi_1^* = \frac{(\tilde{\mu}_t + r - \tilde{r})}{\tilde{\sigma}_t^2} \left(\frac{\xi_\alpha^\beta(t)}{z} - 1 \right), \\ \pi_2^* = \frac{(\tilde{\mu}_t - \tilde{r})}{\tilde{\sigma}_t^2} \left(\frac{\xi_\alpha^\beta(t)}{z} - 1 \right), \end{cases} \quad (38)$$

and the optimal value function is given by

$$\begin{aligned} \mathbf{Var}[Z^{\pi^*}(T)] &= e^{-T(-2\tilde{r} + \delta + 2\mu_t + \sigma_t^2)} z_0^2 + (\omega - \varepsilon)^2 \hat{\varrho} \\ &\quad - 2e^{-T(-\tilde{r} + \delta + \mu_t + \sigma_t^2)} (\omega - \varepsilon) z_0 - (\varepsilon - \hat{\varrho}\omega)^2, \end{aligned} \quad (39)$$

where

$$\xi_\alpha^\beta(t) = e^{(t-T)(\tilde{r}-\mu_t)} (\omega - \varepsilon). \quad (40)$$

Proof. The optimal value function $\hat{V}(0, z_0)$ can be obtained with the equivalence between problem (11) and problem (14). By taking $z_0 = Z(0)$ and setting $t = 0$ in $V(t, z)$, it yields

$$\begin{aligned} \hat{V}(0, z_0) &= V(0, z_0) - \lambda^2 = f(0)z_0^2 + g(0)z_0 + h(0) - \lambda^2 \\ &= e^{-T(-2\tilde{r} + \delta + 2\mu_t + \sigma_t^2)} z_0^2 + ((\alpha + \rho(\beta - \alpha)) - \lambda)^2 \\ &\quad \hat{\varrho} - 2e^{-T(-\tilde{r} + \delta + \mu_t + \sigma_t^2)} ((\alpha + \rho(\beta - \alpha)) - \lambda)z_0 - \lambda^2. \end{aligned} \quad (41)$$

Since $(\delta + \sigma_t^2) > 0$, then $e^{-T(\delta + \sigma_t^2)} < 1$, and it yields

$$\left(e^{-T(\delta + \sigma_t^2)} \delta + \sigma_t^2 \right) < \delta + \sigma_t^2, \quad (42)$$

and then

$$\hat{\varrho} = \frac{(e^{-T(\delta + \sigma_t^2)} \delta + \sigma_t^2)}{\delta + \sigma_t^2} < 1. \quad (43)$$

Note that the optimal value function $\hat{V}(0, z_0)$ is a quadratic function with respect to λ and $\hat{\varrho} - 1$ is the coefficient of quadratic term:

$$\hat{\varrho} - 1 = \frac{(e^{-T(\delta + \sigma_t^2)} \delta + \sigma_t^2)}{\delta + \sigma_t^2} - 1 < 0. \quad (44)$$

Thus, the finite maximum value of $\hat{V}(0, z_0)$ can be obtained at a specific value λ^* , and

$$\lambda^* = \frac{e^{-T(-\tilde{r} + \delta + \mu_t + \sigma_t^2)} z_0}{(1 - \hat{\varrho})} - \frac{\hat{\varrho}(\alpha + \rho(\beta - \alpha))}{(1 - \hat{\varrho})}. \quad (45)$$

Applying the Lagrangian duality, substituting λ^* into $\hat{V}(0, z_0)$, the minimum variance corresponding to an arbitrarily given expectation $(\alpha + \rho(\beta - \alpha))$ is obtained as follows:

$$\begin{aligned} \mathbf{Var}[Z^{\pi^*}(T)] &= e^{-T(-2\tilde{r}+\delta+2\mu_t+\sigma_t^2)} z_0^2 + \left(\frac{(\alpha + \rho(\beta - \alpha))}{(1 - \hat{\varrho})} - \frac{e^{-T(-\tilde{r}+\delta+\mu_t+\sigma_t^2)} z_0}{(1 - \hat{\varrho})} \right)^2 \\ &\quad \cdot \hat{\varrho} - 2e^{-T(-\tilde{r}+\delta+\mu_t+\sigma_t^2)} \left(\frac{(\alpha + \rho(\beta - \alpha))}{(1 - \hat{\varrho})} - \frac{e^{-T(-\tilde{r}+\delta+\mu_t+\sigma_t^2)} z_0}{(1 - \hat{\varrho})} \right) z_0 - \left(\frac{e^{-T(-\tilde{r}+\delta+\mu_t+\sigma_t^2)} z_0}{(1 - \hat{\varrho})} - \frac{\hat{\varrho}(\alpha + \rho(\beta - \alpha))}{(1 - \hat{\varrho})} \right)^2. \end{aligned} \quad (46)$$

Substitute (45) into (21), the optimal investment strategy of problem (10) is given by

$$\begin{cases} \pi_0^* = 1 - \delta \left(\frac{1}{z} \zeta_\alpha^\beta(t) - 1 \right), \\ \pi_1^* = -\frac{(\tilde{\mu}_t + r - \tilde{r})}{\tilde{\sigma}_t^2} + \frac{(\tilde{\mu}_t + r - \tilde{r})}{z\tilde{\sigma}_t^2} \zeta_\alpha^\beta(t), \\ \pi_2^* = -\frac{(\hat{\mu}_t - \tilde{r})}{\tilde{\sigma}_t^2} + \frac{(\hat{\mu}_t - \tilde{r})}{z\tilde{\sigma}_t^2} \zeta_\alpha^\beta(t), \end{cases} \quad (47)$$

where

$$\zeta_\alpha^\beta(t) = e^{(t-T)(\tilde{r}-\mu_t)} \left(\frac{(\alpha + \rho(\beta - \alpha))}{1 - \hat{\varrho}} - \frac{e^{-T(-\tilde{r}+\delta+\mu_t+\sigma_t^2)} z_0}{1 - \hat{\varrho}} \right). \quad (48)$$

□

Remark 2. However, in the obtained results (47) and (46) above, the expected value $(\alpha + \rho(\beta - \alpha))$ of $Z^\pi(t)$ is not fixed at a determined position but can take any value between α and β depending on the value of $\rho \in (0, 1)$. Therefore, the obtained minimum variance depends on the parameter value $\rho \in (0, 1)$ in the expression of expectation $(\alpha + \rho(\beta - \alpha))$. Next, let us turn our attention to the parameter ρ and determine the optimal parameter value ρ^* for minimizing the variance with respect to ρ as well. The optimal ρ^* to achieve the minimum of $\mathbf{Var}[Z^{\pi^*}(T)]$ is given as follows:

$$\rho^* = \frac{\left(\left(\frac{e^{-T(-\tilde{r}+\delta+\mu_t+\sigma_t^2)} z_0}{(1 - \hat{\varrho})} \right) - \alpha \right)}{(\beta - \alpha)}, \quad (49)$$

and the expected value of $Z^\pi(t)$ corresponding to the optimal parameter value ρ^* is given by

$$\alpha + \rho^*(\beta - \alpha) = \frac{z_0}{\hat{\varrho}} e^{-T(-\tilde{r}+\delta+\mu_t+\sigma_t^2)}. \quad (50)$$

Substituting ρ^* into (47), we can get the optimal strategy corresponding to the optimal value $\alpha + \rho^* - \alpha$ of expectation. Moreover, plugging the expression of ρ^* into the formula of $\mathbf{Var}[Z^{\pi^*}(T)]$, the optimal minimum variance $\mathbf{Var}[Z^{\pi^*,\rho^*}(T)]$ can also be determined corresponding to the uniquely optimal value $(z_0/\hat{\varrho})e^{-T(-\tilde{r}+\delta+\mu_t+\sigma_t^2)}$ of the expectation.

3.3. Special Case. The investment regulation is also important for an investment process without liabilities. Let $\mu_t = \sigma_t = 0$, the original problem degenerates to the case of no liability, and the process $Z^\pi(t)$ becomes the dynamics process $\hat{X}^\pi(t)$, as follows:

$$\begin{aligned} d\hat{X}^\pi(t) &= [\tilde{r} + \pi_1(t)(\tilde{\mu}_t + r - \tilde{r}) + \pi_2(t)(\hat{\mu}_t - \tilde{r}) - \pi_1(t)\tilde{\sigma}_t^2 - \tilde{\mu}_t] \\ &\quad \cdot \hat{X}^\pi(t)dt + (\pi_1(t) - 1)\tilde{\sigma}_t \hat{X}^\pi(t)d\tilde{W}(t) + \pi_2(t)\tilde{\sigma}_t \\ &\quad \cdot \hat{X}^\pi(t)d\tilde{W}(t). \end{aligned} \quad (51)$$

Meanwhile, the regulation bound constraints also degenerate into a restriction on the investment dynamics process after eliminating inflation:

$$\alpha < \hat{X}^\pi(T) < \beta. \quad (52)$$

Proposition 1. For the special case $\mu_t = \sigma_t = 0$, the minimum variance of $Z^\pi(t)$ is

$$\begin{aligned} \mathbf{Var}[Z^*(T)] &= e^{-T\tilde{\delta}} \left(\frac{(\alpha + \rho(\beta - \alpha))}{(1 - e^{-T\tilde{\delta}})} - \frac{e^{-T(-r+\delta+\tilde{\sigma}_t^2)} \hat{x}_0}{(1 - e^{-T\tilde{\delta}})} \right)^2 - 2e^{-T(-r+\delta+\tilde{\sigma}_t^2)} \left(\frac{(\alpha + \rho(\beta - \alpha))}{(1 - e^{-T\tilde{\delta}})} - \frac{e^{-T(-r+\delta+\tilde{\sigma}_t^2)} \hat{x}_0}{(1 - e^{-T\tilde{\delta}})} \right) \hat{x}_0 \\ &\quad - \left(\frac{e^{-T(-r+\delta+\tilde{\sigma}_t^2)} \hat{x}_0 - e^{-T\tilde{\delta}}(\alpha + \rho(\beta - \alpha))}{(1 - e^{-T\tilde{\delta}})} \right)^2 + e^{-T(-2r+\delta+2\tilde{\sigma}_t^2)} \hat{x}_0^2, \end{aligned} \quad (53)$$

and the optimal strategy of the special case is given as follows:

$$\begin{aligned}
\pi_1^*(t) &= 1 - \frac{(\tilde{\mu}_t + r - \tilde{r} - \tilde{\sigma}_t^2)}{\tilde{\sigma}_t^2} \left(1 - \frac{e^{(t-T)(-r+\tilde{\delta}+\tilde{\sigma}_t^2)}(\alpha + \rho(\beta - \alpha)) - e^{-T(-r+\tilde{\delta}+\tilde{\sigma}_t^2)}\hat{x}_0}{\hat{x}e^{(t-T)(-2r+\tilde{\delta}+2\tilde{\sigma}_t^2)}(1 - e^{-T\tilde{\delta}})} \right), \\
\pi_2^*(t) &= -\frac{(\hat{\mu}_t - \tilde{r})}{\tilde{\sigma}_t^2} \left(1 - \frac{e^{(t-T)(-r+\tilde{\delta}+\tilde{\sigma}_t^2)}(\alpha + \rho(\beta - \alpha)) - e^{-T(-r+\tilde{\delta}+\tilde{\sigma}_t^2)}\hat{x}_0}{\hat{x}e^{(t-T)(-2r+\tilde{\delta}+2\tilde{\sigma}_t^2)}(1 - e^{-T\tilde{\delta}})} \right), \\
\pi_0^*(t) &= \tilde{\delta} \left(1 - \frac{e^{(t-T)(-r+\tilde{\delta}+\tilde{\sigma}_t^2)}(\alpha + \rho(\beta - \alpha)) - e^{-T(-r+\tilde{\delta}+\tilde{\sigma}_t^2)}\hat{x}_0}{\hat{x}e^{(t-T)(-2r+\tilde{\delta}+2\tilde{\sigma}_t^2)}(1 - e^{-T\tilde{\delta}})} \right),
\end{aligned} \tag{54}$$

where

$$\tilde{\delta} = \frac{(\tilde{\mu}_t + r - \tilde{r} - \tilde{\sigma}_t^2)^2}{\tilde{\sigma}_t^2} + \frac{(\hat{\mu}_t - \tilde{r})^2}{\tilde{\sigma}_t^2}. \tag{55}$$

Proof. The results can be obtained by the same calculation process as the original problem, so the details are omitted here. \square

Remark 3. The expected value $(\alpha + \rho(\beta - \alpha))$ is not fixed at a determined position in the above proposition, but it can take any value between α and β depending on the value of $\rho \in (0, 1)$. Therefore, the obtained minimum variance depends on the parameter value $\rho \in (0, 1)$. Using the first-order necessary condition of extremum value for $\mathbf{Var}[Z^*(T)]$ with respect to the parameter ρ , the optimal value of parameter ρ is given by

$$\rho^* = \frac{e^{-T(-r+\tilde{\sigma}_t^2)}\hat{x}_0 - \alpha}{(\beta - \alpha)}, \tag{56}$$

and the corresponding optimal expectation

$$(\alpha + \rho^*(\beta - \alpha)) = e^{-T(-r+\tilde{\sigma}_t^2)}\hat{x}_0. \tag{57}$$

Substituting ρ^* into the optimal strategy $(\pi_0^*(t), \pi_1^*(t), \pi_2^*(t))$, the optimal strategy corresponding to the optimal value $\alpha + \rho^*(\beta - \alpha)$ of expectation can be obtained. Meanwhile, plugging the expression of ρ^* into $\mathbf{Var}[Z^{\pi^*}(t)]$, the optimal minimum variance $\mathbf{Var}[Z^{\pi^*, \rho^*}(T)]$ can also be determined. It is worth mentioning that the optimal minimum variance for this special case with no liability is zero under the uniquely optimal expectation $(z_0/\hat{\varrho})e^{-T(-r+\delta+\mu_t+\sigma_t^2)}$.

4. Interpretations of the Main Results

The expressions of the obtained results are so abstract that it is necessary to give some interpretations for the main results to clarify their specific meaning, especially for the investment share $\pi_1(t)$ of the inflation-linked index bond and the investment share $\pi_2(t)$ of the risky stock.

Scrutinizing features of the expressions $\pi_1(t)$ and $\pi_2(t)$,

$$\begin{aligned}
\pi_1^* &= \frac{(\tilde{\mu}_t + r - \tilde{r})}{\tilde{\sigma}_t^2} \left(\frac{\zeta_\alpha^\beta(t)}{z} - 1 \right), \\
\pi_2^* &= \frac{(\hat{\mu}_t - \tilde{r})}{\tilde{\sigma}_t^2} \left(\frac{\zeta_\alpha^\beta(t)}{z} - 1 \right),
\end{aligned} \tag{58}$$

where $\zeta_\alpha^\beta(t) = e^{(t-T)(\tilde{r}-\mu_t)}(\omega - \varepsilon)$, $\omega = (\alpha + \rho(\beta - \alpha))/(1 - \hat{\varrho})$, and $\varepsilon = (e^{-T(-\tilde{r}+\delta+\mu_t+\sigma_t^2)}z_0)/(1 - \hat{\varrho})$, it is easy to find that understanding implications of these simplified notations $\delta, \hat{\varrho}, \omega, \varepsilon$, and $\zeta_\alpha^\beta(t)$ is the key point. Subsequently, the interpretations will start from descriptions of these simplified notations.

In notation $\delta = ((\tilde{\mu}_t + r - \tilde{r})^2/\tilde{\sigma}_t^2) + ((\hat{\mu}_t - \tilde{r})^2/\tilde{\sigma}_t^2)$, the one part $(\tilde{\mu}_t + r - \tilde{r})^2/\tilde{\sigma}_t^2$ is the risk compensation rate of inflation-linked index bond and the other part $(\hat{\mu}_t - \tilde{r})^2/\tilde{\sigma}_t^2$ is the risk compensation rate of risky stock. Thus, δ synthetically reflects the risk compensation of two risky assets.

Obviously, $e^{-T(\delta+\sigma_t^2)} < 1$, so the notation $\hat{\varrho} = (e^{-T(\delta+\sigma_t^2)}\delta + \sigma_t^2)/(\delta + \sigma_t^2)$ is less than 1. Notation $\hat{\varrho}$ gives a ratio of one sum with the risk compensation rate δ being converted by $e^{-T(\delta+\sigma_t^2)}$ divided by the other sum in which the risk compensation rate δ has no conversion.

The notation $\omega = (\alpha + \rho(\beta - \alpha))/(1 - \hat{\varrho})$ is an expression of top regulation bound β and bottom regulation bound α , where $\alpha + \rho(\beta - \alpha)$ is a value of the expectation $\mathbf{E}[Z^T(T)]$. Because ρ takes a determined value, the changes of regulation bounds directly result in the changes of ω . Therefore, ω is the key point reflecting the influences of both regulation bounds on investment strategy.

In notation ε , the parameter z_0 has the most obvious economic connotation, since it represents the initial value of asset-liability ratio at time 0. The expression $(e^{-T(-\tilde{r}+\delta+\mu_t+\sigma_t^2)}z_0)/(1 - \hat{\varrho})$ gives a value resulted from the initial asset-liability ratio z_0 influenced by several different factors (return rates and volatilities) and then converted at the constant relative rate $\hat{\varrho}$.

The following expression

$$\frac{\zeta_\alpha^\beta(t)}{z} = \frac{e^{(t-T)(\tilde{r}-\mu_t)}}{(1 - \hat{\varrho})} \cdot \frac{(\alpha + \rho(\beta - \alpha)) - e^{-T(-\tilde{r}+\delta+\mu_t+\sigma_t^2)}z_0}{z} \tag{59}$$

is key to understand the optimal strategy $(\pi_0^*, \pi_1^*, \pi_2^*)$. The denominator z of this fraction $\zeta_\alpha^\beta(t)/z$ represents the current level of asset-liability ratio at time t .

But the numerator $\zeta_\alpha^\beta(t)$ of this fraction is more complicated. The one multiplier $((\alpha + \rho(\beta - \alpha)) - z_0 e^{-T(-\tilde{r} + \delta + \mu + \tilde{\sigma}_t^2)})$ is a difference between the expected value $(\alpha + \rho(\beta - \alpha))$ of $Z(T)$ and the converted value $e^{-T(-\tilde{r} + \delta + \mu + \tilde{\sigma}_t^2)} z_0$ of the initial value z_0 , which characterizes the distance between initial value affected by various economic factors and the expectation of $Z(T)$. The other multiplier $e^{(t-T)(\tilde{r} - \mu_t)/(1 - \tilde{\rho})}$ is a converted value comprehensively influenced by several parameters embodying the states of financial market.

In addition, it is easy to see that if the relative rate $\zeta_\alpha^\beta(t)/z$ is larger than 1, $\pi_1^*(t) = ((\tilde{\mu}_t + r - \tilde{r})/\tilde{\sigma}_t^2)((\zeta_\alpha^\beta(t)/z) - 1)$ and $\pi_2^*(t) = ((\tilde{\mu}_t - \tilde{r})/\tilde{\sigma}_t^2)(\zeta_\alpha^\beta(t)/z - 1)$ are positive; otherwise, both will be negative. Since $\pi_0^*(t) = 1 - \pi_1^*(t) - \pi_2^*(t)$, it indicates that $\pi_0^*(t) = 1 + \delta(1 - (1/z)\zeta_\alpha^\beta(t))$ can be determined as long as the investment shares of risky assets have been determined, so the adjustment of risk-free asset is usually passive. If the relative rate $\zeta_\alpha^\beta(t)/z$ is greater than 1, the investment shares on risk-free asset can be ensured within a much better range between 0 and 1. These observations can serve as an important reference for determining both the top regulation bound β and the bottom regulation bound α .

5. Numerical Examples

This section discusses the variation tendency of optimal strategies π_1^* and π_2^* varying with some important parameters, mainly focusing on the current value z of asset-liability ratio, the top regulation bound β , the bottom regulation bound α , and the parameter ρ . The values of parameters are given in Table 1; unless otherwise specified, other values for the same variable can be consulted from the corresponding graphic legend.

First, let top regulation bound β be fixed; if the bottom regulation bound takes much higher values, then the investment shares of both inflation-linked index bond and risky stock increase significantly. These variety trends can be observed intuitively from Figures 1 and 2, respectively. Meanwhile, Figures 1 and 2 also show that the appropriate bottom regulation bound should be taken at one value between 1 and 2 for an economic environment same as this example. Otherwise, the investment shares on risky assets may increase too much, but the investment share on risk-free asset becomes negative, which may result in leverage risk increases rapidly.

Next, let bottom regulation bound α be fixed; if the top regulation bound β takes much higher value, then the investment shares of both inflation-linked index bond and risky stock should also increase significantly. The variety trends can be perceived from Figures 3 and 4, respectively. Meanwhile, Figures 3 and 4 also show that the appropriate top regulation bound should be taken at one value between 6 and 7 for an economic environment same as this example. Otherwise, the investment shares on risky assets increase too much, but the investment share on risk-free asset becomes negative, which also result in leverage risk increases rapidly.

As can be seen from Figures 3–6, taking the larger value of either the top regulation bound or the bottom regulation

TABLE 1: Values of the parameters.

$\tilde{\mu} = 0.08$	$r = 0.01$	$\tilde{\sigma} = 0.2$	$\alpha = 1$	$\beta = 7$	$z_0 = 3$	$z = 5$
$\tilde{r} = 0.05$	$T = 6$	$\tilde{\mu} = 0.15$	$\tilde{\sigma} = 0.6$	$\mu = 0.03$	$\sigma = 0.1$	$\rho = 0.5$

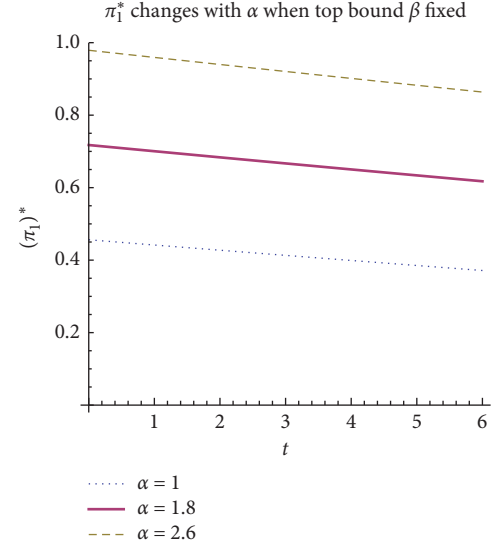


FIGURE 1: Values of π_1^* change with t for different bottom bound α .

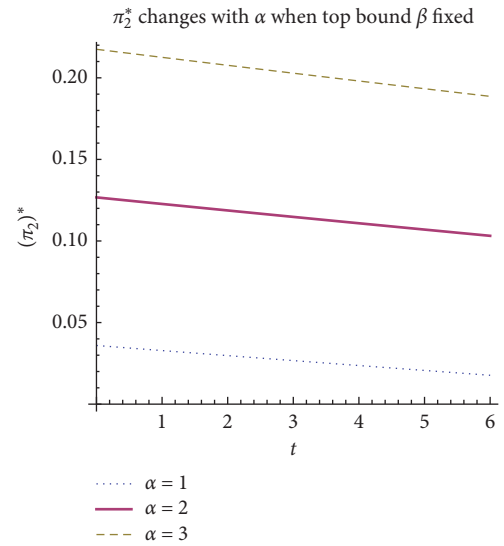


FIGURE 2: Values of π_2^* change with t for different bottom bound α .

bound means allowing a wider volatility range of investment returns, which may lead to the result that assets with greater investment risk are allowed to invest, or greater market risks are incorporated into the wealth process. In addition, the above numerical analysis roughly gives a valuable reference for selecting an appropriate range of top regulation bound and bottom regulation bound according to the relationship between the reasonable values of investment strategy and regulation bounds for an economic environment set by values of parameters.

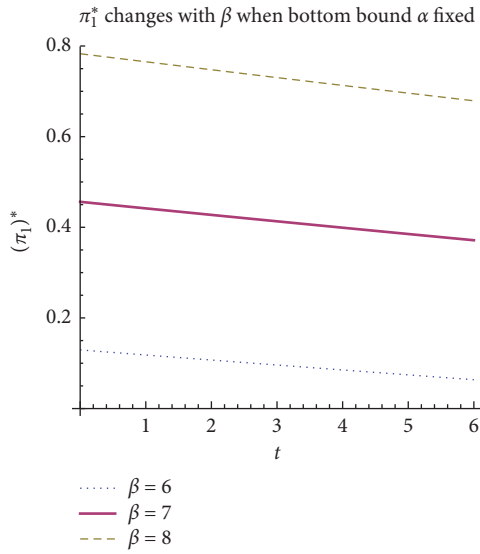


FIGURE 3: Values of π_1^* change with t for different top bound β .

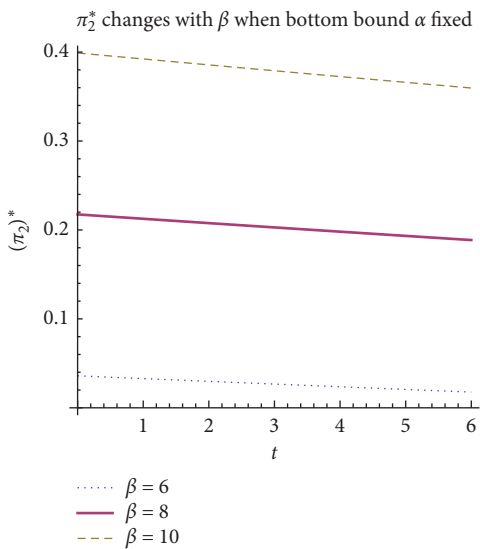


FIGURE 4: Values of π_2^* change with t for different top bound β .

In fact, the regulatory bound cannot be determined arbitrarily. Both the top regulatory bound and bottom regulatory bound determine the fluctuation range that decision-making needs to control. Too narrow fluctuation range of investment returns results in loss of great investment opportunities and makes investment in the financial market meaningless. If the regulatory bound is too large, the potential huge risks cannot be dealt with effectively when the investment return soars fast, which is almost equivalent to situations without regulatory bounds and also lose the significance of using regulatory bounds. It is hard to determine the regulatory bounds only by quantitative calculation. Once there is a model error, it may result in serious deviation between theory and practice in actual management; thus, the determination of upper and lower limits should not be purely theoretical. Nevertheless, it should be

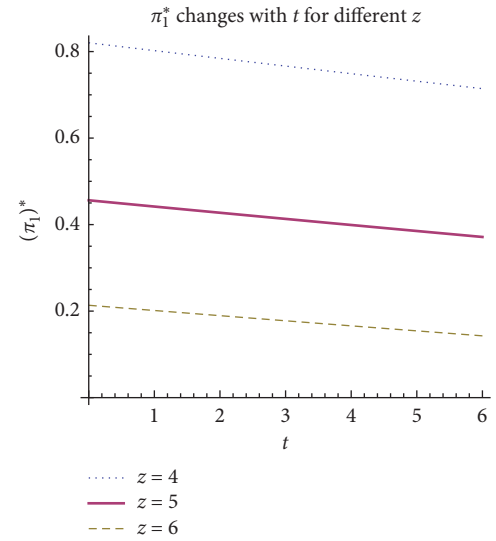


FIGURE 5: Values of π_1^* change with t for different current asset-liability ratio z .

noted that advanced statistical analysis and maturely practical experience of personnel are essential and of great advantage for determining an exact value of α or β .

Assuming that both top regulation bound β and bottom regulation bound α are determined, the impact of asset-liability ratio z at time t on investment strategy can be observed in Figures 5 and 6. If the current asset-liability ratio z at time t takes much greater value, then the investment shares of both inflation-linked index bond and risky stock should be much lower, which shows that increasing shares of investment on risk-free asset is an important measure to reduce investment risk. For the numerical examples shown in Figures 5 and 6, the regulation bounds being fixed at $\alpha = 1$ and $\beta = 7$, the most appropriate asset-liability ratio should take values around 5 which is much more advantageous for the company.

The impact of parameter ρ on investment strategy can be observed in Figures 7 and 8. If the parameter ρ takes a much greater value, then the investment shares of risky assets must increase correspondingly. Since the top regulation bound β and the bottom regulation bound α are predetermined, the expected value $\alpha + \rho(\beta - \alpha)$ of asset-liability ratio entirely depends on the value of parameter ρ . Therefore, the variation tendency of investment shares $\pi_1^*(t)$ and $\pi_2^*(t)$ with parameter ρ actually illustrates the trend of investment shares changing with the expectation of asset-liability ratio $Z(T)$. In one word, the greater the expectation of $Z(T)$, the much higher the shares of the company's wealth to be invested on risky assets.

6. Conclusion

Some empirical studies in the banking literature deal with the effects of prolonged periods of low interest rates in the economy on risk-taking, real effects on the real sector and also on financial stability and so on as given by Chaudron [20]; Bikker; and Vervliet [21]. However, this paper uses

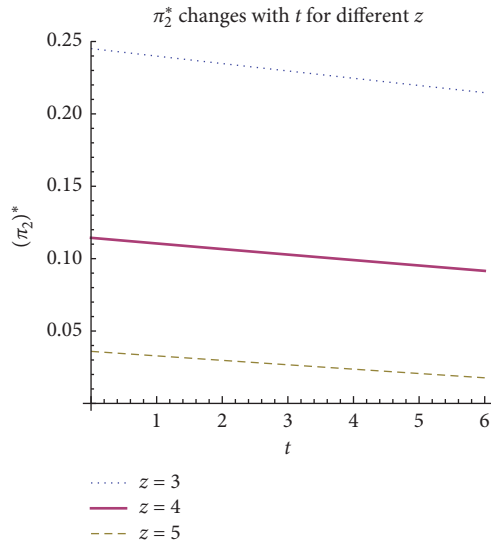


FIGURE 6: Values of π_2^* change with t for different current asset-liability ratio z .

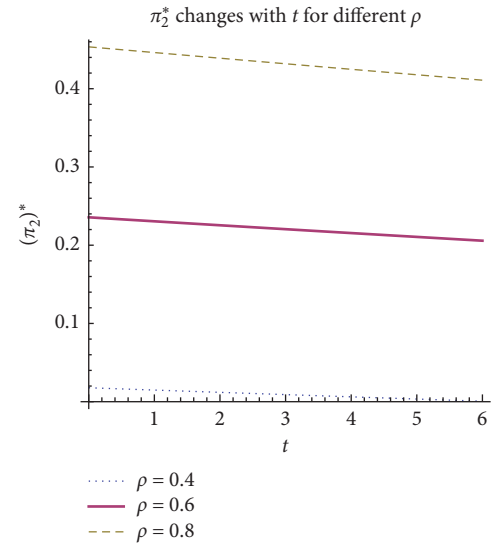


FIGURE 8: Values of π_2^* change with t for different ρ .

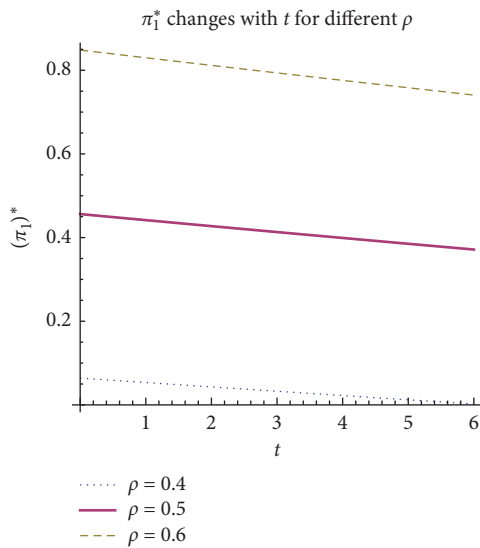


FIGURE 7: Values of π_1^* change with t for different ρ .

mathematical models and methods to study a problem of asset-liability management with financial market risks: risks of too low return rate, risks of the abnormal rapid growth of return rate, and risks of investment leverage. A model of quantitative regulation (both the top regulation bound and the bottom regulation bound being imposed on the asset-liability ratio) is put forward. The efficient strategy and efficient frontier are obtained under the objective of variance minimization with regulation constraints at the regulatory time T . But the most important value of a theoretical research lies in its guidance and reference for practice. Through numerical examples, it is found that the obtained explicit optimal strategy can also provide reference for determination of regulation bounds, which reinforces the theoretical significance of π_1^* of this research work.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was supported by China Postdoctoral Science Foundation funded project (Grant no. 2017M611192), Youth Science Fund of Shanxi University of Finance and Economics (Grant no. QN-2017019), and Youth Science Foundation of National Natural Science Fund (Grant no. 11801179).

References

- [1] W. F. Sharpe and L. G. Tint, "Liabilities- a new approach," *The Journal of Portfolio Management*, vol. 16, no. 2, pp. 5–10, 1990.
- [2] A. Kell and H. Müller, "Efficient portfolios in the asset liability context," *ASTIN Bulletin*, vol. 25, no. 1, pp. 33–48, 1995.
- [3] M. Leippold, F. Trojani, and P. Vanini, "A geometric approach to multiperiod mean variance optimization of assets and liabilities," *Journal of Economic Dynamics and Control*, vol. 28, no. 6, pp. 1079–1113, 2004.
- [4] D. Li and W.-L. Ng, "Optimal dynamic portfolio selection: multiperiod mean-variance formulation," *Mathematical Finance*, vol. 10, no. 3, pp. 387–406, 2000.
- [5] M. C. Chiu and D. Li, "Asset and liability management under a continuous-time mean-variance optimization framework," *Insurance: Mathematics and Economics*, vol. 39, no. 3, pp. 330–355, 2006.
- [6] C. Fu, A. Lari-Lavassani, and X. Li, "Dynamic mean-variance portfolio selection with borrowing constraint," *European Journal of Operational Research*, vol. 200, no. 1, pp. 312–319, 2010.

- [7] J. Pan and Q. Xiao, "Optimal mean-variance asset-liability management with stochastic interest rates and inflation risks," *Mathematical Methods of Operations Research*, vol. 85, no. 3, pp. 491–519, 2017.
- [8] D. Li, Y. Shen, and Y. Zeng, "Dynamic derivative-based investment strategy for mean-variance asset-liability management with stochastic volatility," *Insurance: Mathematics and Economics*, vol. 78, pp. 72–86, 2018.
- [9] J. Wei and T. Wang, "Time-consistent mean-variance asset-liability management with random coefficients," *Insurance: Mathematics and Economics*, vol. 77, pp. 84–96, 2017.
- [10] Y. Zhang, Y. Wu, S. Li, and B. Wiwatanapataphee, "Mean-variance asset liability management with state-dependent risk aversion," *North American Actuarial Journal*, vol. 21, no. 1, pp. 87–106, 2017.
- [11] T. B. Duarte, D. M. Valladão, and Á. Veiga, "Asset liability management for open pension schemes using multistage stochastic programming under solvency-II-based regulatory constraints," *Insurance: Mathematics and Economics*, vol. 77, pp. 177–188, 2017.
- [12] S. M. Guerra, T. C. Silva, B. M. Tabak, R. A. S. Penalosa, and R. C. C. Miranda, "Systemic risk measures," *Physica A*, vol. 442, pp. 329–342, 2015.
- [13] S. R. S. D. Souza, T. C. Silva, B. M. Tabak, and S. M. Guerra, "Evaluating systemic risk using bank default probabilities in financial networks," *Journal of Economic Dynamics and Control*, vol. 66, pp. 54–75, 2016.
- [14] F. Black and M. Scholes, "The pricing of options and corporate liabilities," *Journal of Political Economy*, vol. 81, no. 3, pp. 637–654, 1973.
- [15] R. C. Merton, "On the pricing of corporate debt: the risk structure of interest rates," *The Journal of Finance*, vol. 29, no. 2, pp. 449–470, 1974.
- [16] T. C. Silva, S. R. S. Souza, and B. M. Tabak, "Monitoring vulnerability and impact diffusion in financial networks," *Journal of Economic Dynamics and Control*, vol. 76, pp. 109–135, 2017.
- [17] D. G. Luenberger, *Optimization by Vector Space Methods*, Wiley, New York, NY, USA, 1968.
- [18] H. Pham, *Continuous-time Stochastic Control and Optimization with Financial Applications*, Springer, Berlin, Germany, 2009.
- [19] H. Chang, "Dynamic mean-variance portfolio selection with liability and stochastic interest rate," *Economic Modelling*, vol. 51, pp. 172–182, 2015.
- [20] R. Chaudron, *Bank Profitability and Risk Taking in a Prolonged Environment of Low Interest Rates: A Study of Interest Rate Risk in The Banking Book of Dutch Banks*, DNB Working Papers, Amsterdam, Netherlands, 2016.
- [21] J. A. Bikker and T. M. Vervliet, "Bank profitability and risk-taking under low interest rates," *International Journal of Finance & Economics*, vol. 23, no. 1, pp. 3–18, 2018.

Research Article

Measure User Intimacy by Mining Maximum Information Transmission Paths

Lin Guo ¹ and Dongliang Zhang²

¹School of Economics and Management, Changchun University of Science and Technology, Changchun, Jilin 130022, China

²Institution of Technical Science, Fudan University, Shanghai 200000, China

Correspondence should be addressed to Lin Guo; guolin@cust.edu.cn

Received 17 June 2019; Accepted 29 November 2019; Published 21 March 2020

Guest Editor: Thiago Christiano Silva

Copyright © 2020 Lin Guo and Dongliang Zhang. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The Internet has become an important carrier of information. Its data contain abundant information about hot events, user relations and attitudes, and so on. Many enterprises use high-impact Internet users to promote products, so it is very important to understand the mechanism of information transmission. Mining social network data can help people analyze the complex and changing relationships between users. The traditional method for doing this is to analyze information such as common interests and common friends, but this data cannot truly describe the degree of intimacy between users. What really connects different users on the Internet is the delivery of information. The algorithm proposed in this paper considers the dynamic characteristics of information transmission, finds maximum transmission paths from information transmission results, and finally calculates the intimacy degrees between users according to all the maximum information transmission paths within a certain period.

1. Introduction

Social network data contains a wealth of information about events, relationships, and attitudes. On the basis of fully understanding and analyzing the data, a series of technologies, such as text mining, statistical theory, association analysis, and visualization technologies, are adopted to realize emotional orientation analysis, information extraction, user influence analysis, and so on. Many current methods of computing user intimacy can be applied to static networks. However, users might unfollow certain friends and their interests might shift to new and different topics. In other words, the tie strengths between different users change over time. The algorithm proposed in this paper takes the dynamic nature of data into account to improve information transmission analysis in social networks. After that maximum transmission paths are identified in the information transmission results, and then the intimacy degrees between nodes can be computed according to multiple groups of maximum transmission paths.

The remainder of this paper is organized as follows: Section 2 introduces the related work of this paper. Section 3

proposes the concept of the information transmission matrix. Section 4 introduces the computational process of tie strength. Section 5 states the experimental results. Section 6 introduces the conclusions.

2. Related Works

Many enterprises use influential users to promote new products, but the mechanism of how information spreads through the network still needs to be further studied. It is very important to understand the communication mechanism of information, which can be applied in many fields, such as viral marketing, social behavior prediction, social recommendation, and community detection. These problems attract the attention of researchers from different fields, such as epidemiology, computer science, and sociology, who propose different information diffusion models to describe and simulate the process of information transmission, such as the independent cascade model, linear threshold model, and epidemic model. These models are mainly applied to influence evaluation, influence maximization, and information source detection. Most models recognize that

information is transmitted from a source node set and other nodes can only obtain information from the nodes that neighbor the source node set.

Social networking service providers, such as Twitter and Facebook, have grown rapidly in recent years, with increasing number of users sharing information with their friends. There are more than 2 million active users on Facebook every month from all over the world and about 5 billion new tweets on Twitter every day. Social network analysis can be divided into the following aspects [1, 2]: (a) studying the network structure and trends [3], (b) online learning of complex networks [4], (c) comparing different models, and (d) predicting node status [1, 5]. The focus of social influence study is to investigate neighbors and associations to predict the impact and influence of the occurrence of an action [2, 6].

Researchers have examined information transfers, including the analysis of relationships [7], social action tracking [1], and other types of relationship transfer [8]. The algorithm proposed in this paper constructs a matrix based on the information transmission between users to describe the complex correlation relations. By making certain changes to the matrix, information transmission paths can be identified and the tie strengths between nodes can be calculated. Due to the small computational difficulty involved in constructing a matrix, the algorithm proposed in this paper performs more efficiently than other algorithms.

3. Information Transmission Matrix

A piece of information is very valuable at one time, but after that it may be worthless. From the perspective of information transmission, the degree of interaction between users can be calculated. By analyzing information transmission paths, an information transmission tree can be generated to describe the information transmission rules and be used to analyze the dynamic changes of the correlations between users.

Definition 1. Let \mathbf{G} be a graph with n nodes and e edges. If

$$a_{ij} = \begin{cases} 1, & \text{if } e_j \text{ is associated with } v_i, \\ 0, & \text{if } e_j \text{ is not associated with } v_i, \end{cases} \quad (1)$$

then the $n \times e$ matrix composed of element a_{ij} ($1 \leq i \leq n$, $1 \leq j \leq e$) is constructed. $\mathbf{M} = (a_{ij})_{n \times e}$ is the complete incidence matrix of graph \mathbf{G} , namely, the information transmission matrix.

In this paper, information transmission data is used to construct and update \mathbf{M} . The construction process of \mathbf{M} is given below.

Figure 1 depicts the information transmission relationships between nodes. If there is an edge between nodes, then it means that information has been successfully passed

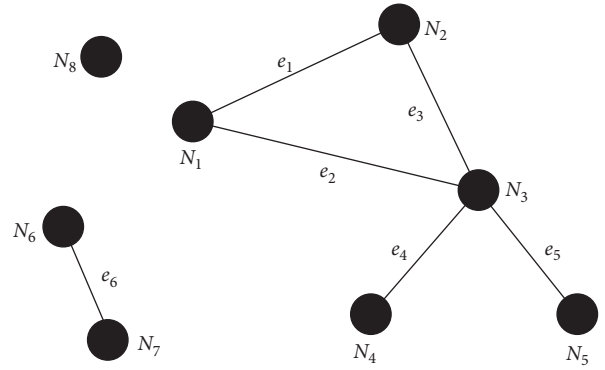


FIGURE 1: Information flow between nodes.

between them. Otherwise, no information has been passed. We construct matrix \mathbf{M} according to Figure 1, which describes the mapping relationship between nodes and edges. If there is an association between N_x and e_y , then $a_{xy} = 1$. Otherwise, $a_{xy} = 0$.

Because there is a large number of inactive nodes, most of the actions of the nodes on the Internet are from browsing information while actions such as commenting and forwarding are rare. Therefore, matrix \mathbf{M} is a sparse matrix. To reduce the negative impact of a large number of meaningless zeros in the matrix on subsequent calculations, further analysis of \mathbf{M} is required to delete redundant nodes. In Section 3.1, we describe a quick and effective way to remove redundant nodes.

3.1. Isolated Nodes

Definition 2. If the determinant of n th order matrix \mathbf{M} is not zero, that is, $|\mathbf{M}| \neq 0$, then \mathbf{M} is called a nonsingular matrix or full rank matrix. Otherwise, \mathbf{M} is called a singular matrix or reduced-rank matrix.

Definition 3. Nodes in graph \mathbf{G} are connected if and only if the rank of the complete incidence matrix is $n - 1$. The matrix whose order is $\min\{p, q\}$ is called a large submatrix of the $p \times q$ matrix.

By calculating whether $|\mathbf{M}|$ is 0, we can judge whether the nodes in \mathbf{G} are connected or not. A reduced matrix \mathbf{D} can be achieved by deleting redundant nodes in \mathbf{M} . \mathbf{D} is a full rank matrix, that is, $|\mathbf{D}| \neq 0$. At this time, \mathbf{D} is the maximum complete incidence matrix. That is, all the nodes in the new graph \mathbf{G} that are formed by \mathbf{D} are reachable, and there are no isolated nodes for information transmission.

Take matrix \mathbf{M} in Figure 2 as an example to illustrate the process of removing isolated nodes. The rank of \mathbf{M} is obtained by calculating the maximum number of linearly-independent crossings (that is, the maximum order of the nonzero submatrix):

$$\begin{aligned}
\mathbf{M} = & \begin{matrix} & e_1 & e_2 & e_3 & e_4 & e_5 & e_6 \\ \begin{matrix} N_1 \\ N_2 \\ N_3 \\ N_4 \\ N_5 \\ N_6 \\ N_7 \\ N_8 \end{matrix} & \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \end{matrix} & \begin{matrix} \text{addition and subtraction} \\ \text{of row vectors} \\ = \end{matrix} & \begin{matrix} & e_1 & e_2 & e_3 & e_4 & e_5 & e_6 \\ \begin{matrix} N_1 \\ N_2 \\ N_3 \\ N_4 \\ N_5 \\ N_6 \\ N_7 \\ N_8 \end{matrix} & \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \end{matrix} \quad (2)
\end{aligned}$$

$$\begin{aligned}
& \begin{matrix} \text{remove redundant nodes} \\ \rightarrow \end{matrix} & \begin{matrix} & e_1 & e_2 & e_3 & e_4 & e_5 & e_6 \\ \begin{matrix} N_1 \\ N_2 \\ N_3 \\ N_4 \\ N_5 \end{matrix} & \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix} \end{matrix} & \begin{matrix} \text{remove redundant edges} \\ \rightarrow \end{matrix} & \begin{matrix} & e_1 & e_2 & e_3 & e_4 & e_5 \\ \begin{matrix} N_1 \\ N_2 \\ N_3 \\ N_4 \\ N_5 \end{matrix} & \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{matrix}
\end{aligned}$$

According to the abovementioned calculation results, $R(\mathbf{M}) = 6$. This indicates the existence of isolated nodes in \mathbf{M} . It can be seen that rows N_7 and N_8 are $\vec{0}$, so N_7 and N_8 are redundant, isolated nodes. Because the original data in line N_6 and N_7 are same and N_7 was determined to be an isolated node to be deleted, N_6 is also an isolated node. In conclusion, N_6 , N_7 , and N_8 are isolated nodes. After removing redundant

nodes, it is necessary to determine whether there are redundant edges in the matrix. Because column e_6 is $\vec{0}$ after the redundant nodes are deleted, e_6 is a redundant edge that needs to be deleted.

Matrix \mathbf{D} is obtained after deleting the redundant nodes in \mathbf{M} . Next, whether the nodes in \mathbf{D} are connected must be calculated as follows:

$$\begin{aligned}
\mathbf{D} = & \begin{matrix} & e_1 & e_2 & e_3 & e_4 & e_5 \\ \begin{matrix} N_1 \\ N_2 \\ N_3 \\ N_4 \\ N_5 \end{matrix} & \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{matrix} & \begin{matrix} \text{addition and subtraction} \\ \text{of row vectors} \\ = \end{matrix} & \begin{matrix} & e_1 & e_2 & e_3 & e_4 & e_5 \\ \begin{matrix} N_1 \\ N_2 \\ N_3 \\ N_4 \\ N_5 \end{matrix} & \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{matrix} \quad (3)
\end{aligned}$$

The result is $R(\mathbf{D}) = 5$. That is, $|\mathbf{D}| \neq 0$, so \mathbf{D} is a full rank matrix. The conclusion is that all nodes in \mathbf{D} are connected. In other words, there are no isolated nodes of information transmission.

To discover all information transmission paths in \mathbf{M} , it is necessary to further determine which nodes can be tentatively considered to be redundant. The deleted redundant nodes are reconstituted into a new matrix \mathbf{M} and the abovementioned operations are repeated to obtain a matrix \mathbf{D} . Finally, multiple matrix \mathbf{D} s are obtained.

3.2. Information Transmission Path. To study the information transmission mechanism, it is necessary to identify all the information transmission paths from the information matrix. Therefore, further processing of the set of \mathbf{D} s is required.

Definition 4. Submatrix \mathbf{A} is obtained by removing one row from the complete incidence matrix \mathbf{D} . For \mathbf{A} to be nonsingular, the edges that correspond to the columns of \mathbf{A} must form a spanning tree of \mathbf{G} .

Definition 4 provides a method for calculating all spanning trees in the connected graph \mathbf{G} . By removing one row from matrix \mathbf{D} and then calculating all the maximized nonsingular submatrices of the newly-generated matrix \mathbf{D} , the edges that correspond to the columns of each nonsingular submatrix form a spanning tree of \mathbf{G} .

The matrix \mathbf{D} obtained in the previous section is taken as an example to illustrate the process of identifying information transmission paths according to Definition 4. Remove one row from \mathbf{D} (delete row 5 here) to get a matrix \mathbf{A} :

$$\begin{array}{c}
e_1 \quad e_2 \quad e_3 \quad e_4 \quad e_5 \quad e_6 \\
\begin{array}{l}
N_1 \\
N_2 \\
N_3 \\
N_4 \\
N_5 \\
N_6 \\
N_7 \\
N_8
\end{array}
\begin{pmatrix}
1 & 1 & 0 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 & 0 \\
0 & 1 & 1 & 1 & 1 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0
\end{pmatrix}
\end{array}$$

FIGURE 2: Matrix \mathbf{M} .

$$\mathbf{A} = \begin{array}{c}
e_1 \quad e_2 \quad e_3 \quad e_4 \quad e_5 \\
\begin{array}{l}
N_1 \\
N_2 \\
N_3 \\
N_4
\end{array}
\begin{pmatrix}
1 & 1 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 \\
0 & 1 & 1 & 1 & 1 \\
0 & 0 & 0 & 1 & 0
\end{pmatrix}
\end{array} \quad (4)$$

By calculating the rank of \mathbf{A} , we can get $R(\mathbf{A}) = 4$. This value indicates that the nodes in \mathbf{A} are connected. Although all nodes are connected to other nodes, there may be redundant edges. For example, the nodes N_1 , N_2 , and N_3 in Figure 1 have three edges, and these three nodes can be completely connected to each other by two of the edges. To remove redundant edges, we apply the following rules to the matrix:

Rule 1: the i th row of the matrix can be added and subtracted to the j th row

Rule 2: repeat the operation of Rule 1 until there is no operable item

Rule 3: the row vectors of the matrix are not interchangeable

Rule 4: the column vectors of the matrix are not interchangeable

Through the transformation of the matrix, the number of 1 in the matrix is reduced and the most concise matrix is finally obtained. At this time, the nodes in the matrix are connected by the minimum number of edges. Take matrix \mathbf{M} in Figure 2 as an example to illustrate the process of removing redundant edges. Matrix \mathbf{D} is obtained by deleting isolated nodes in \mathbf{M} , and matrix \mathbf{A} with 4 rows and 5 columns is obtained after deleting one row from \mathbf{D} . Matrix \mathbf{A} is a full rank matrix. To delete the redundant information transmission path, it is necessary to delete one column from \mathbf{A} to form multiple sets of different column combinations. The different combinations of columns are $\{(e_1, e_2, e_3, e_4); (e_1, e_2, e_3, e_5); (e_1, e_2, e_4, e_5); (e_1, e_3, e_4, e_5); \text{ and } (e_2, e_3, e_4, e_5)\}$. Then, perform row operations on each of the above-mentioned matrices according to the rules.

The first two rows in Table 1 describe the cases in which the constructed matrix does not meet the judgment condition for generating a maximum information transmission path. In the

first combination, edges $(e_1, e_2, e_3, \text{ and } e_4)$ are selected. It is found that in the matrix \mathbf{M}_{N_5, e_1} , \mathbf{M}_{N_5, e_2} , \mathbf{M}_{N_5, e_3} , and \mathbf{M}_{N_5, e_4} are all 0, so this path does not contain N_5 . That is, it is not a maximum information transmission path, so the combination of $(e_1, e_2, e_3, \text{ and } e_4)$ is deleted and the calculation is stopped. Similarly, in the second combination, \mathbf{M}_{N_4, e_1} , \mathbf{M}_{N_4, e_2} , \mathbf{M}_{N_4, e_3} , and \mathbf{M}_{N_4, e_4} is 0, so the calculation result obtained by this structure does not include N_4 , that is, it is not a maximum information transmission path. The ranks of the third, fourth, and fifth matrixes are all 4, so they are full rank matrices that satisfy the condition of generating maximum information transmission paths. The fourth column in rows 3, 4, and 5 in Table 1 show the row transformation process. Number 1 is the lowest in the transformed matrix, so the matrix does not have redundant edges. Column 5 shows the graph structure of the matrix obtained after eliminating the redundant edges. It can be seen from the graphs that the method proposed in this paper can be used to identify all maximum information transmission paths.

4. Tie Strength between Nodes

According to the characteristics of information transmission, it is reasonable to assume that there must be some association between the nodes in the same transmission path. Here, it is assumed that if information is transmitted frequently between two nodes, then the degree of intimacy between these two nodes is high. After a period of data accumulation, data about maximum information transmission paths is added to the correlation strength matrix (denoted as \mathbf{T}). Because the construction of \mathbf{T} is executed according to information transmission flows, matrix \mathbf{T} also keeps changing with the change of information transmission state. In matrix \mathbf{T} , $\mathbf{T}_{i, i}$ represents the occurrence number of node i in the process of information transmission and $\mathbf{T}_{i, j}$ represents the information transfer times between nodes i and j .

The following is the formula for calculating the weight of node i :

$$\text{node}_i = \frac{\mathbf{T}_{i,i}}{\sum_{x=1}^n \mathbf{T}_{x,x}} \quad (5)$$

The following is the formula for calculating the ties between node a and b :

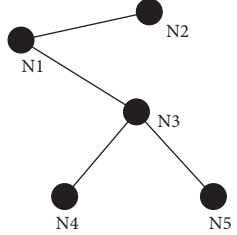
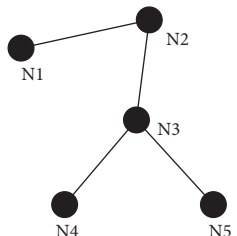
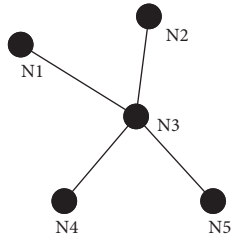
$$\text{edge}_{a,b} = \frac{\mathbf{T}_{a,b}}{\sum_{\substack{x=1 \\ x \neq a}}^n \mathbf{T}_{a,x} - \{\mathbf{T}_{a,a}\}} \quad (a > 1). \quad (6)$$

According to formulas (5) and (6), the degree of intimacy between different users is calculated. The specific algorithm is shown in Algorithm 1.

5. Experiments

Five datasets are used in this paper. For detailed information about the datasets, please refer to our paper [2] published earlier.

TABLE 1: The process of mining maximum information transmission paths.

Matrix	Result	Reasons for failure to meet the criteria/matrix transformation process/the structure corresponding to the calculation result
(e_1, e_2, e_3, e_4)	Not satisfied	Reason: $M_{N_5, e_1}, M_{N_5, e_2}, M_{N_5, e_3}$, and M_{N_5, e_4} in M are 0; therefore, N_5 is excluded, and it is not a maximum information transmission path
(e_1, e_2, e_3, e_5)	Not satisfied	Reason: $M_{N_4, e_1}, M_{N_4, e_2}, M_{N_4, e_3}$, and M_{N_4, e_4} in M are 0; therefore, N_4 is excluded, and it is not a maximum information transmission path
(e_1, e_2, e_4, e_5)	$R = 4$; full rank matrix	$\begin{matrix} & e_1 & e_2 & e_3 & e_4 & e_5 \\ N_1 & 1 & 1 & 0 & 0 & 0 \\ N_2 & 1 & 0 & 0 & 0 & 0 \\ N_3 & 0 & 1 & 1 & 0 & 0 \\ N_4 & 0 & 0 & 1 & 0 & 0 \end{matrix} \Rightarrow \begin{matrix} & e_1 & e_2 & e_3 & e_4 & e_5 \\ N_1 & 0 & 1 & 0 & 0 & 0 \\ N_2 & 1 & 0 & 0 & 0 & 0 \\ N_3 & 0 & 0 & 0 & 0 & 0 \\ N_4 & 0 & 0 & 1 & 0 & 0 \end{matrix}$ 
(e_1, e_3, e_4, e_5)	$R = 4$; full rank matrix	$\begin{matrix} & e_1 & e_2 & e_3 & e_4 & e_5 \\ N_1 & 1 & 0 & 0 & 0 & 0 \\ N_2 & 1 & 1 & 0 & 0 & 0 \\ N_3 & 0 & 1 & 1 & 1 & 0 \\ N_4 & 0 & 0 & 1 & 0 & 0 \end{matrix} \Rightarrow \begin{matrix} & e_1 & e_2 & e_3 & e_4 & e_5 \\ N_1 & 1 & 0 & 0 & 0 & 0 \\ N_2 & 0 & 1 & 0 & 0 & 0 \\ N_3 & 0 & 0 & 0 & 1 & 0 \\ N_4 & 0 & 0 & 1 & 0 & 0 \end{matrix}$ 
(e_2, e_3, e_4, e_5)	$R = 4$; full rank matrix	$\begin{matrix} & e_1 & e_2 & e_3 & e_4 & e_5 \\ N_1 & 1 & 0 & 0 & 0 & 0 \\ N_2 & 0 & 1 & 0 & 0 & 0 \\ N_3 & 1 & 1 & 1 & 1 & 0 \\ N_4 & 0 & 0 & 1 & 0 & 0 \end{matrix} \Rightarrow \begin{matrix} & e_1 & e_2 & e_3 & e_4 & e_5 \\ N_1 & 1 & 0 & 0 & 0 & 0 \\ N_2 & 0 & 1 & 0 & 0 & 0 \\ N_3 & 0 & 0 & 0 & 1 & 0 \\ N_4 & 0 & 0 & 1 & 0 & 0 \end{matrix}$ 

INPUT: matrix M
OUTPUT: intimacy correlation graph

- (1) generate matrix D according to matrix M
- (2) construct matrix A from D
- (3) FOREACH full rank matrix X IN matrix A
- (4) update matrix T according to X
- (5) ENDFOREACH
- (6) calculate *weights* and *ties* in T
- (7) construct a graph describing the degree of intimacy between nodes

ALGORITHM 1: TieCP.

- (1) Coauthor (<https://www.aminer.cn/data>): a dynamic coauthor network from ArnetMiner (<http://www.aminer.cn/>). We collected publications published from 2010 to 2016 by 100,000 authors.
- (2) DBLP (<http://www.vldb.org/dblp/>): the dataset is derived from a snapshot of the bibliography for 10 years, where each vertex represents a scientist and two vertices are connected if they work together on an article.
- (3) Twitter (<https://twitter.com>): we crawled the following links between 19,000,00 users from Twitter at 10 different time stamps from October to December 2017.
- (4) Weibo (<http://code.google.com/p/weibo4j/>): the most popular Chinese microblogging site. The data are crawled from March 8, 2014 when the crash of MH370 happened to April 8, 2014.

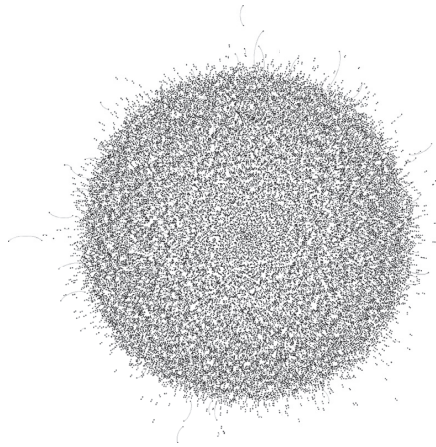


FIGURE 3: Unprocessed graph of Coauthor dataset.

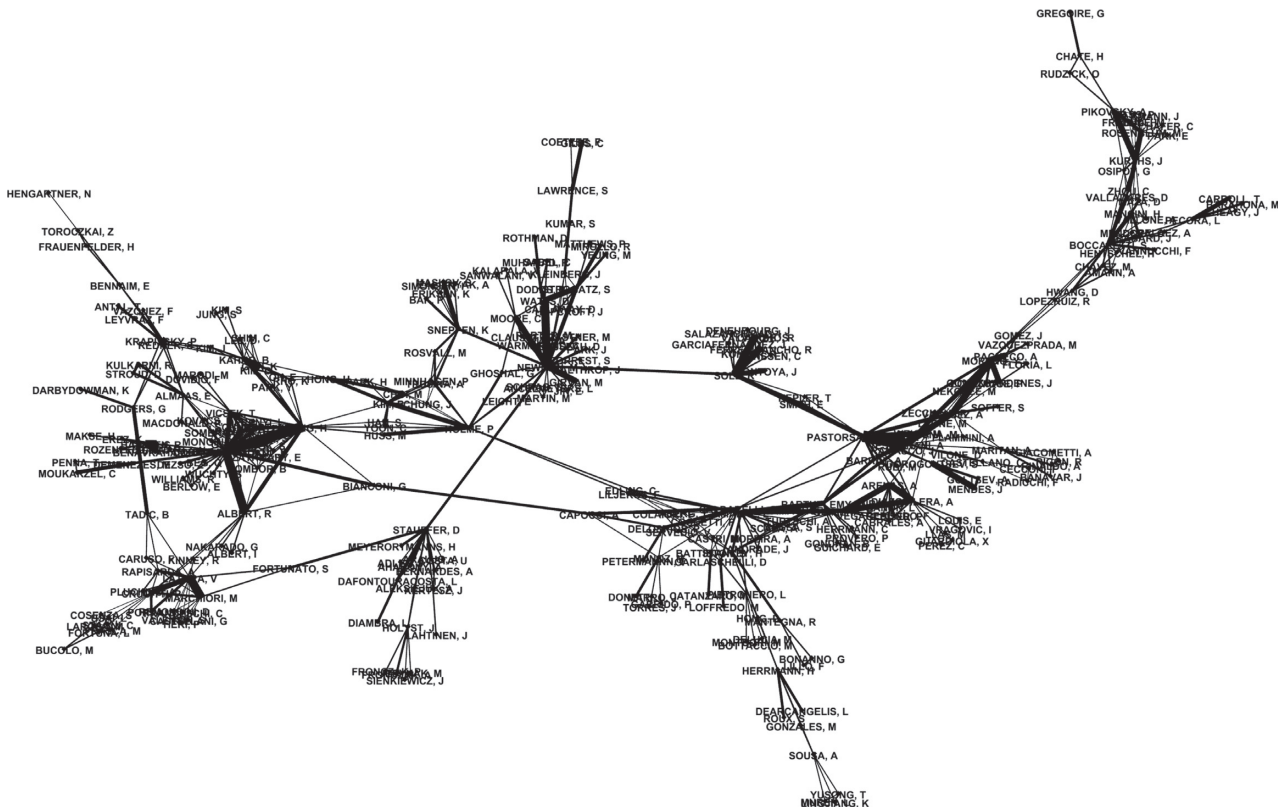


FIGURE 4: The maximum information transmission path in Coauthor dataset.

(5) Dolphin's Associations (<http://www-personal.umich.edu/~mejn/netdata/>): this dataset is an undirected social network of frequent associations between 62 dolphins, which has 62 nodes and 159 edges.

Three sets of baseline approaches are chosen for the experiments:

(1) PTPMF [9]: this method uses neighborhood overlap to approximate tie strength and extend the popular Bayesian Personalized Ranking (BPR) model to incorporate the distinction of strong and weak ties

(2) TrustMF [10]: this is a model-based method that adopts matrix factorization technique that maps users into low-dimensional latent feature spaces in terms of their trust relationship and aims to more accurately reflect the users' reciprocal influence on the formation of their own opinions and to learn better preferential patterns of users for high-quality recommendations.

(3) SBPR: this method presents a generic optimization criterion BPR-Opt for personalized ranking, that is,

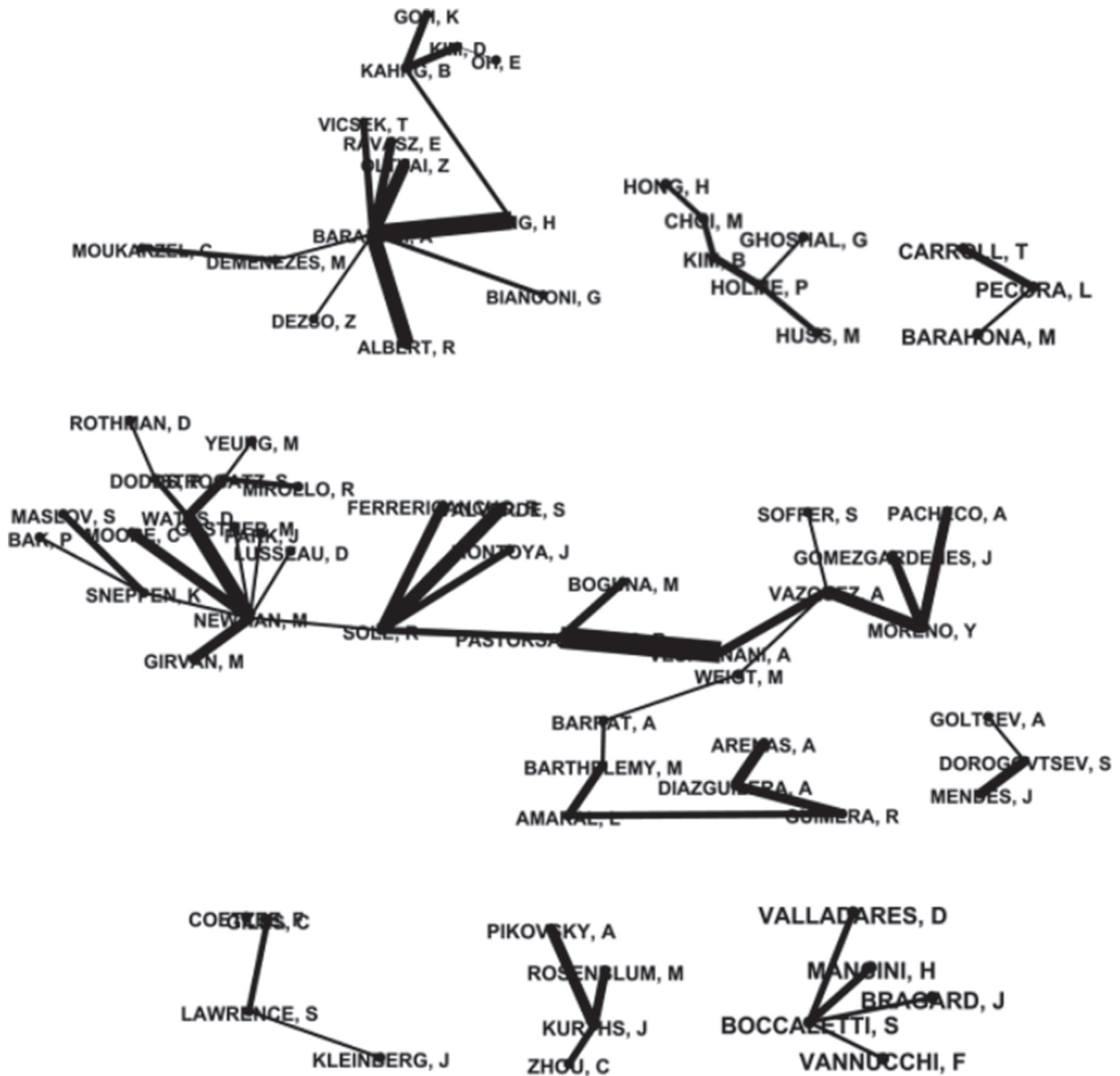


FIGURE 5: The maximum information transmission paths.

the maximum posterior estimator derived from a Bayesian analysis of the problem

Figure 3 shows the information transmission graph without data processing. It contains 38,501 nodes and 20,354 edges. If all nodes in the Coauthor dataset were displayed in Figure 3, then the picture would be black and the structure would not be visible. Therefore, only some of the nodes in the Coauthor dataset are shown in this figure. As can be seen in Figure 3, it is very difficult to process network data.

In the Coauthor dataset, the lengths of most information transmission paths are 2 or 3. Figure 4 shows the path with the maximum length in the Coauthor dataset.

By constructing a matrix according to the structure in Figure 4 and executing the algorithm proposed in this paper

on this matrix, it can be found that several groups of the largest and nonsegmented information transmission paths can be found, as shown in Figure 5. As can be seen from Figure 5, all the paths are loop free and achieve the maximum coverage of all nodes. Therefore, Figure 5 verifies the accuracy of the algorithm from the perspective of visualization.

Figure 6 depicts the degree of all nodes in the maximum information propagation path. It is found that the degree of most nodes is 1, the degree of a few nodes is greater than or equal to 2, and the highest degree value is 13. Figure 6 illustrates that the algorithm achieves the maximum removal of redundant edges.

The tie coefficients between different nodes are calculated according to information transmission paths. Figure 7

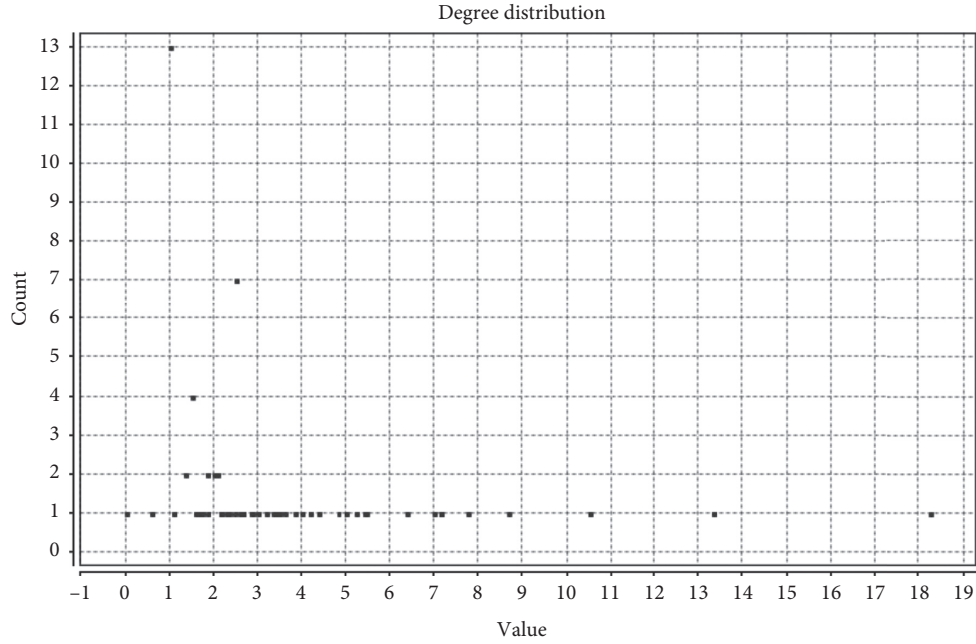


FIGURE 6: The degree of nodes in the maximum information transmission path.

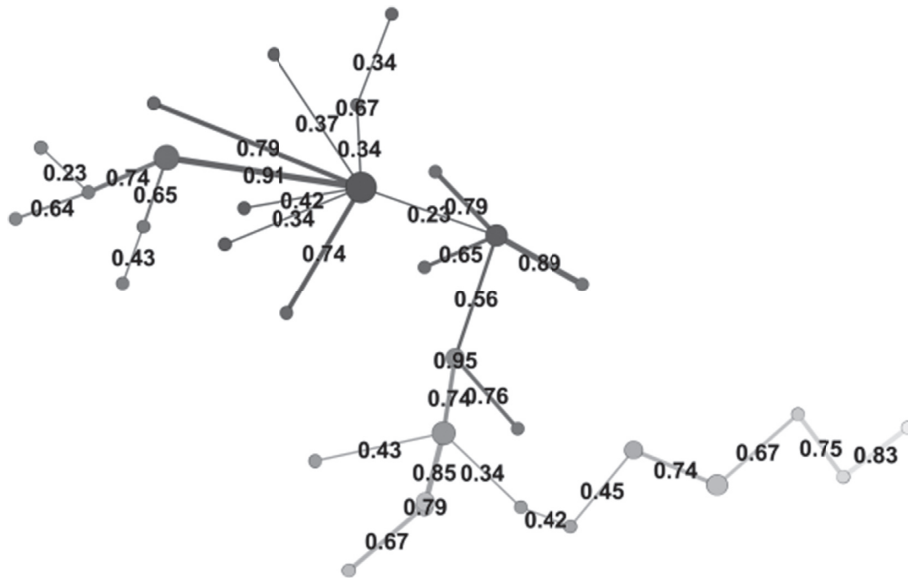


FIGURE 7: Tie coefficient.

shows the tie coefficient of nodes. In it, the darkness of the edges represents the correlation strengths between the node and the ego node. The darker the color is, the stronger the correlation is and vice versa. The number in the edge represents the tie strength between two connected nodes, which is the final result obtained by fusing multiple sets of maximum information transmission paths.

In order to analyze the experimental results, we use the following measurement parameters [10]: Precision calculated by $P = tp / (tp + fp)$, Recall by $R = tp / (tp + fn)$, and F1-score by $F = 2 \times P \times R / (P + R)$. tp is the number of

correctly identified examples, tn is the number of correctly identified nonrelated examples, fn is the number of not correctly identified the related examples, and fp is the number of not correctly identified nonrelated examples. Table 2 shows a comparison of the performances of different clustering algorithms on different datasets. It displays performance comparisons of SBPR, TrustMF, PTPMF, and TieCP using different datasets. According to Table 2, we can conclude that TieCP has the most stable execution effect and the best result regarding F-Score.

TABLE 2: The performance comparisons of different algorithms.

Dataset	SBPR			TrustMF			PTPMF			TieCP		
	<i>R</i>	<i>P</i>	<i>F</i>	<i>R</i>	<i>P</i>	<i>F</i>	<i>R</i>	<i>P</i>	<i>F</i>	<i>R</i>	<i>P</i>	<i>F</i>
DBLP	0.73	0.91	0.81	0.84	0.84	0.84	0.88	0.94	0.91	0.89	0.90	0.89
Twitter	0.63	0.98	0.77	0.76	0.87	0.81	0.80	0.94	0.86	0.83	0.92	0.87
Weibo	0.84	0.80	0.82	0.76	0.86	0.81	0.81	0.86	0.83	0.87	0.89	0.88
Coauthor	0.65	0.77	0.70	0.74	0.75	0.74	0.78	0.86	0.79	0.86	0.89	0.87

6. Conclusion

The algorithm proposed in this paper calculates the intimacy degrees between users according to the information transmission matrix. Compared to some mainstream methods, our method is simple and able to identify all the maximum information transmission paths. Beyond that our algorithm is relatively more stable when dealing with different kinds of data. Due to the small computational difficulty of constructing a matrix, the algorithm proposed in this paper performs more efficiently than other algorithms.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by Youth Program of the National Social Science Fund of China (Project name: Research on Online Behavior Pattern of Customers and Multidimensional Customer Insight Method under Big Data; Grant no. 19CGL024).

References

- [1] C. Tan, J. Tang, and J. Sun, "Social action tracking via noise tolerant time-varying factor graphs," in *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining-KDD'10*, pp. 1049–1058, ACM, Washington, DC, USA, July 2010.
- [2] L. Guo and B. Zhang, "Mining structural influence to analyze relationships in social network," *Physica A: Statistical Mechanics and Its Applications*, vol. 523, pp. 301–309, 2019.
- [3] M. Welling and Y. W. Teh, "Bayesian learning via stochastic gradient langevin dynamics," in *Proceedings of the 28th International Conference on Machine Learning*, pp. 681–688, Bellevue, WA, USA, 2011.
- [4] T. Shi and J. Zhu, "Online Bayesian passive-aggressive learning," *Journal of Machine Learning Research*, vol. 1, pp. 1–48, 2014.
- [5] L. Guo, W. Zuo, and T. Peng, "Inference network building and movements prediction based on analysis of induced dependencies," *IET Software*, vol. 11, no. 1, pp. 12–17, 2017.
- [6] Y. Li, Z. Liu, and H. Yu, "Advisor-advisee relationship identification based on maximum entropy model," *Acta Physica Sinica*, vol. 62, no. 16, 2013.
- [7] J. Chen, Y. Liu, G. Yang, and M. Zou, "Inferring tag co-occurrence relationship across heterogeneous social networks," *Applied Soft Computing*, vol. 66, pp. 512–524, 2018.
- [8] K. Tago and Q. Jin, "Influence analysis of emotional behaviors and user relationships based on twitter data," *Tsinghua Science and Technology*, vol. 23, no. 1, pp. 104–113, 2018.
- [9] X. Wang, S. Hoi, M. Ester, J. Bu, and C. Chen, "Learning personalized preference of strong and weak ties for social recommendation," in *Proceedings of the 26th International Conference on World Wide Web-WWW'17*, ACM, Rio de Janeiro, Brazil, May 2017.
- [10] E. Amigó, J. Gonzalo, J. Artilles, and F. Verdejo, "A comparison of extrinsic clustering evaluation metrics based on formal constraints," *Information Retrieval*, vol. 12, no. 4, pp. 461–486, 2009.

Research Article

Modeling Repayment Behavior of Consumer Loan in Portfolio across Business Cycle: A Triplet Markov Model Approach

Shou Chen  and Xiangqian Jiang 

School of Business and Administration, Hunan University, Changsha 410082, China

Correspondence should be addressed to Xiangqian Jiang; xq_jiang@hnu.edu.cn

Received 30 April 2019; Accepted 28 July 2019; Published 19 January 2020

Guest Editor: Benjamin M. Tabak

Copyright © 2020 Shou Chen and Xiangqian Jiang. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

With a view to develop a more realistic model for credit risk analysis in consumer loan, our paper addresses the problem of how to incorporate business cycles into a repayment behavior model of consumer loan in portfolio. A particular Triplet Markov Model (TMM) is presented and introduced to describe the dynamic repayment behavior of consumers. The particular TMM can simultaneously capture the phases of business cycles, transition of systematic credit risk of a loan portfolio, and Markov repayment behavior of consumers. The corresponding Markov chain Monte Carlo algorithms of the particular TMM are also developed for estimating the model parameters. We show how the transition of consumers' repayment states and systematic credit risk of a loan portfolio are affected by the phases of business cycles through simulations.

1. Introduction

The widespread use of consumer loan has not only alleviated the financial distress of consumers but also benefited credit companies. However, with the increasing consumer loan and credit limit, default risk has risen rapidly and even triggered systemic risk such as the subprime crisis in 2007. Dynamic management of consumer loan risk becomes more and more important in the credit business.

It is an important and practically relevant issue to assess and measure credit risk of a consumer loan in portfolio. In order to model the behavior of credit accounts, a stationary Markov chain approach was introduced [1]. Since then, the line of research has been developed [2–4]. In a previous study [5], a Markov chain model based on behavioral scores was developed for establishing the credit risk of consumer loans in portfolio. Moreover, a Markov model (MM) to measure transition of loan accounts is presented [6], in which the Markov transition probability is a function of account states, collection actions applied and borrower characteristics.

For linking the risk states of a debtor with loan default, Hidden Markov Model (HMM) or Double Chain Markov Model (DCMM), as important extension to the simple MM, has been popular in credit modeling in recent years. Giamperi

et al. [7] introduced an HMM to model the occurrence of defaults within a bond portfolio. They assume that the default probability of each bond in a portfolio depends on the hidden states of an HMM, which are interpreted as risk states of the bonds. In Banachewicz et al. [8], an HMM is developed to model and predict corporate default frequencies; the hidden states in their paper correspond to the industry credit cycles. In Fitzpatrick and Marchev [9], a multivariate DCMM is applied to credit rating dynamics of financial companies where a hidden process is regarded as a broader economy indicator affecting the transition of credit rating of companies. Quirini and Vannucci [10] link HMM with the analysis of credit risk of consumer loans in portfolio; the hidden states, which are regarded as credit market conditions, are endogenously extracted from repayment behavior records to measure loans' creditworthiness.

The hidden states of the model can be interpreted as systematic credit risk states of the loan portfolio when HMM or DCMM is used to measure the credit risk of a portfolio of consumer loan. Based on HMM or DCMM, one can link the systematic credit risk of a loan portfolio with the repayment behavior of consumers. However, the literature mentioned above leaves a fundamental question unanswered: What influences the transition of both repayment behavior of consumers and systematic credit risk of a loan portfolio?

Numerous studies have found a significant relationship between nonperforming loans (NPLs) and business cycles [11–13]. It is common to observe high NPL ratios during business cycle contraction, owing to the contraction in economic activity and the consequent decline in consumers' ability to repay their debts. In contrast, economic growth allows consumers to keep their finances buoyant and pay back their installments on time, resulting in low NPL ratios [14]. The evolution of consumers' repayment behavior and systematic credit risk of a loan portfolio are affected by business cycle, and different portfolios may be affected differently. A realistic model should account for this.

This paper aims to develop a credit risk model for consumer loan portfolio across business cycle. A particular TMM is presented and introduced to describe the dynamic Markov repayment behavior of consumers in the field of consumer loan. The particular TMM can test for and model two paths of impact of business cycle on the transition of consumers' repayment behavior. One of the paths is that the transition consumers' repayment behavior is directly affected by the business cycle. The other path is the business cycle affects the transition of systematic credit risk state of the loan portfolio and then affects the transition of consumers' repayment behavior. Simulation studies are provided to illustrate how the model functions.

TMM, also known as Triplet Markov Chains, was first proposed in Pieczynski et al. [15]. TMM consists of three processes: a hidden process X , an observed process Y and a third process Z . The model is called a TMM if there exists a stochastic process Z , where Z takes its value in a finite set, such that the triplet (X, Y, Z) is a Markov chain. TMM has been applied in the context of signal processing image processing and others [16–19]. To the authors' knowledge, there is little research that links credit risk measurement with TMM. In view of this, the present paper is the first study to introduce TMM into the credit risk analysis of consumer loan in portfolio.

By assessing the consumers' repayment behavior across business cycles, our proposed approach provides more information about the creditworthiness of consumer loans in portfolio, especially at extreme phases of the business cycle, which is very useful for credit risk management. Scenario analysis is more realistic and persuasive since our approach simultaneously captures the phases of a business cycle, transition of hidden risk states of loan portfolio, and Markov repayment behavior of consumers.

The remainder of this paper proceeds as follows. Section 2 describes the architecture of the particular TMM. Section 3 proposes a specific TMM for consumer loan. Section 4 gives simulation studies to illustrate how the model works in consumer loan. Section 5 provides a conclusion.

2. The Particular Triplet Markov Model Architecture

In this section, we present the structure of our particular TMM. For ease of comprehension and comparison, we first introduce two simpler models: HMM and DCMM.

HMM proposed in Baum and Petrie [20] has been widely used in various problems [21–23]. It consists of two stochastic processes X_t and Y_t , where X_t is a Markov chain and not directly

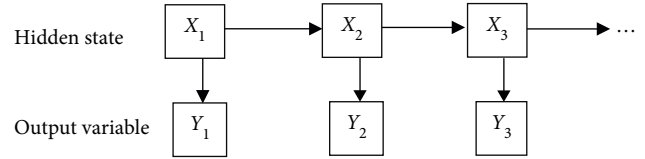


FIGURE 1: Bayesian network representation of HMM.

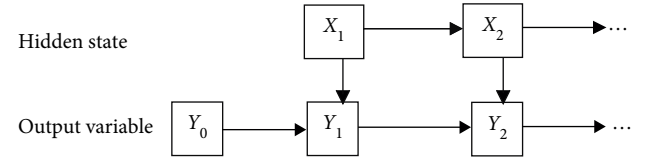


FIGURE 2: Bayesian network representation of DCMM.

visible (hidden), but the output of another variable Y_t whose distribution depends on the hidden process X_t is visible. A specific value of output variable is usually called observation. This is shown in Figure 1. A drawback of HMM is that the outputs are assumed to be conditionally independent. Considering that the conditional independence between outputs of an HMM is not always justified, the literature [24] proposes a DCMM as an extension of HMM to overcome this drawback. Figure 2 gives the Bayesian network representation of DCMM.

The TMM extends DCMM by adding a discrete value process Z_t , such that the triplet (X_t, Y_t, Z_t) is a Markov chain. The particular TMM proposed in our paper consists of three variables: an additional input variable Z_t , an underlying process X_t which is hidden from the output variable Y_t and depends on the input variable, the output variable Y_t which depends on both the input variable and the underlying process. The model is characterized by the following elements:

- (i) $S(Z) = \{1, 2, \dots, U\}$, the set of input variables.
- (ii) $S(X) = \{1, 2, \dots, M\}$, the set of hidden states.
- (iii) $S(Y) = \{1, 2, \dots, L\}$, the set of output variables.
- (iv) $\pi_1 = (\pi_{11}, \pi_{12}, \dots, \pi_{1M})$, the initial probability distribution of a hidden state.
- (v) $A = \{a_{ij}\} = \{P(Z_t = j | Z_{t-1} = i)\}$, $i, j \in S(Z)$, transition probabilities between successive inputs.
- (vi) $B^{(u)} = \{b_{ij}^{(u)}\} = \{P(X_t = j | X_{t-1} = i, Z_t = u)\}$, $i, j \in S(X)$, $u \in S(Z)$, transition probabilities between successive hidden states given a specific value of input variable.
- (vii) $C^{(u,s)} = \{c_{ij}^{(u,s)}\} = \{P(Y_t = j | Y_{t-1} = i, X_t = s, Z_t = u)\}$, $i, j \in S(Y)$, $s \in S(X)$, $u \in S(Z)$, transition probabilities between successive outputs given specific input value and hidden state.

Figure 3 shows the Bayesian network of our particular TMM. We can see that the transition process of X_t and Y_t depends on the specific value of input variable and are all non-homogeneous.

3. Triplet Markov Model for Consumer Loan

In this section, we set up the particular TMM for consumer loan. As described previously, the model consists of three Markov chains. We next construct each of them in the context of consumer loan.

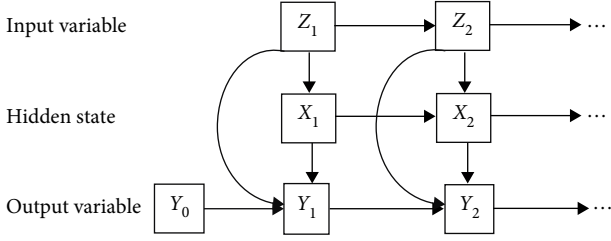


FIGURE 3: Bayesian network representation of our particular TMM.

We take the business cycle as the input variable of the particular TMM. In facing a real case, practitioners can use the business cycle chronology compiled by the National Bureau of Economic Research (NBER) as input variable. Scholars have divided the business cycle into separate phases or regimes, in particular treating expansions separately from contractions [25, 26]. Following this tradition, we present the business cycle in using a very simple two-states model where $S(Z) = \{1, 2\}$ and “1” for expansion, “2” for contraction. The process of two phases of business cycle is governed by the following homogenous Markov transition matrix:

$$A = \{a_{ij}\} = \{P(Z_t = j | Z_{t-1} = i)\}, \quad \text{for } i, j \in \{1, 2\}, \quad (1)$$

In our paper, the hidden states of TMM are regard as systematic credit risk states of a loan portfolio. We assume the number of hidden risk states is m . Taking the input variable Z_t into account, the process of hidden risk states evolves according to a non-homogeneous Markov process:

$$X_t | X_{t-1} \sim \text{Markov}(B^{(u)}, \pi_1), \quad (2)$$

where $B^{(u)} = \{b_{ij}^{(u)}\}$ is the one-step transition probability matrix of the chain with $Z_t = u$, i.e., $b_{ij}^{(u)} = P(X_t = j | X_{t-1} = i, Z_t = u)$ for $i, j \in \{1, 2, \dots, m\}$, and π_1 is the probability distribution at $t = 1$. From (2), it can be seen that the transition of consumer’ hidden risk states switches between two Markov regimes according to the input value of the business cycle $u \in \{1, 2\}$.

It is worth emphasizing that the particular TMM in our paper has only one common hidden sequence, and multiple output sequences are all driven by this common hidden sequence. That is to say, the hidden systemic credit risk is common to all consumers within a loan portfolio, even if each consumer’s repayment behavior is different. This design of hidden sequence in our particular TMM is consistent with that in DCMM represented in Fitzpatrick and Marchev [9].

We now consider the repayment behavior of consumers affected by both the business cycle and the systemic credit risk of the loan portfolio. For a consumer k , let $Y_{k,t}$ denote his/her random repayment state expressed by the number of unpaid installments at time t . In practice, a default state is usually considered and regarded as an absorbing state which means any consumer who has reached this state can never return back. Here, we adopt the assumption in [10, 27] a consumer is assumed to become a defaulter when three accumulative installments are unpaid. Hence, we have:

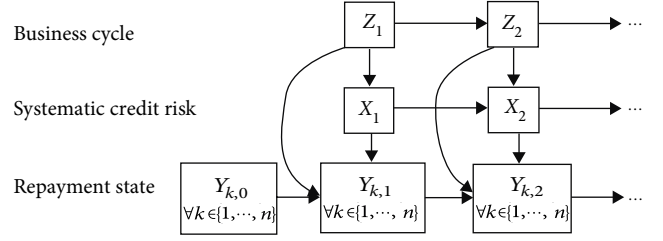


FIGURE 4: Bayesian network representation of TMM for consumer loan.

$$Y_{k,t} \in \{0, 1, 2, 3(\text{default})\}. \quad (3)$$

The process of each consumer’s repayment behavior is a nonhomogeneous Markov chain with the following transition matrices:

$$C^{(u,s)} = \{c_{ij}^{(u,s)}\} = \{P(Y_{k,t} = j | Y_{k,t-1} = i, X_t = s, Z_t = u)\}, \quad (4)$$

for $i, j \in \{0, 1, 2, 3\}$, $s \in \{1, 2, \dots, m\}$, $u \in \{1, 2\}$, $k \in \{1, 2, \dots, n\}$, where n is the number of consumers in portfolio.

Form (3) shows that affected by both business cycle and hidden systemic credit risk, the transition of consumers’ repayment states switches among $2m$ Markov regimes.

The relationship among the business cycles, systematic credit risk of a loan portfolio, and consumers’ repayment state is plotted in Figure 4. We can see that the particular TMM can test for and model two paths of impact of business cycle on the transition of consumers’ repayment state. One of the paths is that the transition of consumers’ repayment state is directly affected by the business cycle. The other path is indirect impact, where the business cycle affects the transition of systematic credit risk state of the loan portfolio, and then affects the transition of consumers’ repayment state.

The impacts of business cycle on consumers’ repayment behavior are reflected in the difference of the transition probabilities in the transition matrix $B^{(u)}$ and $C^{(u,s)}$, respectively. In practical terms, the probability of occurrence of a low/high systematic credit risk state increases when the external business cycle is in the expansion/contraction stage. At the same time, the repayment behavior of consumers in loan portfolio is more likely to ameliorate/deteriorate because of the expansive/contractive business cycle and low/high systemic credit risk.

The particular TMM provides an alternative approach to measure the credit risk of consumer loans in portfolio across business cycles. In practice, credit managers can focus on a portfolio of borrowers from a particular type or market, for example, borrowers from the energy sector. The systematic credit risk of the loan portfolio can then be seen as a proxy variable of credit market condition of the energy sector. Obviously, the credit market condition of the energy sector is affected by external business cycles. TMM provides us an approach to measure the performance of loans in portfolio under different business cycles and credit market conditions of the energy sector, while also considering the impact of a business cycle on the credit market condition of the energy sector.

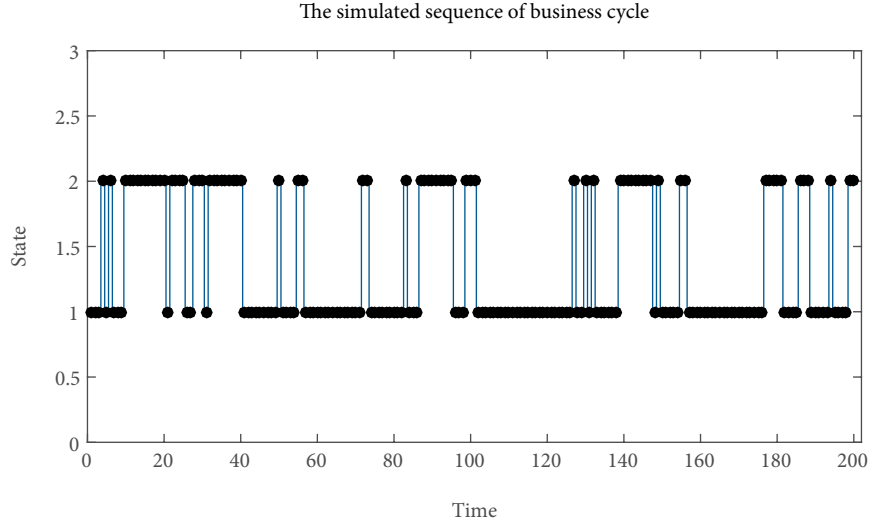


FIGURE 5: Simulated sequence of business cycle.

4. Simulation Studies

4.1. Generation of Simulated Data. For the sake of simplicity and without loss of generality, we consider there are only two hidden risk states since it is a well-accepted theory that risk states fluctuate between two Markov regimes: normal risk and enhanced risk. We denote: “N” for normal risk and “E” for enhanced risk.

Considering the empirical characteristics of transition matrices of credit card loan accounts in Leow and Crook [27], the parameter values in our paper are given as follows:

The transition matrix of business cycle is

$$A = \begin{matrix} & \begin{matrix} 1 & 2 \end{matrix} \\ \begin{matrix} 1 \\ 2 \end{matrix} & \begin{pmatrix} 0.8 & 0.2 \\ 0.3 & 0.7 \end{pmatrix} \end{matrix}, \quad (5)$$

transition matrices of systematic credit risk states are

$$B^{(1)} = \begin{matrix} & \begin{matrix} N & E \end{matrix} \\ \begin{matrix} N \\ E \end{matrix} & \begin{pmatrix} 0.9 & 0.1 \\ 0.5 & 0.5 \end{pmatrix} \end{matrix}, \quad B^{(2)} = \begin{matrix} & \begin{matrix} N & E \end{matrix} \\ \begin{matrix} N \\ E \end{matrix} & \begin{pmatrix} 0.6 & 0.4 \\ 0.2 & 0.8 \end{pmatrix} \end{matrix} \quad (6)$$

transition matrices of consumers' repayment states are

$$C^{(1,N)} = \begin{matrix} & \begin{matrix} 0 & 1 & 2 & 3 \end{matrix} \\ \begin{matrix} 0 \\ 1 \\ 2 \\ 3 \end{matrix} & \begin{pmatrix} 0.98 & 0.02 & 0 & 0 \\ 0.9 & 0.08 & 0.02 & 0 \\ 0.8 & 0.08 & 0.02 & 0.10 \\ 0 & 0 & 0 & 1 \end{pmatrix} \end{matrix}, \quad C^{(2,N)} = \begin{matrix} & \begin{matrix} 0 & 1 & 2 & 3 \end{matrix} \\ \begin{matrix} 0 \\ 1 \\ 2 \\ 3 \end{matrix} & \begin{pmatrix} 0.95 & 0.05 & 0 & 0 \\ 0.84 & 0.11 & 0.05 & 0 \\ 0.71 & 0.11 & 0.05 & 0.13 \\ 0 & 0 & 0 & 1 \end{pmatrix} \end{matrix}, \quad (7)$$

$$C^{(1,E)} = \begin{matrix} & \begin{matrix} 0 & 1 & 2 & 3 \end{matrix} \\ \begin{matrix} 0 \\ 1 \\ 2 \\ 3 \end{matrix} & \begin{pmatrix} 0.92 & 0.08 & 0 & 0 \\ 0.78 & 0.14 & 0.08 & 0 \\ 0.62 & 0.14 & 0.08 & 0.16 \\ 0 & 0 & 0 & 1 \end{pmatrix} \end{matrix}, \quad C^{(2,E)} = \begin{matrix} & \begin{matrix} 0 & 1 & 2 & 3 \end{matrix} \\ \begin{matrix} 0 \\ 1 \\ 2 \\ 3 \end{matrix} & \begin{pmatrix} 0.86 & 0.14 & 0 & 0 \\ 0.66 & 0.20 & 0.14 & 0 \\ 0.44 & 0.20 & 0.14 & 0.22 \\ 0 & 0 & 0 & 1 \end{pmatrix} \end{matrix},$$

We can see that the systematic credit risk states transfer according to whether $B^{(1)}$ or $B^{(2)}$ depend on the value of the input business cycle. The systematic credit risk state is more likely to remain or move to normal risk state when the phase of the external business cycle is expansion. On the contrary, the probability of turning into the enhance risk state increases

during the contractive phase of the business cycle. The elements in matrices $C^{(u,s)}$, $u \in \{1, 2\}$, $s \in \{N, E\}$ show that the probability of recovering overdue installments is larger than the probability of paying nothing, and the assumption is diminished if systematic credit risk state is “E” or the phase of business cycle is “2.”

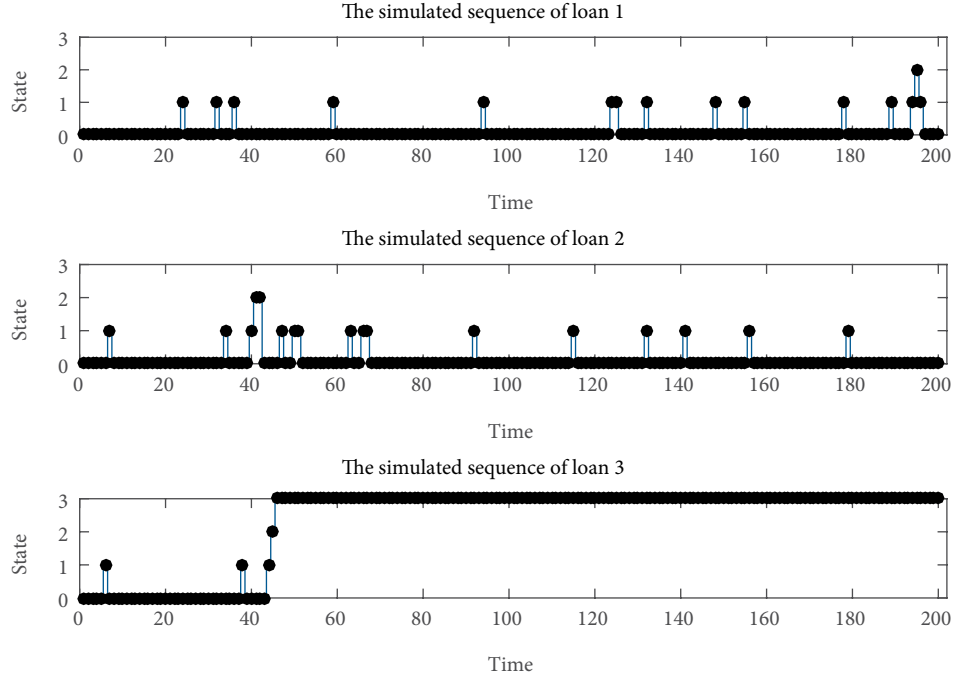


FIGURE 6: Simulated repayment state sequences of three randomly selected loans.

TABLE 1: Posterior inference for parameters of our specific TMM.

	$u = 1$	$u = 2$
$\bar{B}^{(u)}$	$\begin{pmatrix} 0.9205 & 0.0795 \\ 0.4376 & 0.5624 \end{pmatrix}$	$\begin{pmatrix} 0.5933 & 0.4067 \\ 0.1689 & 0.8311 \end{pmatrix}$
$\bar{C}^{(u,N)}$	$\begin{pmatrix} 0.9800 & 0.0200 & 0.0000 & 0.0000 \\ 0.8996 & 0.843 & 0.0161 & 0.0000 \\ 0.7882 & 0.0936 & 0.0249 & 0.0933 \\ 0.0000 & 0.0000 & 0.0000 & 1.0000 \end{pmatrix}$	$\begin{pmatrix} 0.9515 & 0.0485 & 0.0000 & 0.0000 \\ 0.8340 & 0.1182 & 0.0478 & 0.0000 \\ 0.6255 & 0.1331 & 0.0645 & 0.1769 \\ 0.0001 & 0.0001 & 0.0001 & 0.9997 \end{pmatrix}$
$\bar{C}^{(u,E)}$	$\begin{pmatrix} 0.9190 & 0.0810 & 0.0000 & 0.0000 \\ 0.7857 & 0.1376 & 0.0767 & 0.0000 \\ 0.6357 & 0.1567 & 0.0735 & 0.1341 \\ 0.0000 & 0.0000 & 0.0000 & 1.0000 \end{pmatrix}$	$\begin{pmatrix} 0.8594 & 0.1406 & 0.0000 & 0.0000 \\ 0.6674 & 0.1960 & 0.1366 & 0.0000 \\ 0.4293 & 0.2024 & 0.1491 & 0.2192 \\ 0.0000 & 0.0000 & 0.0000 & 1.0000 \end{pmatrix}$

We use the TMM with the given parameter values to generate our simulated data. In addition to containing an input sequence of business cycle, the simulated data also contain repayment states of $N = 1000$ consumers, a total of $T = 200$ time intervals, that is to say, there are 1000 sequences, each containing 200 repayment histories. Figures 5 and 6 show the simulated business cycle and three randomly selected loans, respectively.

4.2. Estimation of Model Parameters. Our paper focuses on estimating the value of transition matrices $B^{(u)}$, $C^{(u,s)}$ from the simulated data set. From a Bayesian perspective, we use Markov chain Monte Carlo (MCMC) algorithms to estimate the parameters of the particular TMM. Since the classic Expectation-Maximization algorithm is sensitive to the starting values and is easy to fall into a local optimal solution. However, MCMC makes posterior risk minimization, and make full use of the experience, history information, and other

TABLE 2: SSE of the full sample and three subsamples.

	SSE1	SSE2	Total SSE
$N = 1000, T = 200$	0.0106	0.0122	0.0228
$N = 1000, T = 190$	0.0157	0.0121	0.02788
$N = 950, T = 200$	0.0107	0.0155	0.0262
$N = 950, T = 190$	0.0157	0.0157	0.0314

Note: SSE is the sum of squared differences of true parameter values used to generate the simulated data and their estimates.

information of samples. Employing the great computational power of MCMC, the model parameters can be quickly extracted. The MCMC algorithms of our particular TMM are similar to the MCMC algorithms of DCMM [9]. For the sake of brevity, all MCMC algorithms are consigned to Appendix A.

We run our MCMC algorithms of TMM 10000 iterations on software R with the first 5000 iterations being discarded as burn-in. The Dirichlet prior parameters in MCMC algorithms

TABLE 3: SSE of four cases with different parameters.

	SSE1	SSE2	Total SSE
Case 1	0.0165	0.0108	0.0273
Case 2	0.0226	0.0062	0.0288
Case 3	0.0231	0.0100	0.0331
Case 4	0.0145	0.0149	0.0294

TABLE 4: Posterior inference for parameters of DCMM

	Estimated parameter values
\bar{B}	$\begin{pmatrix} 0.8247 & 0.1753 \\ 0.3046 & 0.6954 \end{pmatrix}$
$\bar{C}^{(N)}$	$\begin{pmatrix} 0.9736 & 0.0264 & 0.0000 & 0.0000 \\ 0.8772 & 0.0959 & 0.0269 & 0.0000 \\ 0.7179 & 0.1106 & 0.0418 & 0.1297 \\ 0.0000 & 0.0000 & 0.0000 & 1.0000 \end{pmatrix}$
$\bar{C}^{(E)}$	$\begin{pmatrix} 0.8826 & 0.1174 & 0.0000 & 0.0000 \\ 0.7103 & 0.1747 & 0.1150 & 0.0000 \\ 0.5006 & 0.1868 & 0.1226 & 0.1900 \\ 0.0000 & 0.0000 & 0.0000 & 1.0000 \end{pmatrix}$

are all set to 0.1. The estimated values of parameters are obtained from the posterior means of the last 5000 iterations and the results are given in Table 1.

The estimated matrices in Table 1 are close to their true values. The estimated matrices $\bar{B}^{(u)}$, $u \in \{1, 2\}$ give the transition probabilities between the two systematic credit risk states at different phases of the business cycle. Furthermore, by comparing the matrices $\bar{C}^{(u,s)}$, $u \in \{1, 2\}$, $s \in \{N, E\}$, the different transition probabilities among repayment states of consumers in different systematic credit risk states and business cycles can be obtained. The results will assist practitioners for assessing, managing, and monitoring credit risk in their loan portfolio.

As a robustness check, Table 2 lists the sum of squared errors (SSE) of full sample and three subsamples. The SSE of two transition matrices of systematic credit risk states and four transition matrices of consumers' repayment states are denoted by SSE1 and SSE2 respectively.

$$\begin{aligned}
C^{(1,N)} &= \begin{matrix} & \begin{matrix} 0 & 1 & 2 & 3 \end{matrix} \\ \begin{matrix} 0 \\ 1 \\ 2 \\ 3 \end{matrix} & \begin{pmatrix} 0.98 & 0.02 & 0 & 0 \\ 0.9 & 0.08 & 0.02 & 0 \\ 0.8 & 0.08 & 0.02 & 0.10 \\ 0 & 0 & 0 & 1 \end{pmatrix} \end{matrix}, \quad C^{(2,N)} = \begin{matrix} & \begin{matrix} 0 & 1 & 2 & 3 \end{matrix} \\ \begin{matrix} 0 \\ 1 \\ 2 \\ 3 \end{matrix} & \begin{pmatrix} 0.96 & 0.04 & 0 & 0 \\ 0.86 & 0.10 & 0.04 & 0 \\ 0.74 & 0.10 & 0.04 & 0.12 \\ 0 & 0 & 0 & 1 \end{pmatrix} \end{matrix}, \\
C^{(1,E)} &= \begin{matrix} & \begin{matrix} 0 & 1 & 2 & 3 \end{matrix} \\ \begin{matrix} 0 \\ 1 \\ 2 \\ 3 \end{matrix} & \begin{pmatrix} 0.94 & 0.06 & 0 & 0 \\ 0.82 & 0.12 & 0.06 & 0 \\ 0.68 & 0.12 & 0.06 & 0.14 \\ 0 & 0 & 0 & 1 \end{pmatrix} \end{matrix}, \quad C^{(2,E)} = \begin{matrix} & \begin{matrix} 0 & 1 & 2 & 3 \end{matrix} \\ \begin{matrix} 0 \\ 1 \\ 2 \\ 3 \end{matrix} & \begin{pmatrix} 0.90 & 0.10 & 0 & 0 \\ 0.74 & 0.16 & 0.10 & 0 \\ 0.56 & 0.16 & 0.10 & 0.18 \\ 0 & 0 & 0 & 1 \end{pmatrix} \end{matrix} \tag{10}
\end{aligned}$$

TABLE 5: Posterior inference for parameters of MM.

	Estimated parameter values
\bar{C}	$\begin{pmatrix} 0.9412 & 0.0588 & 0.0000 & 0.0000 \\ 0.7813 & 0.1412 & 0.0775 & 0.0000 \\ 0.5800 & 0.1589 & 0.0931 & 0.1680 \\ 0.0000 & 0.0000 & 0.0000 & 1.0000 \end{pmatrix}$

The results of SSE show that a reduction in the number of consumers and the length of time interval has a negative impact on the estimation accuracy of transition matrices. Whereas, on average, the estimated deviations of each element in the transition matrix $\bar{B}^{(u)}$ and $\bar{C}^{(u,s)}$ are only $\sqrt{0.0157/8} = 0.0443$ and $\sqrt{0.0157/64} = 0.0157$, respectively. These small deviations are not enough to have a significant impact on the estimation accuracy of transition matrices. That is to say, a 5% reduction in the number of borrowers and a 5% reduction in the length of time interval have no significant effect on the robustness of TMM.

We now test the robustness of our model with different parameter values. The following four cases are taken into consideration.

Case 1. The transition matrix of business cycle is reset to

$$A = \begin{matrix} & \begin{matrix} 1 & 2 \end{matrix} \\ \begin{matrix} 1 \\ 2 \end{matrix} & \begin{pmatrix} 0.8 & 0.2 \\ 0.2 & 0.8 \end{pmatrix} \end{matrix}, \tag{8}$$

Case 2. Transition matrices of systematic credit risk states are changed to

$$B^{(1)} = \begin{matrix} & \begin{matrix} N & E \end{matrix} \\ \begin{matrix} N \\ E \end{matrix} & \begin{pmatrix} 0.8 & 0.2 \\ 0.4 & 0.6 \end{pmatrix} \end{matrix}, \quad B^{(2)} = \begin{matrix} & \begin{matrix} N & E \end{matrix} \\ \begin{matrix} N \\ E \end{matrix} & \begin{pmatrix} 0.7 & 0.3 \\ 0.3 & 0.7 \end{pmatrix} \end{matrix}, \tag{9}$$

Case 3. Transition matrices of consumers' repayment states are given as

TABLE 6: Predicted values of expected default rate and deviation rate (in parentheses) after 12 time periods.

	Scenario 1			Scenario 2			Scenario 3		
	TMM	DCMM	MM	TMM	DCMM	MM	TMM	DCMM	MM
Repayment state 0	0.0003 (24.25%)	0.0010 (137.29%)	0.0091 (2003.89%)	0.0509 (18.4%)	0.0304 (41.40%)	0.0091 (82.51%)	0.0052 (4.94%)	0.0085 (57.37%)	0.0091 (67.11%)
Repayment state 1	0.0020 (24.19%)	0.0049 (89.79%)	0.0247 (850.37%)	0.0871 (1.96%)	0.0564 (36.46%)	0.0247 (72.13%)	0.0166 (8.71%)	0.0312 (71.73%)	0.0247 (36.16%)
Repayment state 2	0.0961 (6.30%)	0.1366 (33.18%)	0.1946 (89.69%)	0.2996 (0.35%)	0.2432 (18.54%)	0.1946 (34.83%)	0.1498 (15.56%)	0.2216 (24.89%)	0.1946 (9.65%)

Note: The rates of deviation are obtained from the formula $|\bar{\theta} - \theta|/\theta$, where $\bar{\theta}$ is the expected default rate based on the estimated parameter values of each model, θ is the expected default rate based on the true parameter values of TMM in Section 4.1.

Case 4. The parameter values are simultaneously changed to the values given in the above three cases.

We regenerate the simulated data according to the parameters given in above four cases. Again, using MCMC algorithms in Appendix A, we can obtain the estimated parameter values. SSE1, SSE2, and total SSE can then be obtained by comparing them with the true parameter values. The results are presented in Table 3. We can see that the SSE1, SSE2, and total SSE with different parameters are very close to those previously listed in Table 2. This shows that the accuracy of estimation does not seem to decrease with the above parameter values and TMM is robust to these parameter values.

4.3. Comparative Analysis. Now we study other cases for comparison. We estimate the parameter values of DCMM and the MM based on the same simulated data. The MCMC algorithms of DCMM are presented in Fitzpatrick and Marchev [9]. The algorithms of MM can be found in the R package ‘‘MCTM’’ (<https://cran.r-project.org/web/packages/MCTM/index.html>). All codes were run on software R. The estimated parameter values of DCMM and MM are obtained and listed in Tables 4 and 5, respectively.

By comparing the results of Tables 1, 4 and 5, we can see that the estimated matrices $\widetilde{C}^{(s)}$, $s \in (N, E)$ in DCMM can be regarded as a complex combination of $\widetilde{C}^{(1,s)}$ and $\widetilde{C}^{(2,s)}$ in TMM. Further, the estimated matrix \widetilde{C} in MM can be obtained by a complex combination of $\widetilde{C}^{(N)}$ and $\widetilde{C}^{(E)}$ in DCMM. Both DCMM and MM cannot explicitly analyze the influence of business cycle on consumers’ repayment behavior. Whereas, the transition process of consumers’ repayment states across business cycle can be modeled and estimated by TMM. Based on TMM, the analysis of consumer repayment behavior in loan portfolio is more meticulous and persuasive.

Once the estimated parameter values of TMM, DCMM, and MM are obtained, the prediction of future default probability of consumers can be made. We predict the expected default probability of current non-defaulting consumers after 12 time periods based on estimated parameter values. For the sake of simplicity, but without losing generality, we focus on the following three scenarios.

Scenarios 1. All phases of a business cycle are expansion and all systematic credit risk states are normal risk over the next 12 time periods.

Scenarios 2. All phases of a business cycle are contraction and all systematic credit risk states are enhanced risk over the next 12 time periods.

Scenarios 3. The current phase of the business cycle is contraction, but the current systematic credit risk state is normal risk.

Obviously, Scenarios 1 and 2 are the best and worst case of TMM respectively. In Scenario 3, the occurrences of business cycle and systematic risk state in the next 12 time periods are stochastic, and are governed by the estimated transition matrices. Before the process of prediction, we adjust the last row of the estimated matrices $\widetilde{C}^{(u,s)}$, $u \in \{1, 2\}$, $s \in \{N, E\}$ to $(0, 0, 0, 1)$ since the repayment state ‘‘3’’ is an absorbing state. The predicted

values of expected default rate (deviation rates are listed in parentheses) of current nondefaulting consumers after the 12 time periods of TMM, DCMM, and MM are given in Table 6.

The results in Table 6 show that the probability of becoming a defaulter increases with the increase in the number of unpaid instalments. We can also see that the deviation rate of TMM is the smallest among the three models, especially in Scenarios 1 and 2. This suggests that the TMM proposed in this paper can be used to accurately predict and assess the credit risk of consumers in loan portfolio, especially when business cycle stays at one stage for a long time.

Notice that the transition matrices, which characterize the expected changes in credit quality of consumers, are cardinal inputs to portfolio risk assessment. The particular TMM in this paper can integrate business cycle into the analysis of consumers’ repayment behavior and reveal more information about transition matrices of a loan portfolio. We believe our analysis provides a useful method to stress testing a loan portfolio and could assist practitioners in managing credit risk in loan portfolio.

5. Conclusions

The proposed TMM in this paper extends the credit risk measurement of consumer loan portfolio in HMM and DCMM by taking the impact of a business cycle into consideration. The structure of TMM can incorporate two paths of impact of a business cycle on the transition of consumers’ repayment behavior. One is the transition consumers’ repayment behavior is directly affected by the business cycle. The other is the business cycle affects the transition of systematic credit risk of the loan portfolio and then affects the transition of consumers’ repayment behavior. It is the first time that a TMM is applied to consumer loan. We also develop the corresponding MCMC algorithms of the particular TMM for estimating model parameters. Numerical examples illustrated that the proposed TMM is more accurate in assessing and predicting the credit risk of consumers in loan portfolio than DCMM and MM, especially when the business cycle is stuck in one phase for a long time.

Our research still has some limitations, which may be addressed in future research. The simulation studies confirm that the model fitting process can retrieve the original parameters closely. However, the simulated data do not assess potential practical limits and difficulties under real conditions. Concerning the using of TMM in real credit risk situations, an application to real data would substantially improve the quality of our paper. The early payoff case and recovery from default state can also be incorporated according to the real situation. A more complicated data structure such as default correlations among consumers and the missing data case can be further extended.

Appendix

To simplify the notation, for $k \in \{1, \dots, n\}$, $0 \leq t \leq T$, we define:

$$Y(t) := \cup_k \{Y_{k,t}\}, Y^{(t)} := \cup_k \{(Y_{k,0}, \dots, Y_{k,t})\}, \quad (A.1)$$

$$\begin{aligned} Z_t^T &:= (Z_t, \dots, Z_T), X_t^T := (X_t, \dots, X_T), \\ Y_t^T &:= \cup_k \{(Y_{k,t}, \dots, Y_{k,T})\}, \end{aligned} \quad (A.2)$$

$$\lambda = (\pi_0, A^{(u)}, B^{(u,s)}), s = 1, \dots, M, u = 1, \dots, U, \quad (\text{A.3})$$

where n is the number of consumers in portfolio and T is the term of each loan.

The MCMC algorithms of the particular TMM are similar to the MCMC algorithms of DCMM which are given in Fitzpatrick and Marchev [9]. The five main steps of the MCMC algorithms of the particular TMM are as follows.

A. Priors Specification

The priors on λ are Dirichlet as follows:

$$\begin{aligned} \pi_0 &\sim D(\eta_{01}, \dots, \eta_{0M}), a_{m1}^{(u)}, \dots, a_{mM}^{(u)} \\ &\sim D(\eta_{m1}^{(u)}, \dots, \eta_{mM}^{(u)}), m = 1, \dots, M, u = 1, \dots, U, \end{aligned} \quad (\text{A.4})$$

$$\begin{aligned} b_{l1}^{(u,s)}, \dots, b_{lL}^{(u,s)} &\sim D(\eta_{l1}^{(u,s)}, \dots, \eta_{lL}^{(u,s)}), \\ l = 1, \dots, L, s = 1, \dots, M, u = 1, \dots, U. \end{aligned} \quad (\text{A.5})$$

B. Sampling from $P(X_1^T | Y^{(T)}, Z_1^T, \lambda)$

Given the observations $Y^{(T)}$ and the current value of the parameters λ , we wish to simulate a sample path X_1^T of the hidden Markov chain, from its conditional distribution:

$$\begin{aligned} P(X_1^T | Y^{(T)}, Z_1^T, \lambda) &= P(X_T | Y^{(T)}, Z_1^T, \lambda) \\ &\cdot \prod_{t=1}^{T-1} P(X_t | Y^{(T)}, X_{t+1}^T, Z_1^T, \lambda), \end{aligned} \quad (\text{B.1})$$

We draw values for X_T, X_{T-1}, \dots, X_1 backward. The ‘‘typical term’’ in (B.1) can be written as:

$$\begin{aligned} &P(X_t | Y^{(T)}, X_{t+1}^T, Z_1^T, \lambda) \\ &= P(X_t | Y^{(t)}, Y_{t+1}^T, X_{t+1}^T, Z_1^T, \lambda) \\ &= \frac{P(X_{t+1}^T, Y_{t+1}^T | X_t, Y^{(t)}, Z_1^T, \lambda) P(X_t | Y^{(t)}, Z_1^T, \lambda)}{P(X_{t+1}^T, Y_{t+1}^T | Y^{(t)}, Z_1^T, \lambda)} \\ &= P(X_t | Y^{(t)}, Z_1^T, \lambda) \frac{P(X_{t+1}, X_{t+2}^T, Y_{t+1}^T | X_t, Y^{(t)}, Z_1^T, \lambda)}{P(X_{t+1}^T, Y_{t+1}^T | Y^{(t)}, Z_1^T, \lambda)} \\ &= P(X_t | Y^{(t)}, Z_1^T, \lambda) P(X_{t+1} | X_t, Y^{(t)}, Z_1^T, \lambda) \\ &\cdot \frac{P(X_{t+2}^T, Y_{t+1}^T | Y^{(t)}, X_t, X_{t+1}, Z_1^T, \lambda)}{P(X_{t+1}^T, Y_{t+1}^T | Y^{(t)}, Z_1^T, \lambda)} \\ &= P(X_t | Y^{(t)}, Z_1^T, \lambda) P(X_{t+1} | X_t, Y^{(t)}, Z_1^T, \lambda) \\ &\cdot \frac{P(X_{t+2}^T | Y_{t+1}^T, Y^{(t)}, X_t, X_{t+1}, Z_1^T, \lambda)}{P(X_{t+1}^T | Y_{t+1}^T, Y^{(t)}, Z_1^T, \lambda)} \\ &\propto P(X_t | Y^{(t)}, Z_1^T, \lambda) P(X_{t+1} | X_t, Z_{t+1}, A^{(u)}) \end{aligned} \quad (\text{B.2})$$

By Bayes theorem, the mass function of hidden state given information up to t is:

$$\begin{aligned} P(X_t | Y^{(t)}, Z_1^T, \lambda) &= P(X_t | Y^{(t-1)}, Z_1^T, \lambda) \\ &\cdot \frac{P(Y^{(t)} | X_t, Y^{(t-1)}, Z_1^T, \lambda)}{P(Y^{(t)} | Y^{(t-1)}, Z_1^T, \lambda)}. \end{aligned} \quad (\text{B.3})$$

By the law of total probability, we have:

$$\begin{aligned} &P(X_t | Y^{(t-1)}, Z_1^T, \lambda) \\ &= \sum_{l=1}^m P(X_t | X_{t-1} = l, Z_1^T, \lambda) P(X_{t-1} = l | Y^{(t-1)}, Z_1^T, \lambda) \\ &= \sum_{l=1}^m P(X_t | X_{t-1} = l, Z_t, A^{(u)}) P(X_{t-1} = l | Y^{(t-1)}, Z_1^T, \lambda). \end{aligned} \quad (\text{B.4})$$

The initial probability distribution of hidden states is:

$$P(X_1 | Y^{(0)}, Z_1^T, \lambda) = P(X_1 | Y(0), Z_1, \lambda) = \pi_1. \quad (\text{B.5})$$

Using (B.3)–(B.5), we can obtain $P(X_t | Y^{(t)}, Z_1^T, \lambda)$ for $t = 1, \dots, T$.

Combining with (B.1)–(B.2), X_T, X_{T-1}, \dots, X_1 can be simulated.

C. Extra Permutation Step

To improve the convergence properties of the algorithm, a random permutation of the labels is applied. See Fitzpatrick and Marchev [9] for a detailed description.

$$g \in \{\text{permutations of } (1, 2, \dots, M)\}, \quad (\text{C.1})$$

$$g \sim P(g(X_1), \dots, g(X_T)). \quad (\text{C.2})$$

Then we set $X' = gX$.

D. Sampling from $P(\lambda | X_1^T, Y^{(T)}, Z_1^T)$

Given the process of hidden state X , it is rather straightforward to determine the posterior distribution of λ . Using Bayes’ theorem, the posterior distribution can be simulated separately and independently as follows:

$$\pi_0 | X, Y, Z \sim D(\eta_{0,1} + \omega_{0,1}, \dots, \eta_{0,M} + \omega_{0,M}), \quad (\text{D.1})$$

$$\begin{aligned} a_{m1}^{(u)}, \dots, a_{mM}^{(u)} | X, Y, Z &\sim D(\eta_{m1}^{(u)} + \omega_{m1}^{(u)}, \dots, \eta_{mM}^{(u)} + \omega_{mM}^{(u)}), \\ m = 1, \dots, M, u = 1, \dots, U, \end{aligned} \quad (\text{D.2})$$

$$\begin{aligned} b_{l1}^{(u,s)}, \dots, b_{lL}^{(u,s)} | X, Y, Z &\sim D(\eta_{l1}^{(u,s)} + \omega_{l1}^{(u,s)}, \dots, \eta_{lL}^{(u,s)} + \omega_{lL}^{(u,s)}), \\ l = 1, \dots, L, s = 1, \dots, M, u = 1, \dots, U, \end{aligned} \quad (\text{D.3})$$

where $\omega_{0,i} := I(X_1 = i)$, $\omega_{i,j}^{(u)} := \sum_{t=2}^T I(X_{t-1} = i, X_t = j, Z_t = u)$, $\omega_{i,j}^{(u,s)} := \sum_{k=1}^N \sum_{t=2}^T I(Y_{k,t-1} = i, Y_{k,t} = j, X_t = s, Z_t = u)$, I is indicator function.

The steps B, C and D are repeated until a maximum number of iterations is reached.

E. Post-Processing Algorithm

The posterior inference step should be trivial once X is drawn but the draw is complicated by a nonidentifiability problem called *label switching*. This will mean that ergodic averages of component-specific quantities will be identical and thus useless for inference. In dealing with the problem of *label switching*, a post-processing algorithm is applied to ensure the labels of the hidden states are consistent for all iteration. See Figure 3 of Boys and Henderson [28] for a detailed description.

If at iteration i the current estimate of \widehat{X}^* is $\widehat{X}_{(i-1)}^*$ then

- (E.1) Choose a permutation v_i to minimize $-\sum_{t=1}^T I(v_i(X_{t,(i)}) = \widehat{X}_{t,(i-1)}^*)$;
- (E.2) Apply the permutation v_i to output $X_{(i)}$ and $\lambda_{(i)}$;
- (E.3) For $t = 1, 2, \dots, T$, set $\widehat{X}_{t,(i)}^* = \arg \max_{j \in S(X)} \sum_{o=1}^i I(v_k(X_{t,(o)}) = j)$;
- (E.4) Let Q be the size of sampling. We estimate the posterior probabilities of the hidden states X_t along with the sequence by $P(X_t = j|Y, Z) = (1/Q) \sum_{q=1}^L I(\widehat{X}_{t,q}^* = j)$.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (grant numbers 71790593, 71521061, and 71573077). We are grateful to the editor and referees for many valuable comments and suggestions.

References

- [1] R. M. Cyert, H. J. Davidson and G. L. Thompson, "Estimation of the allowance for doubtful accounts by markov chains," *Management Science*, vol. 8, no. 3, pp. 287–303, 1962.
- [2] R. M. Cyert and G. L. Thompson, "Selecting a portfolio of credit risks by markov chains," *The Journal of Business*, vol. 41, no. 1, pp. 39–46, 1968.
- [3] A. W. Corcoran, "The use of exponentially-smoothed transition matrices to improve forecasting of cash flows from accounts receivable," *Management Science*, vol. 24, no. 7, pp. 732–739, 1978.
- [4] J. G. Kallberg and A. Saunders, "Markov chain approaches to the analysis of payment behaviour of retail credit customers," *Financial Management*, vol. 12, no. 2, pp. 5–14, 1983.
- [5] M. Malik and L. C. Thomas, "Transition matrix models of consumer credit ratings," *International Journal of Forecasting*, vol. 28, no. 1, pp. 261–272, 2012.
- [6] P. He, Z. Hua, and Z. Liu, "A quantification method for the collection effect on consumer term loans," *Journal of Banking and Finance*, vol. 57, pp. 17–26, 2015.
- [7] G. Giampieri, M. Davis, and M. Crowder, "A hidden markov model of default interaction," *Quantitative Finance*, vol. 5, pp. 27–34, 2005.
- [8] K. Banachewicz, A. Lucas, and A. van der Vaart, "Modelling portfolio defaults using hidden markov models with covariates," *The Econometrics Journal*, vol. 11, no. 1, pp. 155–171, 2008.
- [9] M. Fitzpatrick and D. Marchev, "Efficient bayesian estimation of the multivariate double chain markov model," *Statistics and Computing*, vol. 23, no. 4, pp. 467–480, 2013.
- [10] L. Quirini and L. Vannucci, "Creditworthiness dynamics and hidden markov models," *Journal of the Operational Research Society*, vol. 65, no. 3, pp. 323–330, 2014.
- [11] D. Anastasiou, "Is ex-post credit risk affected by the cycles? the case of Italian banks," *Research in International Business and Finance*, vol. 42, pp. 242–248, 2017.
- [12] A. T. Vouldis and D. P. Louzis, "Leading indicators of non-performing loans in Greece: the information content of macro-, micro- and bank-specific variables," *Empirical Economics*, vol. 54, no. 3, pp. 1187–1214, 2018.
- [13] D. Anastasiou, H. Louri, and M. Tsionas, "Non-performing loans in the euro area: are core-periphery banking markets fragmented?" *International Journal of Finance and Economics*, vol. 24, no. 1, pp. 97–112, 2019.
- [14] F. Grigoli, M. Mansilla, and M. Saldías, "Macro-financial linkages and heterogeneous non-performing loans projections: an application to Ecuador," *Journal of Banking and Finance*, vol. 97, pp. 130–141, 2018.
- [15] W. Pieczynski, C. Hular, and T. Veit, "Triplet Markov Chains in hidden signal restoration," *Image and Signal Processing for Remote Sensing VIII*, vol. 4885, pp. 58–68, 2003.
- [16] S. Bricq, C. Collet, and J.-P. Armspach, "Triplet markov chain for 3D MRI brain segmentation using a probabilistic atlas 3rd IEEE International Symposium on Biomedical Imaging: Macro to Nano," vol. 1–3, pp. 386–389, 2006.
- [17] P. Lanchantin, J. Lapuyade-Lahorgue, and W. Pieczynski, "Unsupervised segmentation of triplet Markov chains hidden with long-memory noise," *Signal Processing*, vol. 88, no. 5, pp. 1134–1151, 2008.
- [18] M. El Yazid Boudaren, E. Monfrini, W. Pieczynski, and A. Aissani, "Phasic triplet markov chains," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 11, pp. 2310–2316, 2014.
- [19] I. Gorynin, H. Gangloff, E. Monfrini, and W. Pieczynski, "Assessing the segmentation performance of pairwise and triplet Markov models," *Signal Processing*, vol. 145, pp. 183–192, 2018.
- [20] L. E. Baum and T. Petrie, "Statistical inference for probabilistic functions of finite state markov chains," *The Annals of Mathematical Statistics*, vol. 37, no. 6, pp. 1554–1563, 1966.
- [21] L. Zhang, Z. Li, Y. Xu, and Y. Li, "Multi-period mean variance portfolio selection under incomplete information," *Applied Stochastic Models in Business and Industry*, vol. 32, no. 6, pp. 753–774, 2016.
- [22] R. Wang, Y. Li, H. Sun, Y. Zhang, and Y. Sun, "Performance analysis of switched control systems under common-source digital upsets Modeled by MDHMM," *Complexity*, vol. 2018, Article ID 4329053, 12 pages, 2018.

- [23] N. Malešević, D. Marković, G. Kanitz, M. Controzzi, C. Cipriani, and C. Antfolk, "Vector autoregressive hierarchical hidden markov models for extracting finger movements using multichannel surface EMG signals," *Complexity*, vol. 2018, Article ID 9728264, 12 pages, 2018.
- [24] A. Berchtold, "The double chain markov mode," *Communications in Statistics - Theory and Methods*, vol. 28, no. 11, pp. 2569–2589, 1999.
- [25] M. D. Hamilton, "A new approach to the economic analysis of nonstationary time series and the business cycle," *Econometrica*, vol. 57, no. 2, pp. 357–384, 1989.
- [26] A. Bangia, F. X. Diebold, A. Kronimus, C. Schagen, and T. Schuermann, "Ratings migration and the business cycle, with application to credit portfolio stress testing," *Journal of Banking & Finance*, vol. 26, no. 2-3, pp. 445–474, 2002.
- [27] M. Leow and J. Crook, "Intensity models and transition probabilities for credit card loan delinquencies," *European Journal of Operational Research*, vol. 236, no. 2, pp. 685–694, 2014.
- [28] R. J. Boys and D. A. Henderson, "On determining the order of Markov dependence of an observed process governed by a hidden Markov model," *Scientific Programming*, vol. 10, Article ID 683164, 11 pages, 2002.

Research Article

Modeling Investor Behavior Using Machine Learning: Mean-Reversion and Momentum Trading Strategies

Thiago Christiano Silva ^{1,2} Benjamin Miranda Tabak ³
and Idamar Magalhães Ferreira³

¹Universidade Católica de Brasília, Distrito Federal, Brazil

²Department of Computing and Mathematics, Faculty of Philosophy, Sciences, and Literatures in Ribeirão Preto, Universidade de São Paulo, São Paulo, Brazil

³FGV/EPPG Escola de Políticas Públicas e Governo, Fundação Getúlio Vargas, School of Public Policy and Government, Getúlio Vargas Foundation, Distrito Federal, Brazil

Correspondence should be addressed to Benjamin Miranda Tabak; benjaminm.tabak@gmail.com

Received 27 August 2019; Revised 9 November 2019; Accepted 21 November 2019; Published 26 December 2019

Academic Editor: José Manuel Galán

Copyright © 2019 Thiago Christiano Silva et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

We model investor behavior by training machine learning techniques with financial data comprising more than 13,000 investors of a large bank in Brazil over 2016 to 2018. We take high-frequency data on every sell or buy operation of these investors on a daily basis, allowing us to fully track these investment decisions over time. We then analyze whether these investment changes correlate with the IBOVESPA index. We find that investors decide their investment strategies using recent past price changes. There is some degree of heterogeneity in investment decisions. Overall, we find evidence of mean-reverting investment strategies. We also find evidence that female investors and higher academic degree have a less pronounced mean-reverting strategy behavior comparatively to male investors and those with lower academic degree. Finally, this paper provides a general methodological approach to mitigate potential biases arising from *ad-hoc* design decisions of discarding or introducing variables in empirical econometrics. For that, we use feature selection techniques from machine learning to identify relevant variables in an objective and concise way.

1. Introduction

This paper studies the determinants of investors' behavior in the stock market using transaction-level data on buy and sell operations of investors. Our data contains detailed information of the investor's identity and her socioeconomic characteristics, the investment value, and variation due to the buy or sell operation over time. The data is confidential and comes from a large and representative Brazilian bank. With this rich dataset, we are able to study how investors respond to changes in the Brazilian stock market due to variations of its market index, called IBOVESPA. We use historical variations of the IBOVESPA index with different horizons (window length) to test which one better predicts the investors' behavior.

To mitigate potential concerns due to subjective decisions by the analyst—and also to prevent discarding a

potentially relevant predictor—we opt to use an objective approach to identify those horizons that best explain investors' buy or sell operations. For that, we use a robust feature selection technique borrowed from the machine learning literature called elastic net. The great advantage of the elastic net comes by the simplicity of its loss function (just like a regression) and also the robustness in preventing overfitting by optimally using a convex combination of the Lasso and Ridge regularization methods. Overfitting can occur as the algorithm may learn the dynamics of the variable of interest and fit very well the training dataset but with poor predictability in other datasets. Evaluating the potential for overfitting is essential for researchers as it may undermine the model. We understand that our method seeks to avoid, to some extent, the perils of overfitting. The Ridge and the Lasso algorithms impose penalties for large weights in the model [1]. In this way, they tend to reduce the

model's complexity and hence are able to minimize concerns about overfitting.

Investors tend to trade using different strategies, such as buy-and-hold (passive strategy) or an active strategy in which they seek to outperform a benchmark, for example, a market index. If investors trade using active strategies, they may use two different and well-known approaches, a mean-reverting or a momentum strategy. See [2–4]; for seminal contributions in these two strategies.

In the first case, they react to market swings by betting that the market will mean-revert. Therefore, they assume the trend will change and therefore will sell after substantial upward changes and buy after downward changes. In the momentum strategy, they will bet that the trend will persist. Thus, they will increase investments in the stock market after an increase in the market index.

While we understand that they may be other trading strategies, we focus on the mean-reversion and momentum strategies because they are well established in the literature and serve as building blocks of many other strategies. There is a large body of the literature that discusses their use in different contexts [5–11]. In addition, they are easily testable in empirical specifications. Therefore, we seek to understand if investors decide to hold their stocks or sell them after negative/positive shocks.

The issue of how investors will behave on average is empirical. There is a large body of the literature that discusses predictability for the stock market [12–16]. In addition, there is another strand focusing on cognitive biases and excess trading on equity and other financial assets [17–23]. Our data containing transaction-level operations of buy and sell operation permit us to follow each investors' decisions over time and therefore test whether they use mean-reverting or momentum strategy after changes in the stock market index.

It is essential to notice that, if traders use such strategies, they may induce higher volatility in the market with their actions. In theory, market changes should occur as new information arrives, which is economically relevant to estimate future profits and dividend distribution. However, price substantially changes over time and volatility is higher than we would expect in a rational market. Therefore, we assume that the traders' decision to trade excessively will induce higher volatility in the market. Investors' decisions that follow different trading strategies may generate complex patterns in prices and volatility. They may induce long-range correlation, short-term predictability, and chaotic dynamics in prices over time. There is a large body of the literature that attempts to explain complex macrobehavior of systems using a composition of local rules. For that, agent-based modelling has been extensively used to explain price and volatility using artificial markets [24]. Using agent-based modelling, LeBaron [25] explores structural (macro) features that emerge in a market where participants adapt and evolve over time, while Bertella et al. [26] study the effect of investor's behavioral bias in prices. Understanding how investors behave and perform trading strategies is the first step for better understanding the complexity that is intrinsic to financial markets. Our paper also contributes to this matter.

To identify the most relevant predictors that explain investors' behavior, we depart from using traditional panel-data econometric techniques and goodness-of-fit measures and instead employ more robust methodologies borrowed from the machine learning literature. Contrasting to usual econometrics techniques that summarize relationships using linear regression analysis, machine learning offers a set of tools that can potentially capture nonlinear relationships between the data. According to Varian [27], bridging the gap between machine learning and econometrics is a natural tendency mainly because of the presence of large amount of data and the rising complexity—potentially highly nonlinear—between data relationships. Our work contributes to this endeavour by providing a real case study of a financial dataset using machine learning techniques.

Comparatively to econometrics, machine learning techniques have strong model selection techniques, mainly through the use of cross-validation techniques, which are a type of repeated resampling in random subsets of the dataset. Initially, the cross-validation procedure divides the data set into two disjoint and complete subsets: the training set and the test set. All the model's parameters are tuned using only the training set. After the model is selected (tuned) using the training data, we run it against the test set to check its accuracy or some other performance metric. The rationale is that, by training the model with some data and testing against another subset, we are estimating the model's out-of-sample prediction power and not simply learning the data. The test set therefore would be a simulation of real (production) data and the model's performance on this dataset would represent a rough estimate of actual performance of the model in real unseen data.

Since our data set comprises more than 350,000 observations representing individual investor's movements with respect to their investments over 2016 to 2018, we apply regularization techniques to prevent model's overfitting during the feature selection procedure with training data. For that, we apply an elastic net procedure [1] to control for the model's complexity. Elastic net is a generalization of the Ridge (L_2 -norm) and Lasso (L_1 -norm) and hence is more robust. It uses an optimal convex combination of both types of regularization. Lasso tends to shrink the majority of the nonrelevant regressors to zero while keeping only the most important regressors as nonzero. In contrast, Ridge tends to output nonzero coefficients for almost all regressors. By using both regularization schemes, we are able to enjoy the positive characteristics of both schemes.

Regularization is an important issue in large data sets because it prevents methods with high variance and low bias from overfitting [28, 29]. This is the well-known bias-variance trade off in the machine learning literature [30]. While low bias prevents overfitting, it can generate underfitting of the data set. In contrast, high-variance methods can learn noise from the data and let go the true relationships of the data set. Low bias favors low model complexity at the cost of a potential overfitting. High variance tends to successfully capture smoothly nonlinear relationships between the data at the expense of a potential overfitting. Examples of low-bias algorithms are the linear regression or neural networks

with a single layer. Examples of high-variance algorithms are decision trees and multilayer neural networks. It is important to first set the rationale behind the regularization process from the viewpoint of our financial data set of buy and sell operations. On the one hand, a strand in the economics literature advocates that the agents' decisions are completely rational, in that decisions are taken by considering all information from the market (complete information) [31]. On the other hand, another body of the literature argues that investors cannot potentially consider every single information from the market when taking their decisions because (i) the agent does not have complete information and (ii) even if the agent did have complete information, she would be unable to perform all required calculations. In this way, they would naturally focus on the most relevant variables. In this case, we say that investors have a bounded rationality, term first coined by Simon [32]. We can frame these two theories into the two types of regularization frameworks used in this paper. Investors with unbounded rationality, i.e., that consider all potential variables, would better be modeled by a Ridge regularization procedure because it does not tend to place zero importance on any variables. In contrast, investors with bounded rationality would be better modeled by a Lasso regularization because it would choose a few (and more relevant) variables and set the remainder as zero. By using a weighted convex combination of both Ridge and Lasso regularization procedures, we are effectively considering both cases in our estimation process.

While Brazil does not have well-developed stock markets as advanced economies, it is an important emerging country that, due to its size and relative importance to its peers, deserves to be studied. In addition, capital markets have been increasing in the last years (according to the BM&FBovespa, which is the Brazilian stock exchange, the number of investors increased almost 20% from 2017 to 2018.), which strengthen the relevance of our work. Our main results suggest that investors use a mean-reverting trading strategy. Therefore, they reduce their investments after positive changes in the IBOVESPA and increase it after negative changes.

We also test whether investors' biological and socio-economic characteristics explain their trading behavior. In terms of schooling, educated investors, in theory, should behave in more rational ways and trade less frequently when there is no new information arriving continuously in the market, at least those that are not relevant regarding potential for future profits. Therefore, we would expect these investors to have a smaller reaction to price fluctuations. We also test dissimilarities in investment decision making arising from the gender. Neyse et al. [33] and Lundeborg et al. [34] partly attribute investment differences among males and females due to systematic changes in overconfidence. Excessive overconfidence is associated with higher levels of testosterone, which is more pronounced in males. Overconfidence may induce investors to take on higher risks, leading them to look for higher returns in the short term. In this way, we would expect a less sensitive behavior of females to changes of past IBOVESPA variations

as they would value more fundamentals and look for yields in the longer term. Our empirical analyses corroborate these views.

Several papers have studied investor behavior. Onishchenko and Ülkü [35] show a change in foreign investors, which have become more sophisticated. They find that foreign investors in Korea do not chase returns as the previous literature normally reports. Their results suggest a transition from positive to negative feedback trading over time. Abreu [36] finds that investors that buy warrants have specific characteristics, such as young age and less educated, or investors with gambling attitudes (and overconfidence) (see also [37–41]). To the best of our knowledge, our paper is one of the first that uses machine-learning techniques to unveil what are the characteristics that matter the most for explaining investor behavior at the disaggregated level. We study the reaction of investors to market changes and test whether they employ momentum or mean-reverting strategies.

2. Data

We collect and match several unique proprietary and public datasets. Our sample consists of public information from the IBOVESPA index, investor-specific information, and a proprietary customer database from a large Brazilian bank with investor-specific matched daily transactions on buy and sell operations in the IBOVESPA stock exchange market. The last two datasets are confidential.

The first source is the IBOVESPA index of the Brazilian stock exchange (BM & FBovespa). This index is considered the stock market benchmark for Brazil. We have 747 days in our sample spanning over the years of 2016 to 2018.

The second source is the investor registration information, such as their profession, degree of education, and equity. Information is from the database of the home broker and customer relationship management (CRM) solution. Our data set is comprehensive and encompasses 13,634 investors.

The last source provides each transaction made by each investor, on BM & FBOVESPA and on each of the days between January 2, 2016, and December 31, 2018, as well as their daily holdings. We observe their daily trading activities on investment decisions. This rich data set enables us to keep track of investors' buy and sell operations over time and therefore permits us to test whether they use the mean-reverting trading strategy or the momentum trading strategy as a response to IBOVESPA index changes. These are two common trading strategies that have been discussed in the literature [42, 43]. Other strategies exist, which may be more complex in nature and difficult to model, and they are not the object of our analysis. One such example would be rational traders that employ fundamental analysis and forecast future profits of traded companies to estimate their potential to distribute dividends and can value these stocks. The sample has 1,099,985 trading decisions (change in the investment volume). We also have 358,176 customer holdings over time.

Table 1 reports summary statistics of our data on investor's daily decisions on their investments. We can see that

TABLE 1: Summary statistics of our panel data on investors' daily trading decisions over the period of 2016 to 2018.

Statistic	N	Mean	St. Dev.	Min	Pctl (25)	Median	Pctl (75)	Max
Investment variation (%)	356,172	9.267	68.389	-100.000	-6.680	0.510	8.770	499.879
1-day IBOVESPA variation	356,172	0.145	1.528	-4.870	-0.740	0.110	1.010	6.600
2-day IBOVESPA variation	355,796	0.323	2.160	-6.550	-1.140	0.290	1.600	9.130
3-day IBOVESPA variation	355,419	0.472	2.598	-7.950	-1.140	0.540	2.020	10.880
5-day IBOVESPA variation	354,588	0.781	3.274	-8.250	-1.230	0.770	2.750	16.870
30-day IBOVESPA variation	343,592	4.863	8.282	-19.060	-0.740	5.240	10.500	28.770
IBOVESPA index	356,176	0.145	1.528	-4.870	-0.740	0.110	1.010	6.600

there is a large range of daily investment variations, going from -100% to almost 500%. On average, we see a positive investment variation (9.267%). We also show the IBOVESPA index level and its variations in the last 1, 2, 3, 5, and 30 days. We will use these IBOVESPA index changes to check how they correlate with the investment variations variable. One underlying hypothesis is that investors look at the IBOVESPA index to decide on their trading decisions.

Figures 1(a) and 1(b) portray weekday heatmaps showing the average daily investment changes in 2016 broken down by investor's gender and education level (see [44]). First, we observe the richness of our data set in which there is a large heterogeneity of investor's decisions on their investment on a daily basis. Second, though there is a similarity on how investors decide on their investments for males and females and for those with higher and lower education, we observe some discrepancies in some occasions, suggesting that these are two important features that we should study in our empirical analysis. Besides this subjective analysis, our feature selection procedure will corroborate such vision using an objective and quantitative approach. For instance, we observe that, on average, investors mostly buy by the beginning or end of the week while they sell on Wednesdays. There is evidence of behavioral changes of investors over weekdays in the stock market. For instance, Pena [45] studies the effect of reform on the Spanish stock exchange market. They find that, before such reform, there were positive abnormal excess returns on Mondays, effect of which disappeared following that reform.

Figure 2 shows how investments are split across Brazilian states over time. As we can see, there is also some heterogeneity across investors residing in different states, which suggests that we may have to control for state origin of investors. For instance, there are some large investment variations in the Northern region of Brazil.

Figure 3 depicts the distribution of investment variation across different Brazilian states broken down by investor's gender (female or male). Each distribution conditioned to the state and gender integrates to one. Interestingly, most of the distributions have three persist modals that occur not only across different states but also for different genders. The modals are centered at the zero (no investment variation) and at $\pm 30\%$ investment variation marks. While, in most cases, the profiles of investment changes of both male and female largely coincide, there are some notable exceptions. For instance, in less developed regions—such as the North and Northeast—the distributions of investment decisions of

males and female significantly differ at some dates. Overall, males tend to change more their investment positions relatively to females. However, such feature is even more pronounced in less developed regions.

Figure 4 displays the same distribution of investment variation across different Brazilian states but now broken down by the investor's academic education. We consider investors with higher education and those with high school or below. Again, the three-modal distribution found when we broke down the distribution by investor's gender also show up when we look at their educational levels. In more developed regions—such as the Southeast and South—investors' decision are roughly the same regardless of their academic educational levels. Such similarity reinforces the divergence of academic degree and the level of financial literacy, especially in trading. In contrast, we observe a large heterogeneity in the North region; less educated investors tend to vary their investment positions more than investors those with higher education.

3. Feature Selection Using Machine Learning

In this section, we analyze the predictive power of our attributes in explaining investor responses to changes in the Brazilian stock market index. We use different time aggregations of changes in the IBOVESPA index, which is the financial index that is carefully looked by investors when deciding their investment strategies in Brazil. We use 2-, 3-, and 5-day IBOVESPA index variations, as well as 3- and 5-day IBOVESPA index averages. This analysis sheds light on how investors look at IBOVESPA changes when deciding their trading strategies in the stock market. It is an empirical open question to test whether investors take the very short-term changes, such as 2- or 3-day, or a more prolonged window, such as 5-day changes.

To test the predictive power, we use data-driven machine learning methods to identify the most relevant attributes [46–48]. Since we have data from 13,247 investors from January 1, 2016, to December 31, 2018, on a daily basis, we need first to purge out any macroeconomic factor that could be affecting all investors *in the same manner* over this time frame. This becomes even more important due to the fact that Brazil was facing a recession from 2014Q4 to 2016Q4 and therefore our sample contains part of that period. We perform this preprocessing to homogenize the data distribution, since machine learning methods best perform on cross-sectional data [30, 49].

To remove time factors homogeneously faced by investors in a period, we use a static panel-data specification

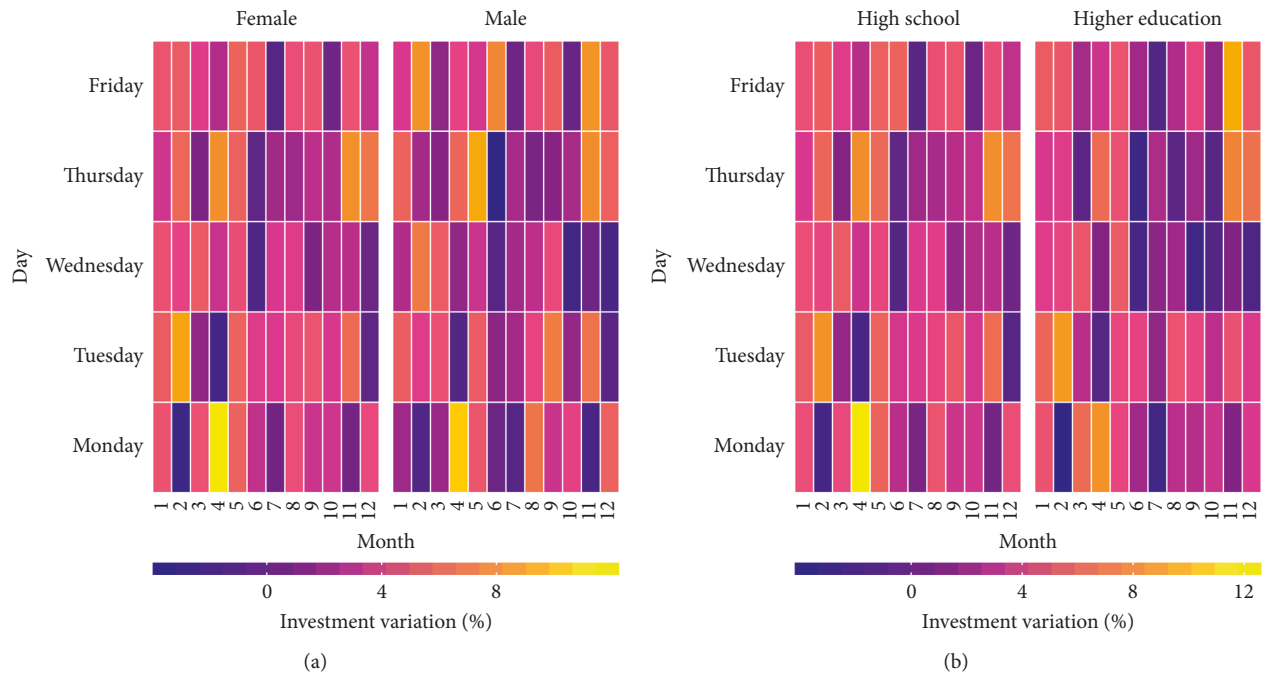


FIGURE 1: Heatmap showing average daily investment changes (%) in 2016 broken down by investor’s gender (a) and education degree (b). We take average values of investment decisions that took place in every weekday within the month. Due to the existence of some large absolute values and to improve readability, we winsorize the investment variation distribution by 5% at each side before taking the average.

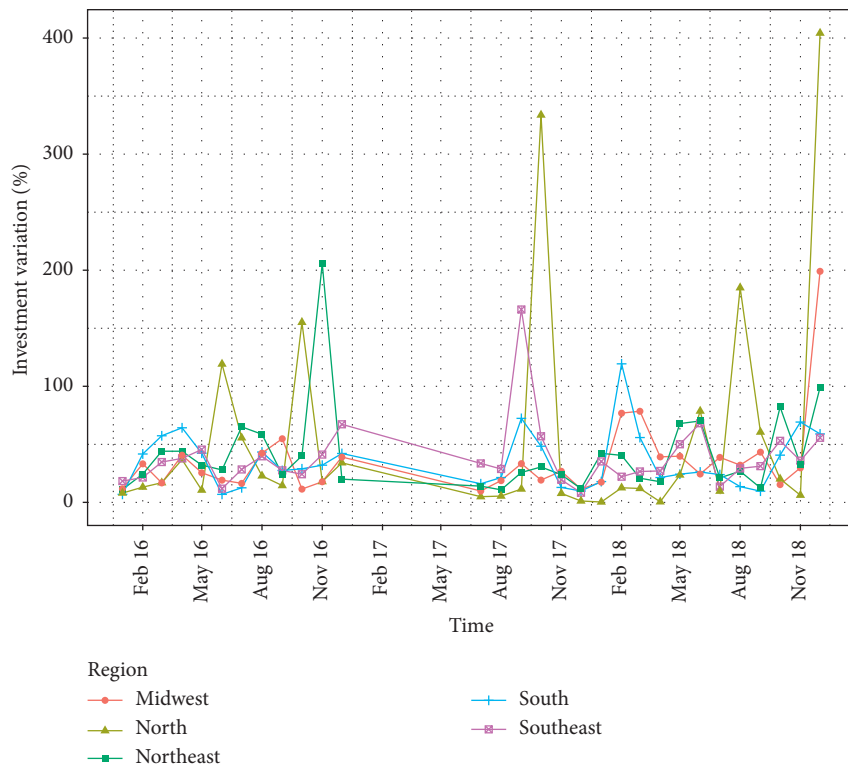


FIGURE 2: In this figure, we aggregate investment made by all investors by states to which they belong. As we can we have investors from all parts of Brazil in our database.

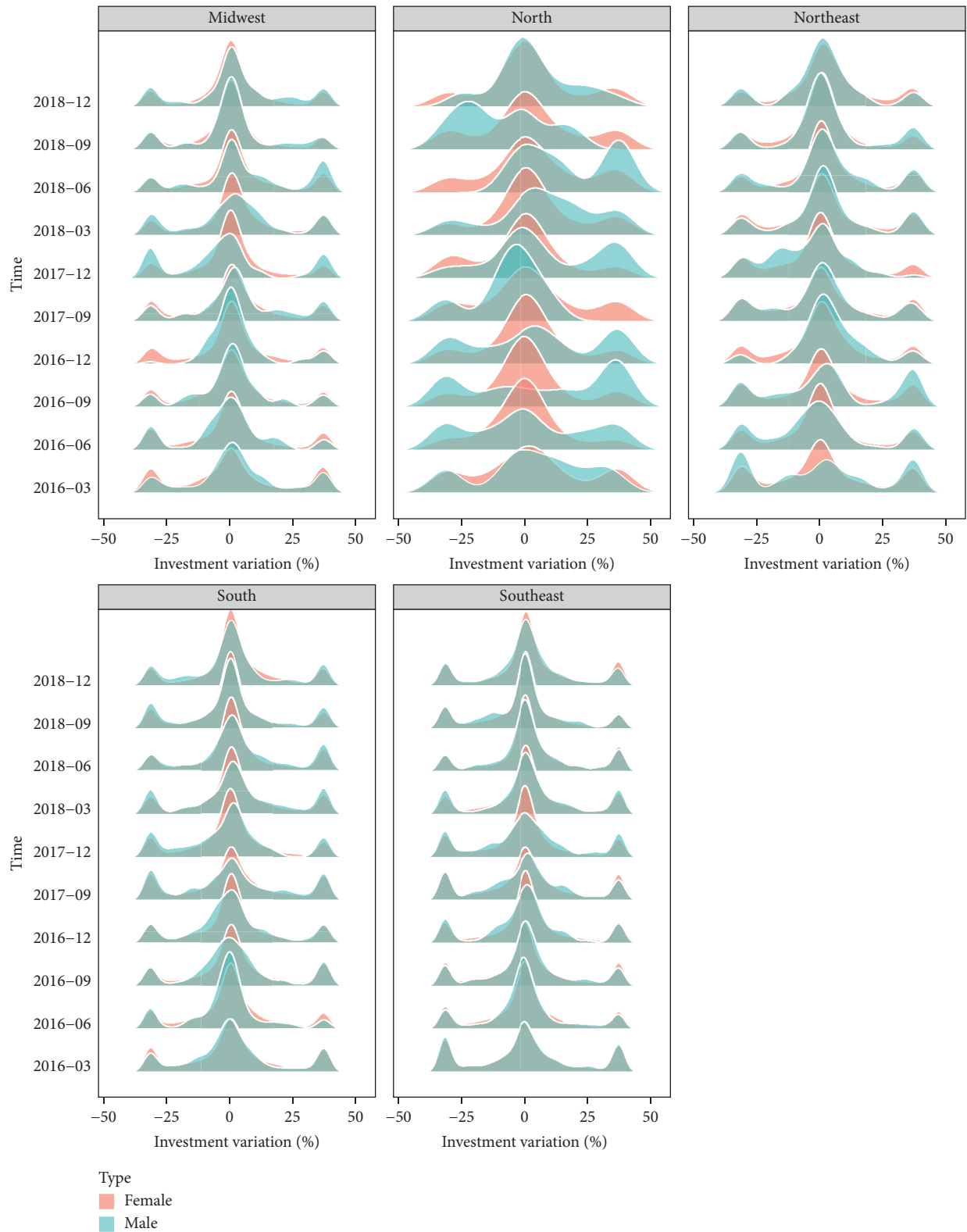


FIGURE 3: Distribution of investment variation across different Brazilian states (Midwest, North, Northeast, South, and Southeast) broken down by investor's gender (female or male). Due to the existence of some large absolute values and to improve readability, we winsorize the investment variation distribution by 5% at each side.

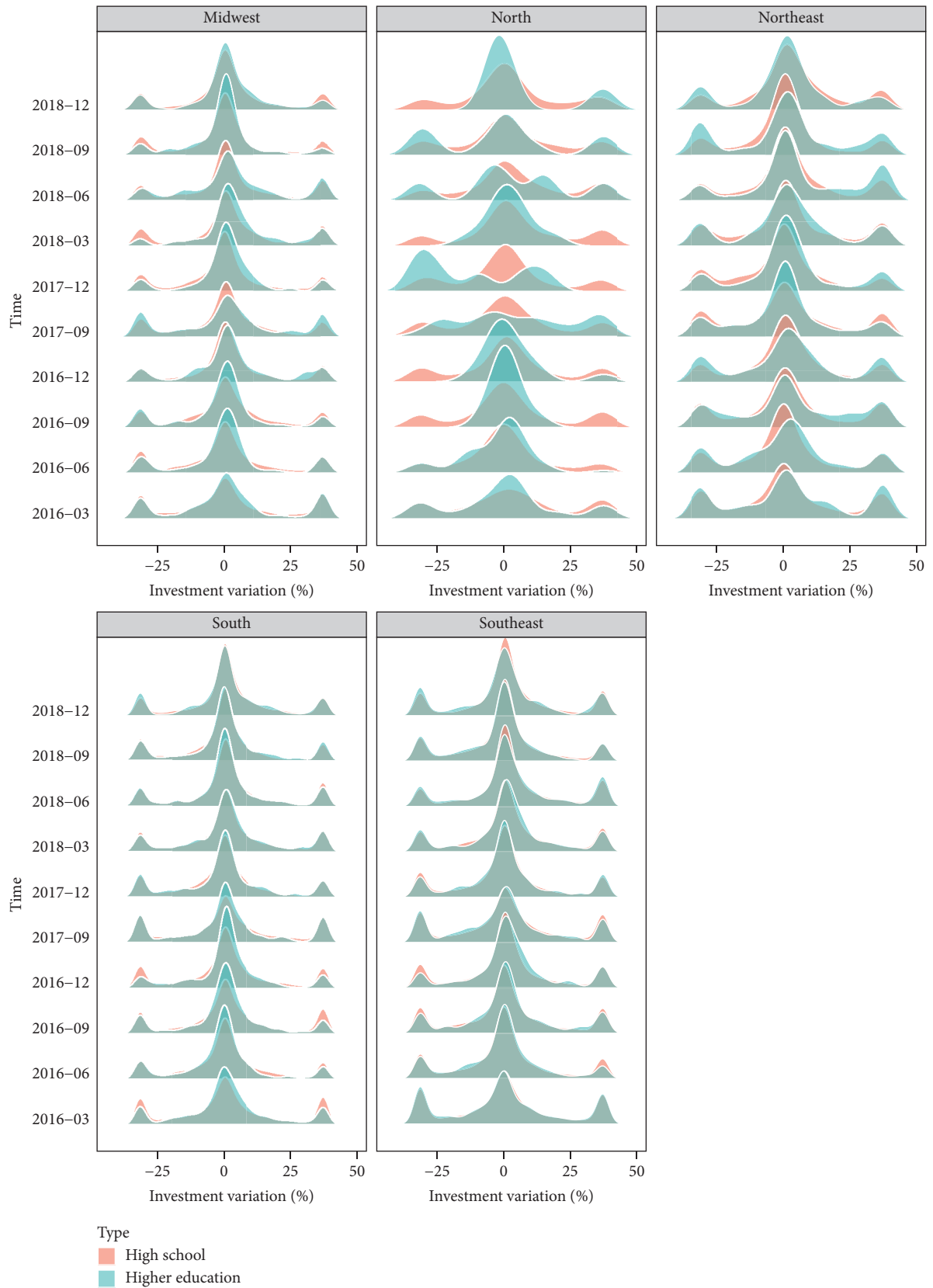


FIGURE 4: Distribution of investment variation across different Brazilian states (Midwest, North, Northeast, South, and Southeast) broken down by investor's education (higher education or high school). Due to the existence of some large absolute values and to improve readability, we winsorize the investment variation distribution by 5% at each side.

with time fixed effects to purge out macroeconomic components as follows:

$$\Delta y_{it} = \alpha_t + \epsilon_{it}, \quad (1)$$

in which Δy_{it} denotes the portfolio volume variation in the stock market of investor i at time t , α_t represents time fixed effects, and ϵ_{it} is the residual. In this specification, we interpret the residual ϵ_{it} as any variation of investor i 's portfolio volume at time t that is not due to any time common factor, such as the underlying macroeconomic scenario. By using ϵ_{it} instead of Δy_{it} , we can effectively treat the data as a large cross-sectional unit. Hence, we are able to fully use machine learning methods at their best setup, which we discuss further.

We choose an elastic net regression to estimate the importance of each attribute in the model. Such regression optimally combines L_2 -norm (Ridge) and L_1 -norm (Lasso) regularization. Therefore, we are able to prevent any overfitting in our empirical model. Moreover, we use a convex combination of L_1 -norm, which tends to shrink the majority of the nonrelevant regressors to zero and keep the most important nonzero, and L_2 -norm, which tends to output nonzero and approximate coefficients for all the similar regressors. By using both regularization schemes, we are able to enjoy the positive characteristics of both schemes.

To select the most important attributes, we use the residual ϵ_{it} , the investment volume variation of investor i at time t not due to common time factors, as dependent variable and different IBOVESPA index time aggregations and investors' biological and education characteristics as independent variables as follows:

$$\epsilon_{it} = \beta^T \cdot \mathbf{X}_{it} + \text{error}_{it}, \quad (2)$$

in which \mathbf{X}_{it} is a vector composed of past IBOVESPA changes with different windows (1-, 2-, 3-, 5-, 10-, 20-, and 30-day IBOVESPA changes) and investors' characteristics (state of residency, gender, and level of schooling). The term error_{it} is the standard error. According to the elastic net procedure, we select β that minimizes the following loss function $L(\beta)$:

$$L(\beta) = \sum_{t=1}^T \sum_{i=1}^N \left(\epsilon_{it} - \sum_{j=1}^p \beta_j x_{it}^{(j)} \right)^2 + \lambda \left[(1 - \alpha) \frac{\|\beta\|_2^2}{2} + \alpha \|\beta\|_1^2 \right], \quad (3)$$

in which $t \in \{1, \dots, N\}$ index times on a daily basis (from January 1, 2016, to December 31, 2018,) and i index investors. The term $x_{it}^{(j)}$ indexes the j th regressor of investor i at time t . The operators $\|\cdot\|_1$ and $\|\cdot\|_2$ indicate L_1 - and L_2 -norms taken over the vector input.

The first expression in (3) denotes the traditional data fitting error (residuals), while the second represents the regularization term. Parameter λ modulates the importance of the traditional and regularization terms. The term α controls the convex mixture of L_1 and L_2 regularization. The regularization works by penalizing large β coefficients. Therefore, it shrinks the estimated coefficients and the

overall fit function becomes smoother over the data distribution.

In the elastic net regression, a takes values in between 0 and 1. We optimally tune a and λ using a nested cross-validation procedure with $k = 10$ folds and 100 independent repeats for statistical robustness [29, 49]. In this procedure, we use $k - 1 = 9$ folds for training and the remaining fold for testing. This procedure is cycled k times such that each fold appears exactly once for testing. Such methodology enables us to tune the regularization parameters while preventing overfitting of the model. We optimize a over the grid search space $\{0, 0.05, 0.10, \dots, 1\}$ and λ over $\{0, 0.1, 0.2, \dots, 5\}$. As standard practice, we preprocess all regressors by applying a Z-score standardization over all the data points using predetermined values extracted only from the training data (so as to prevent data leakage from the test set).

Figure 5 shows our results for the importance of different time aggregations of the IBOVESPA index in explaining investors' behavior. The optimal regularization parameters were $\lambda = 0.1$ and $\alpha = 0.35$. We normalize the coefficients in terms of the most important attribute. The attribute "1-day IBOVESPA variation" is the most powerful predictor for explaining investors' behavior, followed by "2-day IBOVESPA variation" and "5-day IBOVESPA variation." This suggests that investors prefer to base their investment decisions using short-term variations of the stock market index. Even though more prolonged periods of IBOVESPA index changes are important—such as 10-, 20-, and 30-day variations—they are much less important than the short-term variations. In addition, we find that investors' gender and schooling level are also important characteristics for explaining buy and sell operations over the Brazilian stock exchange market in the period from 2016 to 2018. We also observe that some regional variables are important, such as Santa Catarina, Rio de Janeiro, Distrito Federal, Minas Gerais, and Paraná and São Paulo. This may suggest a different mass of investors' composition across different states.

Our feature selection procedure gives us an objective way of identifying potentially important variables that should be accounted for in our econometric exercise. Such tool taken together with the analyst's expertise to assess their validity in terms of relationship with the analyzed measure is an important step in producing econometric methods in a more reliable manner. Our results point that we should control for investors' characteristics (gender and schooling level) and also past IBOVESPA variations. The investor's state is not important because we will employ a panel-data analysis with fixed effects at the investor level. Therefore, the investor's state is collinear with the investor fixed effect and would be dropped during the estimations.

4. Econometric Analysis with Selected Variables

In the previous section, we have found that short-term variations of the IBOVESPA index are better predictors for buy or sell operations in the Brazilian stock exchange market than long-term variations. The feature selection procedure is a transparent way of choosing relevant variables in an objective way. However, such method does not provide an

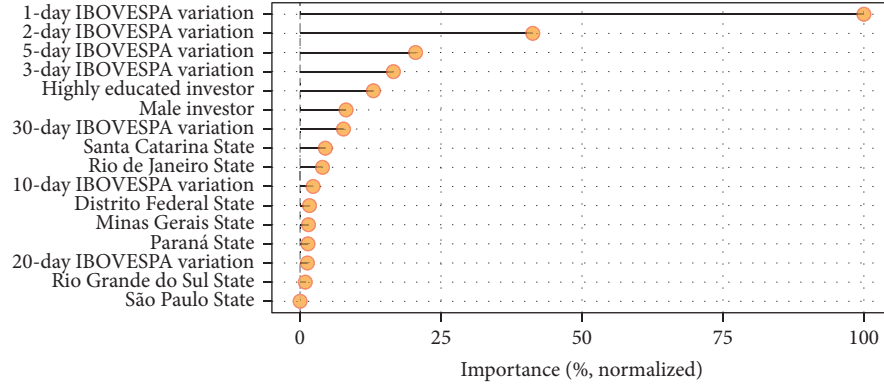


FIGURE 5: Feature selection results using an elastic net procedure with L_2 and L_1 regularization. Coefficients are normalized in terms of the most important attribute (“1-day IBOVESPA variation”).

answer as to whether each variable contributes positively or negatively to the target variable, i.e., the investment decision of the investor (buy or sell). In this section, we look at such direction and estimate the magnitude of the most relevant variables found by our feature selection technique.

In Section 4.1, we first test whether investment decisions of Brazilian investors better fit to a mean-reversal or momentum strategy. For that, we regress total investment variations of investors against past variations of the IBOVESPA index. For robustness, we use 1-, 2-, 3-, 5-, and 30-day variations of the IBOVESPA index. Our regressions are at the investor level, which enables us to control for unobserved time-invariant characteristics of each Brazilian investor, which would otherwise be impractical in case we had aggregate data like most existing studies. Therein, we find that the mean-reversal technique better explains buy or sell operations in the Brazilian stock market during the period from 2016 to 2018. Our results corroborate the findings of our feature selection technique: short-term variations explain more buy or sell operations than long-term variations.

In Sections 4.2 and 4.3, we study the determinants that either soften or exacerbate the mean-reversal behavior of Brazilian investors by looking at the role of gender and level of schooling, respectively, of investors. These exercises connect with the existing literature on the influence of socioeconomic and biological features in shaping the behavior of economic agents.

4.1. Do Investors Use a Mean-Reversion or a Momentum Strategy in Their Buy and Sell Operations? To answer how investors respond to changes in the IBOVESPA index, we run the following econometric specification:

$$\Delta y_{it} = \alpha_i + \alpha_{t^*} + \beta \Delta \text{IBOVESPA}_t + \eta_{it}, \quad (4)$$

in which Δy_{it} is the portfolio volume variation of investor i at time t . There is a positive variation ($\Delta y_{it} > 0$) when investor i buys more stocks at time t and a negative variation ($\Delta y_{it} < 0$) when she sells. Alternatively, when the investor holds her investment over time, then ($\Delta y_{it} = 0$). The factor η_{it} is the standard error term.

Our coefficient of interest is β , which captures investors’ responses to variations of the IBOVESPA index, denoted as $\Delta \text{IBOVESPA}_t$. We test whether investors use a mean-reversal or momentum strategy as follows (we discard the hypothesis that investors’ buy and sell decisions are unrelated to variations of the IBOVESPA index because our feature selection technique identified past variations of the IBOVESPA index as the most relevant predictors of investor-specific investment variations):

- (i) If investors use a mean-reversal strategy, then increases in the IBOVESPA index—i.e., $\Delta \text{IBOVESPA}_t > 0$ —are followed by sell operations in such a way that the investment volume of investors, on average, decreases ($\Delta y_{it} < 0$). Therefore, a mean-reversal strategy is translated by a negative β coefficient ($\beta < 0$).
- (ii) If investors use a momentum strategy, then increases in the IBOVESPA index—i.e., $\Delta \text{IBOVESPA}_t > 0$ —are followed by buy operations in such a way that the investment volume of investors, on average, increases ($\Delta y_{it} > 0$). Therefore, a momentum strategy is translated by a positive β coefficient ($\beta > 0$).

As there is persistence of the past IBOVESPA index variations by construction, we test how investors’ investment volume respond to 1-, 2-, 3-, 5-, and 30-day variations of the IBOVESPA index in an *independent manner*. This empirical design strategy prevents standard errors to get overly inflated due to high pairwise correlation of these regressors.

The term α_i represents investor fixed effects and absorbs any nonobserved time-invariant characteristic of each investor in the sample. This mitigates potential omitted variables that could bias our results, such as investors’ skill, which is hard to measure. We should note that any omitted variable that is time variant would not be absorbed by the investor fixed effect. Therefore, while the introduction of such fixed effect mitigates omitted variable bias, it does not completely avoid it. For instance, if investors’ skill significantly increases over time, then we would have an omitted variable bias. Since our panel spans a relatively small time period—2016 to 2018—it is fair to assume that investors’ skill remains roughly constant. The

TABLE 2: Output from Regression (4). We ask how investors respond to changes in the IBOVESPA index. We only use changes rather than past averages because the former has greater prediction power as reported by our feature selection procedure. The dependent variable is the variation of portfolio investment volume of investor i at time t in the Brazilian stock market from the beginning of 2016 to the end of 2018. Regressors are 1- (1), 2- (2), 3- (3), 5- (4), and 30-day (5) IBOVESPA index variations. The panel is on a daily frequency basis. Following Petersen [50], we double-cluster standard errors at the investor and time levels. Significance levels: * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$.

Dependent variable	Investor portfolio volume variation (Δy_{it})				
	(1)	(2)	(3)	(4)	(5)
Regressor $\Delta \text{IBOVESPA}_t$ with					
1-day variation	-9.693*** (1.580)				
2-day variation		-4.656*** (1.160)			
3-day variation			-2.400** (0.964)		
5-day variation				-2.265*** (0.852)	
30-day variation					0.058 (0.680)
Fixed effects					
Investor	Yes	Yes	Yes	Yes	Yes
Month-year	Yes	Yes	Yes	Yes	Yes
Observations	356,172	355,796	355,419	354,588	343,592
R^2	0.037	0.036	0.036	0.035	0.033
Error clustering	Investor Time	Investor Time	Investor Time	Investor Time	Investor Time

term α_{it} connotes time-fixed effects at the year-month level, which absorbs any homogeneous time-variant effect, such as the Brazilian recession or month-wise exchange rate fluctuations. Since our panel frequency is on a daily basis, we cannot add a time fixed effect at the same frequency because our coefficient of interest— β —would get absorbed by the time fixed effects as it only varies across time. To prevent such problem, we use a less granular time fixed effects, namely, month-year.

Our data set contains 13,247 investors in a large and representative bank in Brazil and 610 time points. Due to this configuration, we follow Petersen [50] and double-cluster standard errors at the investor and time levels. This is a robust strategy that is important for panels with a large number of individuals and time points because it mitigates heteroscedasticity and serial correlation. Finally, our data is in percentage terms.

Table 2 reports our estimates of Regression (4). We observe that a 1 percent increase of the IBOVESPA index associates with an average decrease of 9.693% of the investor portfolio volume when we look at the 1-day IBOVESPA variation. The results remain with a statistically significant coefficient across different lengths of past IBOVESPA variations (2-, 3-, and 5-day variations), except for 30-day variations, in which the statistical significance vanishes. Moreover, the magnitude of the coefficient reduces as we use less recent past variations of the IBOVESPA index, which is consistent with the view that investors in our sample are more concerned with short-term rather long-term variations of the IBOVESPA index. The negative and statistically significant sign corroborates the hypothesis that investors use mean-reverting trading strategies, in which they tend to

sell after substantial upward changes of the IBOVESPA index, and tend to buy after downward changes.

4.2. Does Gender Impact Investors Responsiveness to IBOVESPA Index Changes? We have showed empirical evidence that investors' strategy, on average, better fit to a mean-reverting behavior in the Brazilian stock market. That is, they tend to sell after positive changes of the IBOVESPA index and buy after negative changes. In this section, we ask whether the sensitiveness of investors to the IBOVESPA index depends on their biological characteristics, in special their gender. Biological factors—especially gender—have been extensively explored in investment decision-making. Notable works relating biological factors, including gender, are Hira and Loibl [51]; Lundeberg et al. [34]; Neyse et al. [33]; and Sunden and Surette [52]. This paper provides further evidence of the existence of such gender gap in investment decisions using a microdata on investor-matched buy and sell operations.

In this line of research, Neyse et al. [33] and Lundeberg et al. [34] partly attribute behavioral differences among males and females due to systematic changes in overconfidence. Excessive overconfidence is associated with higher levels of testosterone, which is more pronounced in males. Overconfidence may induce investors to take on higher risks, leading them to look for higher returns in the short term. In this way, we would expect that females be less sensitive to changes of past IBOVESPA variations as they would value more fundamentals and look for yields in the longer term. Therefore, short-term variations of the IBOVESPA indices would explain less their buy or sell operations comparatively

TABLE 3: Output from Regression (5). We ask whether female investors have different sensitiveness with respect to their investment portfolio to IBOVESPA index changes. We only use changes rather than past averages because the former has greater prediction power as reported by our feature selection procedure. The dependent variable is the variation of portfolio investment volume of investor i at time t in the Brazilian stock market from the beginning of 2016 to the end of 2018. Regressors are 1- (1), 2- (2), 3- (3), 5- (4), and 30-day (5) IBOVESPA index variations, as well as their interaction with the investor's gender. The panel is on a daily frequency basis. Following Petersen [50], we double-cluster standard errors at the investor and time levels. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Dependent variable	Investor portfolio volume variation (Δy_{it})				
	(1)	(2)	(3)	(4)	(5)
Regressor $\Delta IBOVESPA_t$ with					
1-day variation	-10.345*** (1.754)				
2-day variation		-5.019*** (1.275)			
3-day variation			-2.650** (1.055)		
5-day variation				-2.315*** (0.886)	
30-day variation					-0.001 (0.693)
Interactions of $\Delta IBOVESPA_t$ with gender					
1-day variation · Female	6.543*** (2.392)				
2-day variation · Female		3.708*** (1.264)			
3-day variation · Female			2.585** (1.212)		
5-day variation · Female				0.517 (1.538)	
30-day variation · Female					0.572 (0.831)
Fixed effects					
Investor	Yes	Yes	Yes	Yes	Yes
Month-year	Yes	Yes	Yes	Yes	Yes
Observations	356,172	355,796	355,419	354,588	343,592
R^2	0.039	0.037	0.036	0.035	0.034
Error clustering	Investor Time	Investor Time	Investor Time	Investor Time	Investor Time

to males. To empirically answer this question, we construct the following econometric specification:

$$\Delta y_{it} = \alpha_i + \alpha_t + \beta_1 \Delta IBOVESPA_t + \beta_2 \Delta IBOVESPA_t \times \text{Female}_i + \eta_{it}, \quad (5)$$

in which Female_i is a dummy variable that takes the value of 1 when investor i is female and 0 otherwise. We do not add the investor's gender alone in (5) because it would be absorbed by the investor fixed effects α_i . Our coefficient of interest is β_2 , which captures any behavioral deviation of females to changes of the IBOVESPA index with respect to the average of the entire sample (male and female). If $\beta_2 > 0$, then the mean-reversal strategy is less pronounced to females, while $\beta_2 < 0$ indicates a more accentuated behavior towards the mean-reversal strategy. In the case $\beta_2 = 0$, then females and males respond, on average, equivalently to changes of the IBOVESPA index. Following the discussion on overconfidence and its influence on short-term decisions over males and females, our hypothesis is that $\beta_2 > 0$.

Table 3 reports our estimates of Regression (5). Our previous results relating the mean-reversal strategy of investors in the Brazilian stock market remain the same. We observe that the interaction of changes in the IBOVESPA index and the dummy female is positive and statistically significant. This empirical finding corroborates the view that female investors have a less pronounced mean-reversal strategy than males as they look at longer-term returns and are less attentive to short-term variations of the IBOVESPA index, which could arise due to noisy information. For instance, looking at Specification (1), a 1 percent positive change in the IBOVESPA index associates with a decrease of $-10.345 + 6.543 = -3.802\%$ of the invested volume of female investors. In contrast, the entire sample (males and females) decreases their portfolio volume, on average, by -10.345% for a 1 percent positive change in the IBOVESPA index. Interestingly, even though statistically insignificant, 30-day variations of the IBOVESPA index are positively associated with investment volumes for females, suggesting traits of a momentum strategy. This is also suggestive evidence that

TABLE 4: This table reports output from Regression (6). We ask whether investors with higher academic degree have different sensitiveness with respect to their investment portfolio to IBOVESPA index changes. We only use changes rather than past averages because the former has greater prediction power as reported by our feature selection procedure. The dependent variable is the variation of portfolio investment volume of investor i at time t in the Brazilian stock market from the beginning of 2016 to the end of 2018. Regressors are 1- (1), 2- (2), 3- (3), 5- (4), and 30-day (5) IBOVESPA index variations, as well as their interaction with the investor's academic degree. The panel is on a daily frequency basis. Following Petersen [50]; we double-cluster standard errors at the investor and time levels. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Dependent variable	Investor portfolio volume variation (Δy_{it})				
	(1)	(2)	(3)	(4)	(5)
Regressor $\Delta\text{IBOVESPA}_t$ with					
1-day variation	-10.347*** (1.795)				
2-day variation		-5.136*** (1.299)			
3-day variation			-2.750** (1.076)		
5-day variation				-2.565*** (0.931)	
30-day variation					0.085 (0.695)
Interactions of $\Delta\text{IBOVESPA}_t$ with academic degree					
1-day variation · Higher education	5.347*** (1.520)				
2-day variation · Higher education		3.915*** (1.040)			
3-day variation · Higher education			2.864* (1.573)		
5-day variation · Higher education				2.471* (1.398)	
30-day variation · Higher education					-0.237 (0.647)
Fixed effects					
Investor	Yes	Yes	Yes	Yes	Yes
Month-year	Yes	Yes	Yes	Yes	Yes
Observations	356,172	355,796	355,419	354,588	343,592
R^2	0.038	0.036	0.036	0.035	0.035
Error clustering	Investor Time	Investor Time	Investor Time	Investor Time	Investor Time

females tend to look at longer horizons when taking investment decisions.

4.3. Does Formal Education Impact Investors Responsiveness to IBOVESPA Index Changes? In this section, we look at how formal education (academic degree or level of schooling) can influence investors' sensitiveness to IBOVESPA index variations. There are several works in the behavioral finance literature that attempt to establish a connection between level of schooling and investors' awareness of stock markets and their decision-making determinants. We highlight the research studies of Grinblatt et al. [53] and Guiso and Jappelli [54]. In theory, educated investors should behave in more rational ways and trade less frequently when there is no new relevant information arriving in the market but noises. Therefore, we would expect these investors to have a smaller reaction to price fluctuations as they are able to better identify information from noise. To empirically test this behavior, we run the following specification:

$$\Delta y_{it} = \alpha_i + \alpha_{t*} + \beta_1 \Delta\text{IBOVESPA}_t + \beta_2 \Delta\text{IBOVESPA}_t \times \text{Higher Education}_i + \eta_{it}, \quad (6)$$

in which $\text{Higher Education}_i$ is a dummy variable that takes the value of 1 when investor i has a higher education (at least college degree) and 0 otherwise (high school or a lower degree). Our coefficient of interest is β_2 , which captures any behavioral deviation of investors with higher formal education to changes of the IBOVESPA index with respect to the average of the entire sample. The hypothesis is that $\beta_2 < 0$, in which more educated investors tend to better discern information from noise out of variations of the IBOVESPA index and therefore the mean-reversal strategy would be less pronounced.

Table 4 reports our estimates of Regression (6). On average, the mean-reversal strategy remains. We note that the interaction of changes of the IBOVESPA index and the dummy higher education is positive and statistically significant. This suggests that investors with higher academic

degree have a less pronounced mean-reversal strategy than less educated investors, which corroborates our hypothesis. Looking at Specification (3), we observe a positive, though marginally significant, relationship between IBOVESPA changes and investment volume ($-2.750 + 2.864 = 0.114$) for more educated investors, suggesting traits of a momentum strategy.

5. Conclusions

We employ machine learning techniques together with econometrics techniques to model investor behavior using a unique dataset for investors that focus on stock market investments. We propose a methodological approach to link machine learning methods widely used in computer science to standard econometric techniques commonly employed in social sciences.

Using the unique data set with high-frequency daily investment decision of a broad set of investors in Brazil, we provide evidence that investors look at past performance of a benchmark stock index in order to decide their investment decisions. Investors seem to prefer mean-reverting strategies in the short-run, rather than momentum. This may be associated with the disposition effect - investors prefer to sell the winners and buy the losers [55, 56]. Furthermore, research could exploit alternative explanations for this behavior.

In addition, we study the determinants that either soften or exacerbate the mean-reversal behavior of Brazilian investors by looking at the role of gender and level of schooling. We find that females and more educated investors are less sensitive to changes of past IBOVESPA variations, which is consistent with the literature on behavioral finance.

This paper highlights the importance of using non-traditional methods in econometric analysis. The use of machine learning methods permits us to automate the often subjective process of choosing which variables are important in any econometric analysis. By using a feature selection scheme—such as the elastic net in this paper—we are able to identify those attributes that best describe how investors decide to buy or sell their positions in an objective and statistically correct manner. In addition to that, the business specialist can always assess these variables pointed out as most important to analyze their economic meaning.

Data Availability

The data is confidential.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

Thiago C. Silva (grant no. 408546/2018-2) and Benjamin M. Tabak (grant no. 310541/2018-2, 425123/2018-9) gratefully acknowledge financial support from the CNPq foundation.

References

- [1] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.
- [2] E. F. Fama and K. R. French, "Permanent and temporary components of stock prices," *Journal of Political Economy*, vol. 96, no. 2, pp. 246–273, 1988.
- [3] A. W. Lo and A. C. MacKinlay, "Stock market prices do not follow random walks: evidence from a simple specification test," *Review of Financial Studies*, vol. 1, no. 1, pp. 41–66, 1988.
- [4] J. M. Poterba and L. H. Summers, "Mean reversion in stock prices," *Journal of Financial Economics*, vol. 22, no. 1, pp. 27–59, 1988.
- [5] J. Conrad and G. Kaul, "An anatomy of trading strategies," *Review of Financial Studies*, vol. 11, no. 3, pp. 489–519, 1998.
- [6] C. Engel and J. D. Hamilton, "Long swings in the dollar: are they in the data and do markets know it?," *The American Economic Review*, vol. 80, pp. 689–713, 1990.
- [7] N. Jegadeesh and S. Titman, "Momentum," *Annual Review of Financial Economics*, vol. 3, no. 1, pp. 493–509, 2011.
- [8] R. S. J. Koijen, J. C. Rodríguez, and A. Sbuelz, "Momentum and mean reversion in strategic asset allocation," *Management Science*, vol. 55, no. 7, pp. 1199–1213, 2009.
- [9] M. Morrin, J. Jacoby, G. V. Johar, X. He, A. Kuss, and D. Mazursky, "Taking stock of stockbrokers: exploring momentum versus contrarian investor strategies and profiles," *Journal of Consumer Research*, vol. 29, no. 2, pp. 188–198, 2002.
- [10] J. Okunev and D. White, "Do momentum-based strategies still work in foreign currency markets?," *The Journal of Financial and Quantitative Analysis*, vol. 38, no. 2, pp. 425–447, 2003.
- [11] D. Schiereck, W. De Bondt, and M. Weber, "Contrarian and momentum strategies in Germany," *Financial Analysts Journal*, vol. 55, no. 6, pp. 104–116, 1999.
- [12] D. O. Cajueiro and B. M. Tabak, "Testing for predictability in equity returns for European transition markets," *Economic Systems*, vol. 30, no. 1, pp. 56–78, 2006.
- [13] D. O. Cajueiro and B. M. Tabak, "Testing for time-varying long-range dependence in real state equity returns," *Chaos, Solitons & Fractals*, vol. 38, no. 1, pp. 293–307, 2008.
- [14] E. J. Chang, E. J. A. Lima, and B. M. Tabak, "Testing for predictability in emerging equity markets," *Emerging Markets Review*, vol. 5, no. 3, pp. 295–316, 2004.
- [15] A. Sensoy, K. Ozturk, E. Hacıhasanoglu, and B. M. Tabak, "Not all emerging markets are the same: a classification approach with correlation based networks," *Journal of Financial Stability*, vol. 33, pp. 163–186, 2017.
- [16] B. M. Tabak and E. J. A. Lima, "Market efficiency of Brazilian exchange rate: evidence from variance ratio statistics and technical trading rules," *European Journal of Operational Research*, vol. 194, no. 3, pp. 814–820, 2009.
- [17] C. M. Boya, "From efficient markets to adaptive markets: evidence from the French stock exchange," *Research in International Business and Finance*, vol. 49, pp. 156–165, 2019.
- [18] R. Ding and P. Cheng, "Speculative trading, price pressure and overvaluation," *Journal of International Financial Markets, Institutions and Money*, vol. 21, no. 3, pp. 419–442, 2011.
- [19] E. Lee and N. Piqueira, "Behavioral biases of informed traders: evidence from insider trading on the 52-week high," *Journal of Empirical Finance*, vol. 52, pp. 56–75, 2019.
- [20] T.-Y. Pak and P. Babiarz, "Does cognitive aging affect portfolio choice?," *Journal of Economic Psychology*, vol. 66, pp. 1–12, 2018.

- [21] T. Suzuki and Y. Ohkura, "Financial technical indicator based on chaotic bagging predictors for adaptive stock selection in Japanese and American markets," *Physica A: Statistical Mechanics and its Applications*, vol. 442, pp. 50–66, 2016.
- [22] A. Urquhart and F. McGroarty, "Are stock markets really efficient? evidence of the adaptive market hypothesis," *International Review of Financial Analysis*, vol. 47, pp. 39–49, 2016.
- [23] X. Xiong, Y. Meng, X. Li, and D. Shen, "An empirical analysis of the adaptive market hypothesis with calendar effects: evidence from China," *Finance Research Letters*, vol. 31, 2019.
- [24] H. Takahashi and T. Terano, "Analyzing the influence of overconfident investors on financial markets through agent-based model," in *Intelligent Data Engineering and Automated Learning-IDEAL 2007*, H. Yin, P. Tino, E. Corchado, W. Byrne, and X. Yao, Eds., Springer, Berlin, Heidelberg, Germany, 2007.
- [25] B. LeBaron, "Empirical regularities from interacting long- and short-memory investors in an agent-based stock market," *IEEE Transactions on Evolutionary Computation*, vol. 5, no. 5, pp. 442–455, 2001.
- [26] M. A. Bertella, F. R. Pires, L. Feng, and H. E. Stanley, "Confidence and the stock market: an agent-based approach," *PLoS One*, vol. 9, no. 1, Article ID e83488, 2014.
- [27] H. R. Varian, "Big data: new tricks for econometrics," *Journal of Economic Perspectives*, vol. 28, no. 2, pp. 3–28, 2014.
- [28] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: With Applications in R*, Springer Publishing Company, Incorporated, Switzerland, 2014.
- [29] T. C. Silva and L. Zhao, *Machine Learning in Complex Networks*, Springer Publishing Company, Incorporated, Switzerland, 1st edition, 2016.
- [30] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Springer, Berlin, Heidelberg, Germany, 2nd edition, 2009.
- [31] K. J. Arrow and G. Debreu, "Existence of an equilibrium for a competitive economy," *Econometrica*, vol. 22, no. 3, pp. 265–290, 1954.
- [32] H. A. Simon, *Models of Man: Social and Rational- Mathematical Essays on Rational Human Behavior in a Social Setting*, Wiley, Hoboken, NJ, USA, 1957.
- [33] L. Neyse, S. Bosworth, P. Ring, and U. Schmidt, "Overconfidence, incentives and digit ratio," *Scientific Reports*, vol. 6, no. 1, 2016.
- [34] M. A. Lundeborg, P. W. Fox, and J. Pun coha, "Highly confident but wrong: gender differences and similarities in confidence judgments," *Journal of Educational Psychology*, vol. 86, no. 1, pp. 114–121, 1994.
- [35] O. Onishchenko and N. Ülkü, "Foreign investor trading behavior has evolved," *Journal of Multinational Financial Management*, vol. 51, pp. 98–115, 2019.
- [36] M. Abreu, "How biased is the behavior of the individual investor in warrants?," *Research in International Business and Finance*, vol. 47, pp. 139–149, 2019.
- [37] J.-C. Li, Y.-X. Li, N.-S. Tang, and D.-C. Mei, "The roles of mean residence time on herd behavior in a financial market," *Physica A: Statistical Mechanics and Its Applications*, vol. 462, pp. 350–357, 2016.
- [38] C. Liu and X. Li, "Media coverage and investor scare behavior diffusion," *Physica A: Statistical Mechanics and Its Applications*, vol. 527, p. 121398, 2019.
- [39] K. W. Park, S. H. Jeong, and J. Y. J. Oh, "Foreigners at the gate? foreign investor trading and the disposition effect of domestic individual investors," *The North American Journal of Economics and Finance*, vol. 49, pp. 165–180, 2019.
- [40] Y. Shi, Y.-r. Tang, and W. Long, "Sentiment contagion analysis of interacting investors: evidence from China's stock forum," *Physica A: Statistical Mechanics and its Applications*, vol. 523, pp. 246–259, 2019.
- [41] J. R. Wei, J. P. Huang, and P. M. Hui, "An agent-based model of stock markets incorporating momentum investors," *Physica A: Statistical Mechanics and its Applications*, vol. 392, no. 12, pp. 2728–2735, 2013.
- [42] R. J. Balvers and Y. Wu, "Momentum and mean reversion across national equity markets," *Journal of Empirical Finance*, vol. 13, no. 1, pp. 24–48, 2006.
- [43] D. B. Chaves and V. Viswanathan, "Momentum and mean-reversion in commodity spot and futures markets," *Journal of Commodity Markets*, vol. 3, no. 1, pp. 39–53, 2016.
- [44] B. M. Barber and T. Odean, "Boys will be boys: gender, overconfidence, and common stock investment," *The Quarterly Journal of Economics*, vol. 116, no. 1, pp. 261–292, 2001.
- [45] J. I. Peña, "Daily seasonalities and stock market reforms in Spain," *Applied Financial Economics*, vol. 5, no. 6, pp. 419–423, 1995.
- [46] T. C. Silva and L. Liang Zhao, "Network-based high level data classification," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, no. 6, pp. 954–970, 2012.
- [47] T. C. Silva and L. Liang Zhao, "Network-based stochastic semisupervised learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, no. 3, pp. 451–466, 2012.
- [48] T. C. Silva and L. Liang Zhao, "Stochastic competitive learning in complex networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, no. 3, pp. 385–398, 2012.
- [49] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*, Springer-Verlag, Berlin, Heidelberg, Germany, 2006.
- [50] M. A. Petersen, "Estimating standard errors in finance panel data sets: comparing approaches," *Review of Financial Studies*, vol. 22, no. 1, pp. 435–480, 2009.
- [51] T. K. Hira and C. Loibl, "Gender differences in investment behavior," in *Handbook of Consumer Finance Research*, J. J. Xiao, Ed., Springer, New York, NY, USA, 2008.
- [52] A. E. Sunden and B. J. Surette, "Gender differences in the allocation of assets in retirement savings plans," *The American Economic Review*, vol. 88, pp. 207–211, 1998.
- [53] M. Grinblatt, M. Keloharju, and J. Linnainmaa, "IQ and stock market participation," *The Journal of Finance*, vol. 66, no. 6, pp. 2121–2164, 2011.
- [54] L. Guiso and T. Jappelli, "Awareness and stock market participation," *Review of Finance*, vol. 9, no. 4, pp. 537–567, 2005.
- [55] N. Barberis and W. Xiong, "What drives the disposition effect? an analysis of a long-standing preference-based explanation," *The Journal of Finance*, vol. 64, no. 2, pp. 751–784, 2009.
- [56] H. Shefrin and M. Statman, "The disposition to sell winners too early and ride losers too long: theory and evidence," *The Journal of Finance*, vol. 40, no. 3, pp. 777–790, 1985.

Research Article

Public Procurement Announcements in Spain: Regulations, Data Analysis, and Award Price Estimator Using Machine Learning

Manuel J. García Rodríguez , **Vicente Rodríguez Montequín** ,
Francisco Ortega Fernández, and **Joaquín M. Villanueva Balsera** 

Project Engineering Area, University of Oviedo, Oviedo 33004, Spain

Correspondence should be addressed to Vicente Rodríguez Montequín; montequi@uniovi.es

Received 27 June 2019; Revised 13 September 2019; Accepted 27 September 2019; Published 14 November 2019

Guest Editor: Benjamin M. Tabak

Copyright © 2019 Manuel J. García Rodríguez et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The largest project managers and adjudicators of a country, both by number of projects and by cost, are public procurement agencies. Therefore, knowing and characterising public procurement announcements (tenders) is fundamental for managing public resources well. This article presents the case of public procurement in Spain, analysing a dataset from 2012 to 2018: 58,337 tenders with a cost of 31,426 million euros. Many studies of public procurement have been conducted globally or theoretically, but there is a dearth of data analysis, especially regarding Spain. A quantitative, graphical, and statistical description of the dataset is presented. Mainly, the analysis is of the relation between the award price and the bidding price. An award price estimator is proposed that uses the random forest regression method. A good estimator would be very useful and valuable for companies and public procurement agencies. It would be a key tool in their project management decision making. Finally, a similar analysis, employing a dataset from European countries, is presented to compare and generalise the results and conclusions. Hence, this is a novel study which fills a gap in the literature.

1. Introduction

Every year, public authorities in European countries spend around 14% of GDP on public procurement, about 1.9 trillion euros [1], which is the latest estimate (2017) not including spending by utility companies. Spain is also similar, which spends around 10% to 20% of GDP [2]. Public procurement is very important in sectors such as civil construction, energy, transport, defence, IT, or health services. Therefore, it is crucial to analyse the public procurement notices, also called requests for tenders or simply tenders, to understand their behaviour in terms of prices, bidding companies, duration of projects, types of work, etc.

The growing awareness of public procurement as an innovative policy tool has recently sparked the interest of both policy makers and researchers [3]. The open data associated with public procurement and other open government data initiatives [4] are increasing mainly due to the following factors:

- (i) Technological factors: software tools to manipulate big data and machine learning algorithms to analyse data (e.g., to make predictions) [5, 6].
- (ii) Bureaucratic factors: standardisation of contracting language e-procurement [7, 8] and the benefits of the digitalisation of public procurement agencies [9].
- (iii) Political factors: greater transparency in political decision making and design of methods of selecting suppliers for public procurement [10].
- (iv) Economic factors: accuracy of the estimation of the cost [11], contract renegotiation [12], risk and uncertainty in the contracts [13], estimation of bidder participation in tenders [14] and its impact on prices [15], and globalisation—companies competing in markets far away from their origin [1].
- (v) Social factors: less tolerance for inefficient political management or political irregularities in the procedure [16] and greater transparency and flexibility

in award mechanisms between public procurement agencies and private companies [17].

The layout of this paper is connected with the method employed in the research, as depicted in Figure 1. Section 2 summarises the legislation regarding public procurement notices. A tender is organised in fields, but nevertheless, it is necessary to preprocess the information to produce the dataset. The data fields involved in the process as well as how the data are preprocessed are described. Section 3 analyses the dataset (main characteristic values, correlation, dispersion, etc.), lists the evaluation metrics used (types of errors), and makes a quantitative and graphical analysis of two fundamental fields: the tender price and the award price. The competition in public tenders and its impact on savings have been analysed: how the award price is affected by the competitiveness of the companies. In Section 4, an estimator of the award price is proposed using the machine learning algorithm random forest for regression. Several fields of the tender (the name of the public procurement agency, type of contract, geographical location, type of work or service, duration, date, etc.) have been used to make the prediction. The success of the estimator is analysed based on the evaluation metrics defined previously. Furthermore, a similar analysis employing a dataset from other European countries is presented. Lastly, some concluding remarks and avenues for future research are presented in Section 5.

As far as we know, this article is the first attempt to provide an award price estimator for all types of tenders in a country using machine learning algorithms. Similar articles dealing with this topic [18, 19] have been published recently but only for construction projects and small datasets. It is typical to find literature only applied to construction projects; this is mainly because they are the biggest public procurement projects. On the contrary, the approach of this article is from a multidisciplinary perspective, and it analyses a large volume of data using machine learning techniques.

2. Spanish Public Tenders (2012–2018): Description of the Dataset

In this section, the origin and nature of the Spanish public procurement processes are analysed. Section 2.1 presents a summary of the legislation associated with public procurement and the reuse of public information. Section 2.2 lists the fields of the public procurement notice with information that appears in the announcement. Section 2.3 explains how the original information has been preprocessed to finally obtain a dataset which is valid for statistical and mathematical analysis.

2.1. European and Spanish Legislation on Public Procurement and on the Reuse of Public Information. At the European and Spanish levels, laws have been developed related to the reuse of public sector information and procurement or contracting in the public sector. They are summarised in Table 1. According to *Spanish Law 20/2013*, the website of the Public Sector Contracting Platform (P.S.C.P.) of Spain has to publish the public procurement notices and their resolutions

of all contracting agencies belonging to the Spanish Public Sector.

With regard to official announcements of Spanish tenders outside Spain, Article 135 of *Law 9/2017* establishes that when tenders are subject to harmonised regulations (those with an amount greater than a threshold or with certain characteristics, stipulated in Articles 19 to 23), tenders have to also be published in *The Official Journal of the European Union* (OJEU) [20]. When the public contracting authority considers it appropriate, tenders not subject to harmonised regulations can be announced in the OJEU. The Europe Union (EU) has an Open Data Portal [21] which was set up in 2012, following *Commission Decision 2011/833/EU* on the reuse of commission documents. All EU institutions are invited to make their data publicly available whenever possible.

Furthermore, there is a portal called Tenders Electronic Daily (TED) [22] dedicated to European public procurement. It provides free access to business opportunities in the EU, the European Economic Area, and beyond.

2.2. Data Fields of Spanish Public Procurement Notices. The information of public procurement notices is defined in *Spanish Law 9/2017*, Annex III “Information that has to appear in the announcements.” P.S.C.P. has an open data section for the reuse of this information (in compliance with the publicity obligations established in *Law 9/2017*) which will be used in this article to generate the dataset. The information is provided by the Ministry of Finance (link in the Data Availability section) and has been published as open data since 2012 and updated monthly in XML format.

The fields of the public procurement notices are numerous, and they can completely define the tender. The most important fields are as follows (more details in Table 2):

- (i) Announcement fields: tender status, contract file number, object of the contract, tender price (budget), duration of the contract, CPV classification, contract type, contract subtype, place of execution, lots, type of procedure, contracting system, type of processing, contracting body, place and deadline for submission of tenders, participation requirements, award criteria, subcontracting conditions, contract modifications, etc.
- (ii) Award fields: award result, identity of the winning company (CIF and company name), award price, number of received offers, maximum and minimum received bids, etc.

Not all fields have been selected (last column in Table 2) to mathematically analyse the tenders for several reasons:

- (1) Some fields are usually empty or have inconsistent data or errors.
- (2) Not all fields have the same importance. For example, the tender price is more important than the language of the tender document.
- (3) The content of many of these fields is textual, which makes their mathematical modelling very complex.

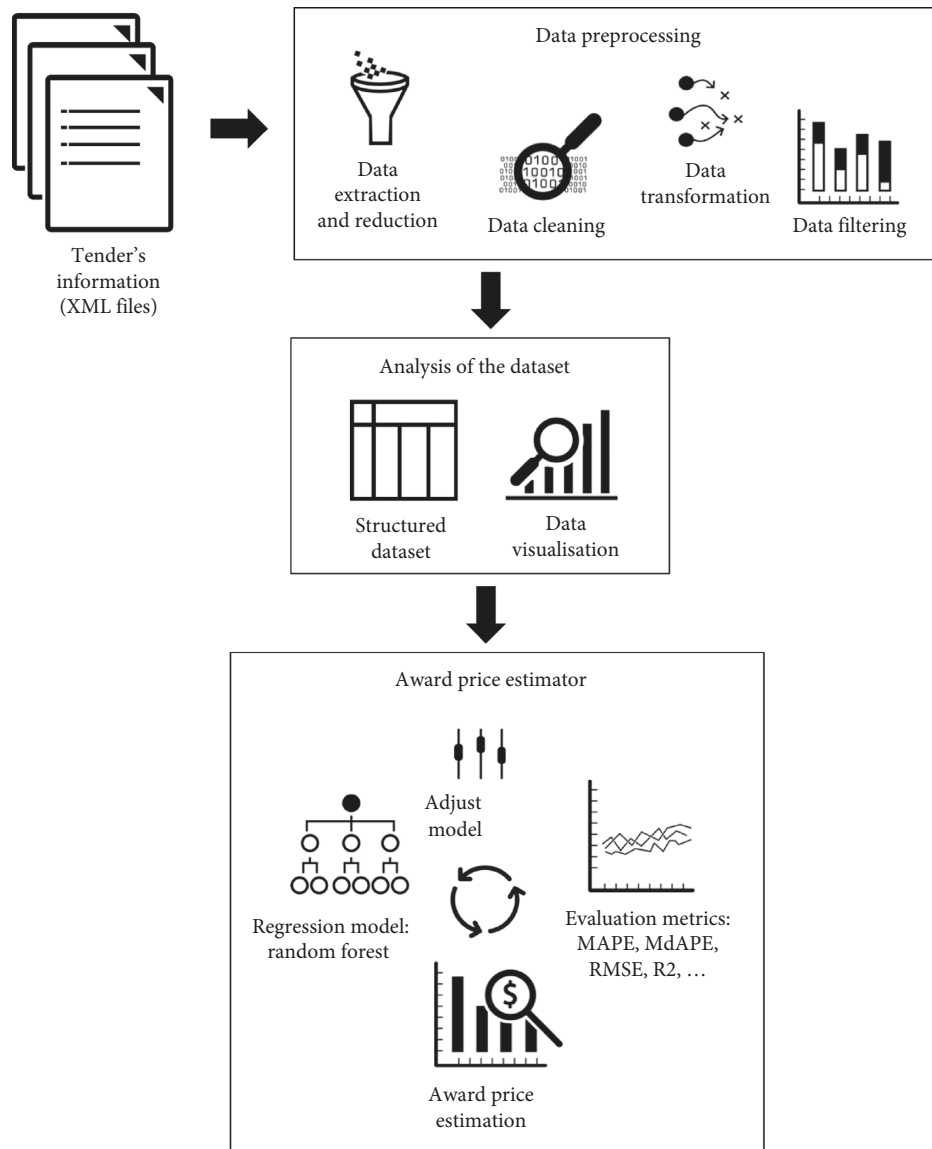


FIGURE 1: Flowchart of the data analysis and award price estimator.

2.3. Data Preprocessing. It is necessary to carry out several steps to preprocess the data. This is a laborious task because the tender's information has not been verified automatically to correct human errors. The preprocessing can be divided into the following 5 consecutive tasks:

- (1) **Data Extraction.** Structured data are stored in text files (XML format). A script has been created to read the fields recursively, saving in the database one tender per row and as many columns as there are fields to be stored.
- (2) **Data Reduction.** Around 60 fields are selected; a priori they are interesting for the performance of a statistical and mathematical analysis.
- (3) **Data Cleaning.** The data are cleaned. For example, deleting spaces, punctuation marks, and special characters, conversion to capital letters, deleting data

with fixed structure (postal code, CPV, CIF, etc.) which do not obey the structure's rules, etc.

- (4) **Data Transformation.** Basically, four types of transformations are carried out:
 - (a) **Normalisation.** This consists of homogenising the fields. For example, converting dates to time stamps.
 - (b) **Aggregation.** This consists of adding new useful fields for the analysis. For example, creating a new field which is the first two numbers of the CPV classification (common procurement vocabulary).
 - (c) **Data Enhancement.** It serves to create fields with external information and thus enables checking the consistency of the extracted data. For example, employing the postal code of the tender, it has generated its geographical

TABLE 1: Laws about public procurement and the reuse of public sector information.

Law	Description	Level	Permanent link
<i>Directive 2003/98/EC</i>	Reuse of public sector information	Europe	http://data.europa.eu/eli/dir/2003/98/oj
<i>Directive 2013/37/EU</i>	Modifying previous directive 2003/98/EC	Europe	http://data.europa.eu/eli/dir/2013/37/oj
<i>Directive 2007/2/EC</i>	Establishing an Infrastructure for Spatial Information in the European Community (INSPIRE)	Europe	http://data.europa.eu/eli/dir/2007/2/oj
<i>Law 37/2007</i>	Transposing into Spanish law the European directive 2003/98/EC	Spain	https://boe.es/eli/es/l/2007/11/16/37
<i>Royal Decree-Law 1495/2011</i>	Developing the Spanish law 37/2007	Spain	https://boe.es/eli/es/rd/2011/10/24/1495
<i>Commission Decision 2011/833/EU</i>	On the reuse of commission documents	Europe	http://data.europa.eu/eli/dec/2011/833
<i>Law 19/2013</i>	Transparency, access to public sector information and good governance	Spain	https://boe.es/eli/es/l/2013/12/09/19
<i>Law 20/2013</i>	Market unit guarantee	Spain	https://boe.es/eli/es/l/2013/12/09/20
<i>Law 18/2015</i>	Transposing into Spanish law the European directive 2013/37/EU	Spain	https://boe.es/eli/es/l/2015/07/09/18
<i>Directive 2014/23/EU</i>	Award of concession contracts	Europe	http://data.europa.eu/eli/dir/2014/23/oj
<i>Directive 2014/24/EU</i>	Public procurement	Europe	http://data.europa.eu/eli/dir/2014/24/oj
<i>Law 9/2017</i>	Transposing into Spanish law the previous European directives 2014/23/UE and 2014/24/UE	Spain	https://boe.es/eli/es/l/2017/11/08/9

TABLE 2: Most relevant data fields in the public procurement notices (tenders) used in the dataset

Name	Description	Name column dataset
Tender status	Status of the tender during the development of the procedure: prior notice, in time, pending adjudication, awarded, resolved or cancelled	Not used (similar to Result_code)
Contract file number	Unique identifier for a contract file	Not used
Object of the contract	Summary description of the contract	Not used (unstructured textual information)
Public procurement agency	Public procurement agency that made the tender: name, identifier (NIF or DIR3), website, address, postal code, city, country, contact name, telephone, fax, e-mail, etc	Name_Organisation Postalzone Postalzone_CCAA Postalzone_Province Postalzone_Municipality
Tender price	Amount of bidding budgeted (taxes included)	Tender_Price
Duration	Time (days) to execute the contract	Duration
CPV classification	CPV (common procurement vocabulary) is a European system for classifying the type of work in public contracts defined in the Commission Regulation (EC) No 213/2008: http://data.europa.eu/eli/reg/2008/213/oj The numerical code consists of 8 digits, subdivided into divisions (first 2 digits of the code), groups (first 3 digits), classes (first 4 digits), and categories (first 5 digits)	CPV CPV_Aggregated (first 2 digits of the code)
Contract type	Type of contract defined by legislation (Law 9/2017): works, services, supplies, public works concession, works concession, public services management, services concession, public sector and private sector collaboration, special administrative, private, patrimonial, or others	Type_code
Contract subtype	Code to indicate a subtype of contract. If it is a type of service contract: based upon the 2004/18/CE Directive, Annex II. If it is a type of work contract: works contract codes defined by the Spanish DGPE	Subtype_code

TABLE 2: Continued.

Name	Description	Name column dataset
Contract execution place	Contract's execution has a place through the Nomenclature of Statistical Territorial Units (NUTS), created by Eurostat [23]	Not used (assumed equal to Postalzone)
Type of procedure	Procedure by which the contracts was awarded: open, restricted, negotiated with advertising, negotiated without publicity, competitive dialogue, internal rules, derived from framework agreement, project contest, simplified open, association for innovation, derivative of association for innovation, based on a system dynamic acquisition, bidding with negotiation, or others	Procedure_code
Contracting system	The contracting system indicates whether it is a contract itself or a framework agreement or dynamic acquisition system	
Type of processing	Type of processing: ordinary, urgent, or emergency	Urgency_code
Award result	Type of results: awarded, formalised, desert, resignation, and withdrawal	Result_code
Winner identifier	Identifier of the winning bidder (called CIF in Spain) and its province (region)	CIF_Winner Winner_Province
Award price	Amount offered by the winning bidder of the contract (taxes included)	Award_Price
Date	Date of agreement in the award of the contract	Date
Number of received offers	Number of received offers (bidders participating) in each tender	Received_Offers

location (latitude and longitude), the municipality, the province, and the autonomous community.

- (d) *Conversion*. This consists of converting fields from one format to another. For example, conversions of text fields (strings) to a unique numeric identifier (integers) because the regression algorithm used only works with numeric variables: $\text{string}_1 \Rightarrow 1$, $\text{string}_2 \Rightarrow 2$, ..., $\text{string}_N \Rightarrow N$.
- (5) *Data Filtering*. The data are filtered to discard useless data for our analysis. Basically, this involves the following:
- Only formalised or awarded tenders are selected.
 - A tender is removed when it has one or several empty fields.
 - A tender is removed when it has an abnormally large positive price (award price or tender price) to remove outliers.
 - A tender which is formed by several different contracts (called lots) is removed. This is because it does not give the tender price for each contract, and this is a fundamental field for further analysis.

At first, there were 232,175 tenders. After data preprocessing, there were 58,337 tenders.

3. Statistical Analysis of the Dataset

In Section 3.1, a quantitative description of the dataset and a correlation analysis between fields of dataset are presented.

In Section 3.2, nine evaluation metrics are defined. In Section 3.3, they are used to calculate the error between two very important fields: tender price versus award price.

3.1. General Description. These data preprocessing operations prepare a structured and organised dataset ready for the data analysis. There are 58,337 tenders from 2012 to 2018 spread across Spain. Table 3 shows the quantitative description of the dataset: total numbers, means, medians, maximum, etc. The dataset has 19 fields or variables: 15 announcement fields and 4 award fields. Special emphasis is placed on Tender_Price and Award_Price. The amount is one of the most important variables in any project. Furthermore, the amount is fundamental in this article because an award price estimator is made.

Looking at Table 3, the following issues are observed:

- There are a lot of winning companies and bidding organisations. On average, each public procurement agency makes 16.46 tenders and each company wins 3.37 tenders.
- There is a great dispersion of prices (for both Tender_Price and Award_Price) looking at the median, the mean, and the maximum.
- There is a big difference between Tender_Price and Award_Price looking at the differences between both medians (€14,897) and means (€135,812.48). Therefore, it makes sense to propose a predictor of Award_Price because Tender_Price is not an accurate estimator.
- The 5 types of CPV with greater weight add up to 48.55% of the total number of tenders.

TABLE 3: Quantitative description of the dataset.

Topic	Description	Value
General values	Total number of tenders in the dataset	58,337
	Temporal range of tenders	2012/01/01–2018/12/28
	Total number of tendering organisations	3,544
	Total number of winning/award companies	17,305
	Mean number of offers received per tender	4.55
Dataset's variables	Mean duration of tender's works	382.21 days
	Input variables of tender's notice: Procedure_code, Urgency_code, Type_code, Subtype_code, Result_code, Name_Organisation, Postalzone, Postalzone_CCAA, Postalzone_Province, Postalzone_Municipality, Tender_Price, CPV, CPV_Aggregated, Duration, and Date	15 input variables (description in Table 2)
	Output variables of tender's resolution: Award_Price, Winner_Province, CIF_Winner, and Received_Offers	4 output variables (description in Table 2)
Tender price (taxes included)	Mean tender price	€538,707.39
	Median tender price	€86,715.00
	Maximum tender price	€3,196,970,000
	Aggregated tender price of all tenders	€31,426,572,936
Award price (taxes included)	Mean award price	€402,894.91
	Median award price	€71,818.00
	Maximum award price	€786,472,000
	Aggregated award price of all tenders	€23,503,680,419
Number of tenders by CPV	Tenders with CPV = 45: construction work	12,166 (20.85%)
	Tenders with CPV = 50: repair and maintenance services	5,174 (8.87%)
	Tenders with CPV = 79: business services (law, marketing, consulting, recruitment, printing, and security)	3,992 (6.84%)
	Tenders with CPV = 72: IT services (consulting, software development, Internet, and support)	3,725 (6.39%)
	Tenders with CPV = 34: transport equipment and auxiliary products to transportation	3,264 (5.60%)
Number of tenders by type code	Tenders with Type_code = 1: goods/supplies	17,876 (30.64%)
	Tenders with Type_code = 2: services	28,363 (48.62%)
	Tenders with Type_code = 3: works	12,008 (20.58%)

To obtain new relevant information through the variables, the Spearman correlation method was used; Figure 2 shows the Spearman correlation matrix (a symmetric matrix with respect to the diagonal). Among the three typical correlation methods (Pearson, Kendall, and Spearman), the Spearman correlation method is chosen because it evaluates the strength of a monotonic relationship between two variables. A monotonic function preserves order (increasing or decreasing). The Spearman correlation coefficient (r_s) is defined for a sample of size n , and the n raw scores X_i, Y_i are converted to ranks rg_{X_i}, rg_{Y_i} :

$$r_s = \frac{\text{cov}(rg_X, rg_Y)}{\sigma_{rg_X} \sigma_{rg_Y}}, \quad (1)$$

where $\text{cov}(rg_X, rg_Y)$ is the covariance of the rank variables and σ_{rg_X} and σ_{rg_Y} are the standard deviations of the rank variables.

Looking at Figure 2, the greatest correlations are the following:

- (i) Tender_Price vs. Award_Price (0.97): this high correlation is in accordance with common sense since high bids are associated with high awards and low bids with low awards.
- (ii) Type_code vs. Subtype_code (0.74): each type of contract has its associated subtypes of contract. This is the reason for the high correlation.
- (iii) Name_Organisation vs. Postalzone_Municipality (0.42): each public procurement agency has a location associated with a postal code.
- (iv) Type_code vs. CPV (0.38): each type of contract is usually used for certain types of works.
- (v) Procedure_code vs. Tender_Price (−0.38) and Award_Price (−0.36): each type of contract procedure tends to correspond to a range of bidding and adjudication amounts.
- (vi) CPV vs. Duration (0.34): each type of work is usually associated with a temporal range (duration) for its realisation.

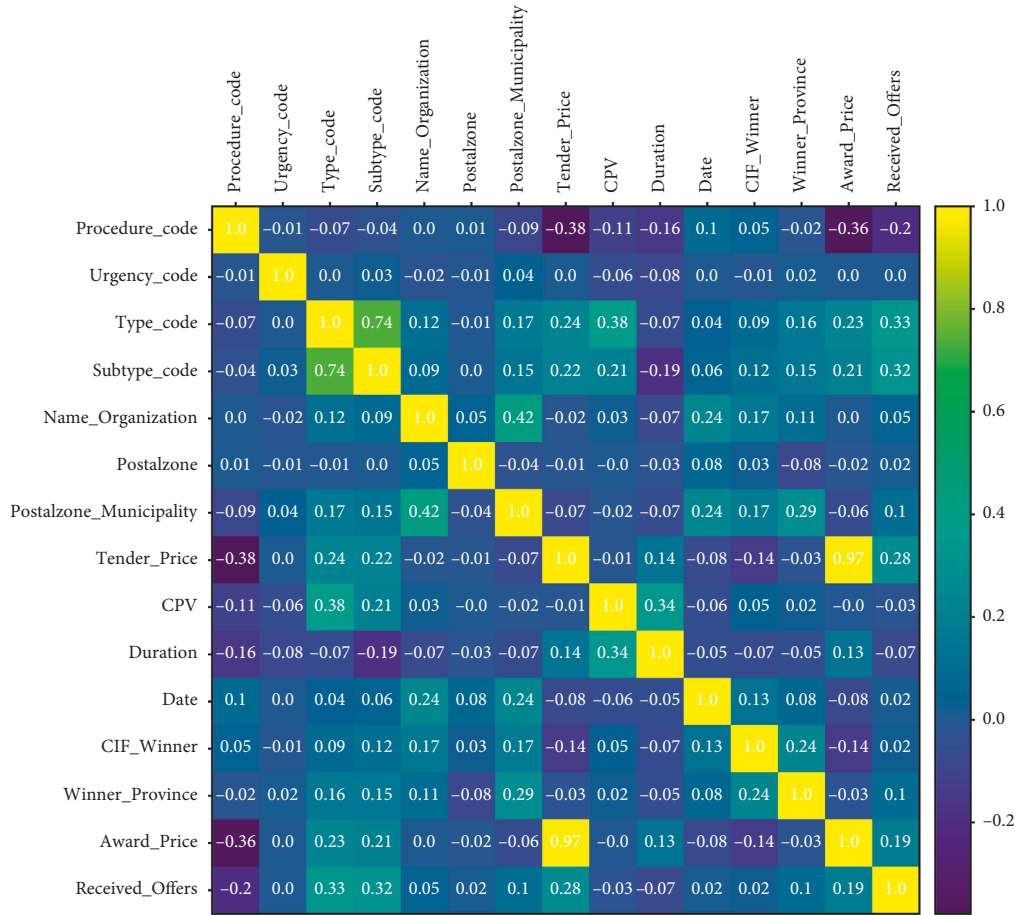


FIGURE 2: Correlation matrix between the variables of the dataset. Spearman's rank correlation coefficient is the method applied.

- (vii) Received_Offers vs. Type_code (0.33) and Subtype_code (0.32): the number of received offers by tender has a correlation with the type and subtype of the contract.
- (viii) Winner_Province vs. Postalzone_Municipality (0.29): there is a correlation between the origin (province) of the winning company and the location (municipality) of the tender. In general, tenders from a specific geographical region are won by companies from the same region. There are different socioeconomic reasons for this.

Higher correlation values have not been obtained due to the numerical form of expressing the information and the limitations of the correlation method (all methods have disadvantages). For example, Name_Organization and Postalzone_Municipality have a direct relation: an organisation usually has a unique assigned postal code. However, this relation can follow any mathematical pattern or function.

Another way to analyse the data is through the scatter matrix (see Figure 3) where the variables are plotted two by two and the matrix's diagonal is the probability density function of the corresponding variable. Although it cannot be appreciated in detail by the large amount of data and variables, the following relations are seen:

- (i) Procedure_code, Urgency_code, Type_code, and Subtype_code generate straight lines because they are variables with few values (they are codes) but have great dispersion when they are confronted with the rest of the variables.
- (ii) Name_Organisation, Postalzone, and Postalzone_Municipality have a large dispersion. In the probability density function of Postalzone, a great maximum is seen in Madrid's postal codes. This is because many tenders in Spain have been put forward by agencies located in the capital (Madrid).
- (iii) The CPVs show that some codes have high tender and award prices, a longer duration, and more received offers. This is true because each type of work has certain characteristics such as price, duration, or competence in the sector.
- (iv) The relation between Tender_Price and Award_Price will be analysed in detail later, but a certain relation can be seen. It had already appeared in the correlation matrix.

3.2. *Evaluation Metrics.* To compare the variables and calculate the errors or deviations of the prediction algorithms, first it is necessary to define some error metrics. The use of

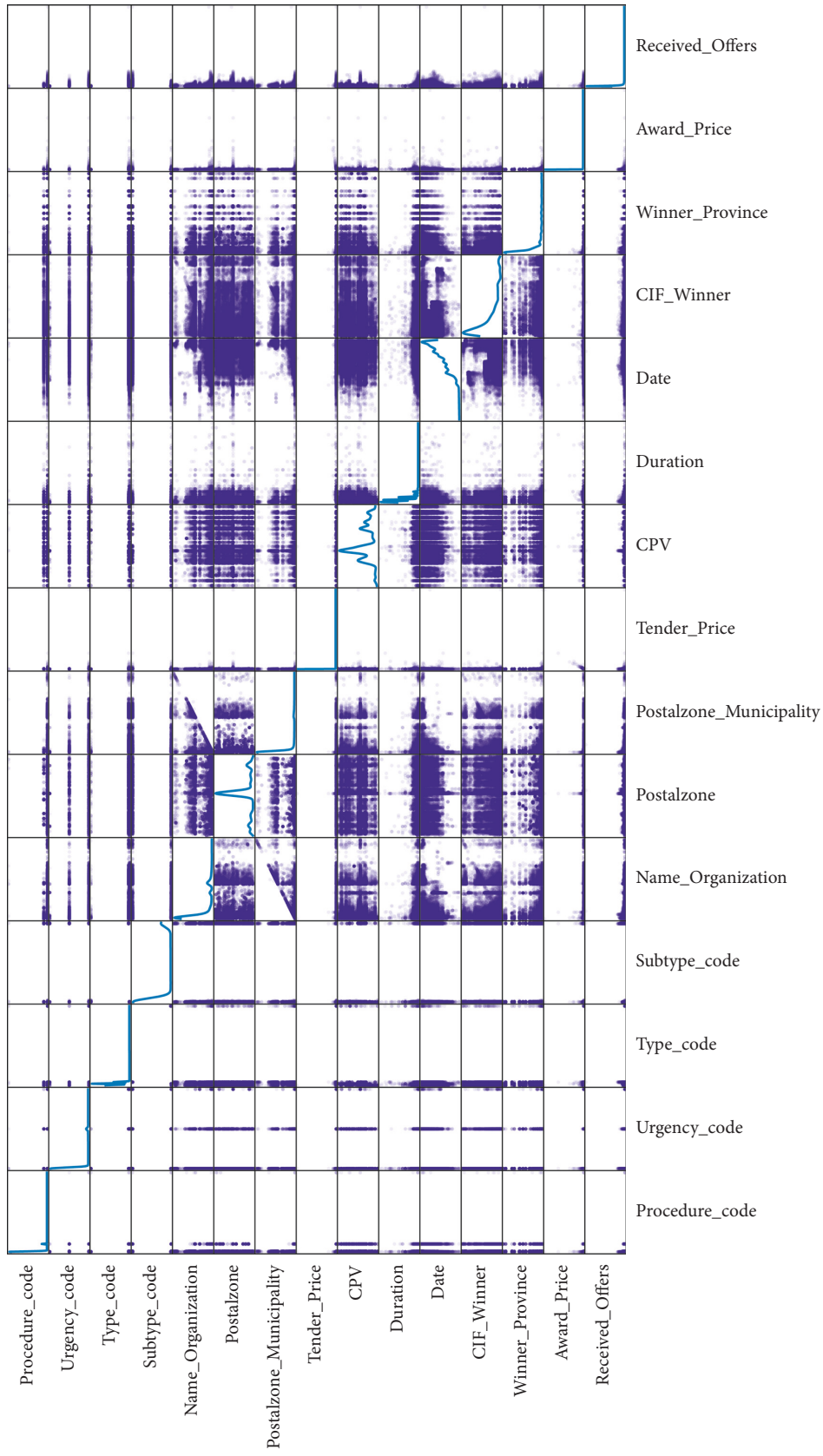
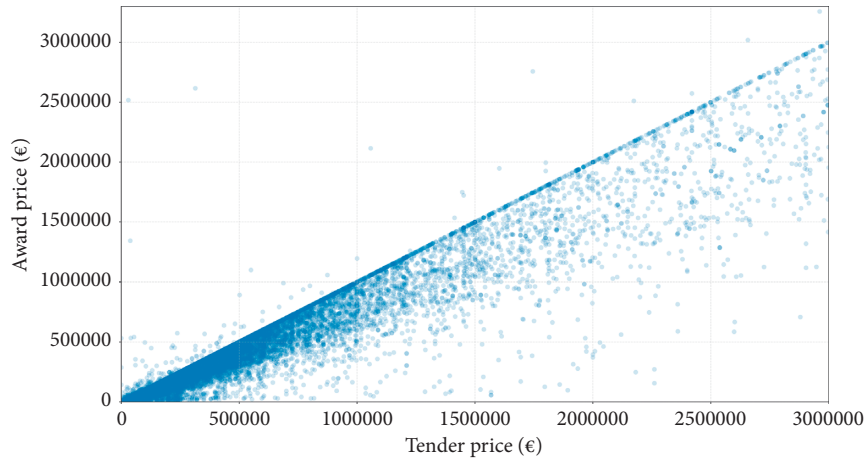
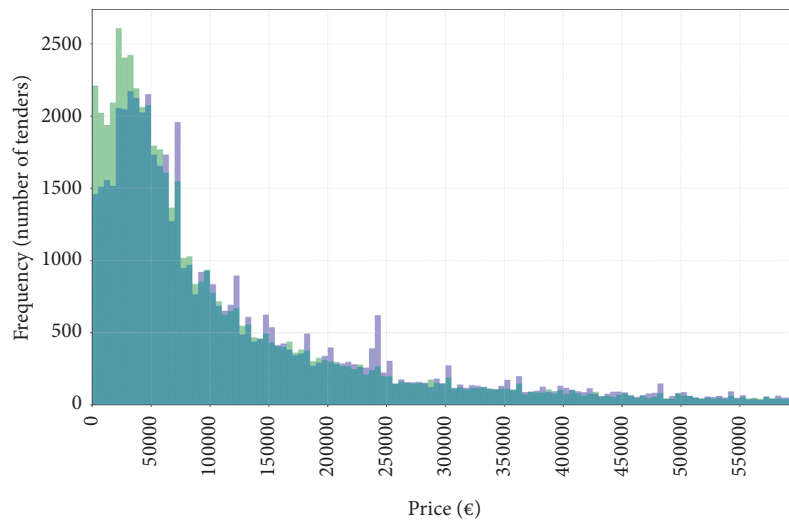


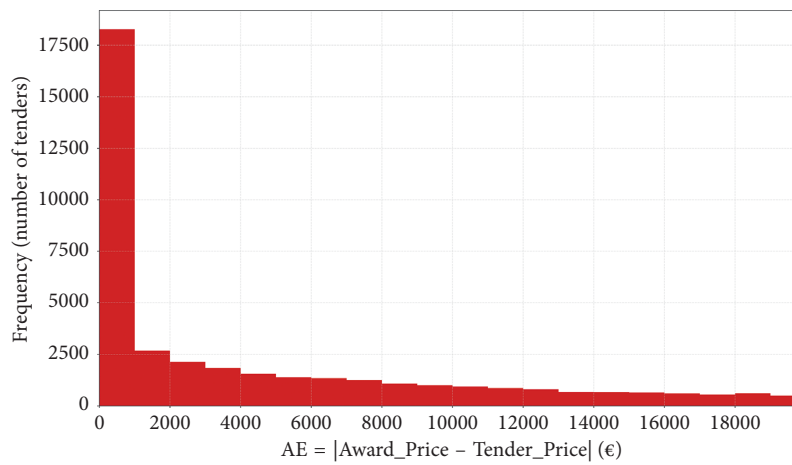
FIGURE 3: Scatter matrix between the variables of the dataset.



(a)



(b)



(c)

FIGURE 4: Relation between tender price and award price. (a) Scatter plot. (b) Histograms of frequency (number of tenders). (c) Absolute error (AE) histogram.

metrics based on medians and relative percentage is useful in this survey because the dataset has outliers of great weight, and the use of such metrics helps us to counteract the effect of these outliers.

Absolute error (AE), absolute percentage error (APE), mean absolute error (MAE), mean absolute percentage error (MAPE), median absolute error (MdAE), median absolute percentage error (MdAPE), root mean square error (RMSE), normalised root mean square error (NRMSE), and coefficient of determination (R^2) were selected as evaluation criteria (2)–(10): A_t is the actual value for period t , F_t is the expected or estimated value for period t , and n is the number of periods.

$$AE_t = |A_t - F_t|, \quad (2)$$

$$APE_t (\%) = 100 \left| \frac{AE_t}{A_t} \right| = 100 \left| \frac{A_t - F_t}{A_t} \right|, \quad (3)$$

$$MAE = \frac{1}{n} \sum_{t=1}^n AE_t = \frac{1}{n} \sum_{t=1}^n |A_t - F_t|, \quad (4)$$

$$MAPE (\%) = \frac{100}{n} \sum_{t=1}^n APE_t = \frac{100}{n} \sum_{t=1}^n \left| \frac{A_t - F_t}{A_t} \right|, \quad (5)$$

$$MdAE = \frac{1}{n} \text{median}(|A_1 - F_1|, |A_2 - F_2|, \dots, |A_n - F_n|), \quad (6)$$

$$MdAPE (\%) = \frac{100}{n} \text{median} \left(\left| \frac{A_1 - F_1}{A_1} \right|, \left| \frac{A_2 - F_2}{A_2} \right|, \dots, \left| \frac{A_n - F_n}{A_n} \right| \right), \quad (7)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{t=1}^n |A_t - F_t|^2}, \quad (8)$$

$$NRMSE = \frac{RMSE}{\max(A_t) - \min(A_t)} = \frac{\sqrt{(1/n) \sum_{t=1}^n |A_t - F_t|^2}}{\max(A_t) - \min(A_t)}, \quad (9)$$

$$R^2 = 1 - \frac{\sum_{t=1}^n |A_t - F_t|^2}{\sum_{t=1}^n |A_t - \bar{A}|^2}, \quad (10)$$

where \bar{A} is the mean: $\bar{A} = (1/n) \sum_{t=1}^n A_t$.

3.3. Tender Price vs. Award Price. Figure 4(a) shows graphically the variable tender price versus award price for all tenders when tender price is less than €3,000,000. This threshold is 3.5 times the median of tender price. A line at 45 degrees can be seen; its points satisfy the condition that tender price is equal to award price. Therefore, in this line there, is no error between the two variables, and so the

TABLE 4: Error metrics between tender price and award price.

Error	Value
Absolute error (AE)	See Figure 4(c)
Absolute percentage error (APE)	See Figure 5
Median absolute error (MdAE)	€6,955.00
Median absolute percentage error (MdAPE)	11.84%
Mean absolute error (MAE)	€137,778.64
Mean absolute percentage error (MAPE)	39.79%
Root mean square error (RMSE)	101,451,609,620,714
Coefficient of determination	-3.10

tender price would be a perfect estimator. Below this line, there is a large dispersion of points. When the distance between a point and the line is high, the error is also high. Finally, there are few points above the line. This is because only rarely is the award price higher than the tender price. This can happen due to special conditions of the contract or, alternatively, it can be wrong data. There is no information about how the public procurement agencies calculate the tender price or if it is validated before entering the dataset.

Figure 4(b) shows the frequency histogram of both variables. The frequency is the number of tenders for each bar of €5,000. For example, the most frequent range for the tender price is €30,000–€35,000; for the award price, it is €20,000–€25,000. Figure 4(c) shows the frequency histogram of the AE between both variables by ranges of €1,000. It can be observed that approximately 18,000 tenders (30% of the total) have less than €1,000 error. There is a big difference with the rest of the bars.

Table 4 presents the error metrics (or evaluation metrics) calculated between the variables tender price and award price for the entire dataset. An error between tender price and award price, in terms of project management, means that there is a budget deviation between the tender price and the price finally awarded.

An interesting analysis is how the award price is affected by the competitiveness of the companies (see Table 5). It is necessary to group the tenders according to the number of offers received. For this purpose, 4 groups have been created: no competitiveness (1 offer), low competitiveness (2–4 offers), medium competitiveness (5–10 offers), and high competitiveness (more than 10 offers). As competitiveness increases, the difference between the award price and tender price is greater because MdAE, MdAPE, MAE, and MAPE are greater. This shows that companies are more aggressive (bid lower prices) to win the tender. Consequently, the award price is lower in a scenario with less competitiveness or, in other words, public procurement agencies save money.

Figure 5 shows the APE boxplot grouped by CPV. Box diagrams are a standard method to graphically represent numerical data through their quartiles. The outliers of the dataset have not been represented because they are values very far out, which would make it difficult to scale the axes. MAPE (red colour) and MdAPE (green colour) for each CPV group are marked. The great differences of APE,

TABLE 5: Description of the dataset and the errors between tender price and award price by number of received offers.

Description	Groups by competitiveness			
	No competitiveness	Low Received offers (2-4)	Medium Received offers (5-10)	High Received offers >10
Total number of tenders in the dataset	18,790	22,714	11,553	5,271
Total number of tendering organisations	1,956	2,553	2,135	1,053
Total number of winning/award companies	7,550	9,555	5,222	2,402
Mean received offers by tender	1.0	2.80	6.73	20.01
Mean duration of tender's works	401.07 days	396.65 days	370.95 days	277.50 days
Mean tender price	€354,882.49	€388,526.27	€785,455.49	€1,301,031.70
Median tender price	€60,500.00	€75,000.00	€121,000.00	€254,376.00
Mean award price	€341,874.79	€323,611.87	€460,548.68	€836,188.79
Median award price	€58,984.50	€64,833.00	€90,689.00	€174,986.00
Median absolute error (MdAE)	€93.50	€7,661.50	€22,854.00	€76,420.00
Median absolute percentage error (MdAPE)	0.12%	13.39%	29.63%	45.94%
Mean absolute error (MAE)	€13,966.65	€68,244.60	€326,698.33	€464,907.75
Mean absolute percentage error (MAPE)	10.02%	25.65%	54.48%	77.98%

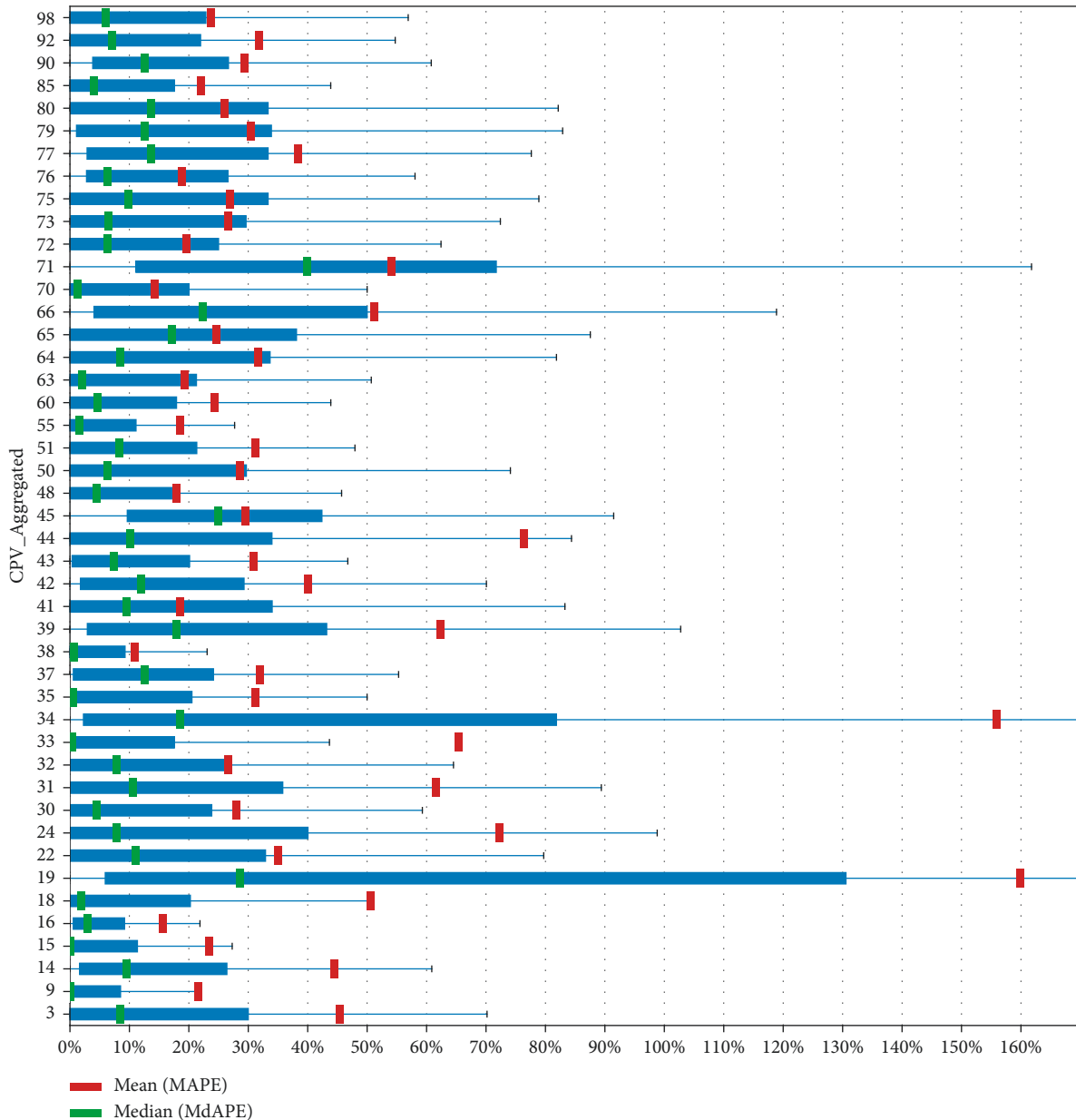


FIGURE 5: Boxplot of absolute percentage error (APE) between award price and tender price grouped by CPV.

MAPE, and MdAPE according to the CPV can be clearly seen. In general, MdAPE is between 20% and 40% and MAPE is higher than 40%. The total value of MAPE and MdAPE (without dividing by CPV) has already been calculated, as shown in Table 4.

In conclusion, in view of the graphical and quantitative results, it can be affirmed that tender price is a bad estimator of award price. Perhaps it is not excessively bad in median (11.84%) but it is so in mean (39.79%). This is certainly due to the high dispersion between both prices (as seen in Figure 4(a)). This is the reason to create an award price estimator in the following section.

4. Award Price Estimator

A good award price estimator would be very useful and valuable for companies and public procurement agencies. It would be a key tool in their project management decision making because it reduces the economic risks. Due to the complexity involved, machine learning techniques have been chosen to create the estimator, in particular, random forest. In Section 4.1, random forest for regression is presented, from the theoretical framework to its application to the Spanish tenders' dataset. In Section 4.2, the empirical results and analysis are presented, for example, the error metrics of the award price estimator created. In Section 4.3 a similar analysis is presented using a dataset from other countries, creating a new award price estimator.

4.1. Random Forest for Regression. Random forests (RF), introduced by Breiman [24] in 2001, is an ensemble learning method for regression or classification that operates by constructing a multitude of decision trees at training time and outputting the class which is the mode of the classes (classification) or mean prediction (regression) of the individual trees. It is a popular learning algorithm that offers excellent performance [25, 26], no overfitting [27], and a versatility of applicability to large-scale problems and in handling different types of data [25, 28]. It provides its own internal generalisation error estimate, called out-of-bag (OOB) error.

Simplified algorithm of RF for regression [29]:

- (1) For $b = 1$ to B :
 - (a) Draw a bootstrap sample Z^* of size N from the training data.
 - (b) Grow a random forest tree T_b to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size n_{\min} is reached.
 - (i) Select m variables at random from the p variables.
 - (ii) Pick the best variable/split-point among the m .
 - (iii) Split the node into two daughter nodes.
- (2) Output the ensemble of trees $\{T_b\}_1^B$.

To make a prediction at a new point x , $\hat{f}_{rf}^B(x) = (1/B)\sum_{b=1}^B T_b(x)$.

At each split in each tree, the improvement in the split criterion is the measure of the importance attributed to the splitting variable and is accumulated over all the trees in the forest separately for each variable. It is called variable importance [24].

There are other implementations of RF algorithms, such as Boruta [30], regularised random forest (RRF) [31], conditional forest [32], quantile regression forest (QRF) [33], or extremely randomised regression trees (extraTrees) [34]. The last one was tested with this dataset, but it has a worse accuracy than random forest, so finally it was discarded. The reason is because the function to measure the quality of a split is the Gini index, which is worse than MAE (mean absolute error) or MSE (mean squared error). A comparison between the use of MAE and MSE is shown in Figure 6 for 30 to 1000 trees generated in RF. MAE used as the quality function has clearly better values for the error metrics (especially MAPE and NRMSE) than the MSE quality function for this dataset. Therefore, the function selected is MAE.

The random forest method has been used for multiple and different real-world applications [25], such as the estimation of traffic car issues [35–37], wind speed prediction [38], classification of protein sequences [39], discrimination between seismic events and nuclear explosions [40], pedestrian detection [41], aggregated recommender systems [42], bed occupancy predictor in hospitals [43], classification of phishing e-mail [44], network intrusion detection [45], and employee turnover prediction [46].

Figure 7 shows different ratios between the training and testing subsets (train : test in percentage): 65 : 35, 70 : 30, 75 : 25, 80 : 20, 85 : 15, and 90 : 10. The most important errors for this study, MdAPE and MAPE, are constantly in the order of 9% and 30%, respectively. OBB and NMRSE do not change significantly. Hence, the train : test ratio is not relevant. The typical ratio 80 : 20 will be used in this article.

RandomForestRegressor from *Scikit-learn*, which is a machine learning library for the Python programming language, with 400 trees is the function used in this article. The 14 input variables used in RF are Tender_Price, Date, Duration, Name_Organisation, CPV, CPV_Aggregated, Procedure_code, Type_code, Subtype_code, Urgency_code, Postalzone, Postalzone_CCAA, Postalzone_Province, and Postalzone_Municipality. The variable to perform the regression is Award_Price, and the output generated by RF (prediction) will be called Forecast_Price.

This article does not use the other 3 variables of the tender's resolution (Winner_Province, CIF_Winner, and Received_Offers; Table 3) because they are not variables of the tender's notice. In a real scenario, the award price estimator only can use the variables of the tender's notice. However, if these 3 output variables are used in RF plus 14 input variables, the errors would decrease logically. This is demonstrated as shown in Figure 8: MdAPE is about 5% and MAPE 25%. MdAPE and MAPE are, respectively, 4% and 5% lower than the real scenario with only variables of the tender's notice (see Figure 7). The variable importances (RF

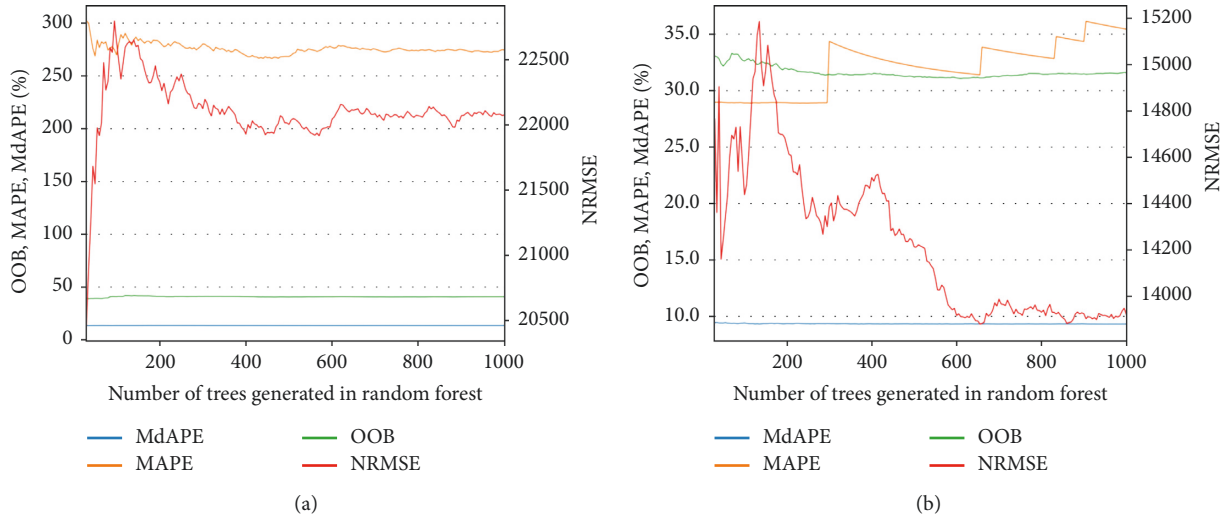


FIGURE 6: Relationship between trees in random forests (number of estimators) and error metrics (MdAPE, MAPE, OOB, and NRMSE) for two functions to measure the quality of a split. (a) The quality function is mean squared error (MSE). (b) The quality function is mean absolute error (MAE).

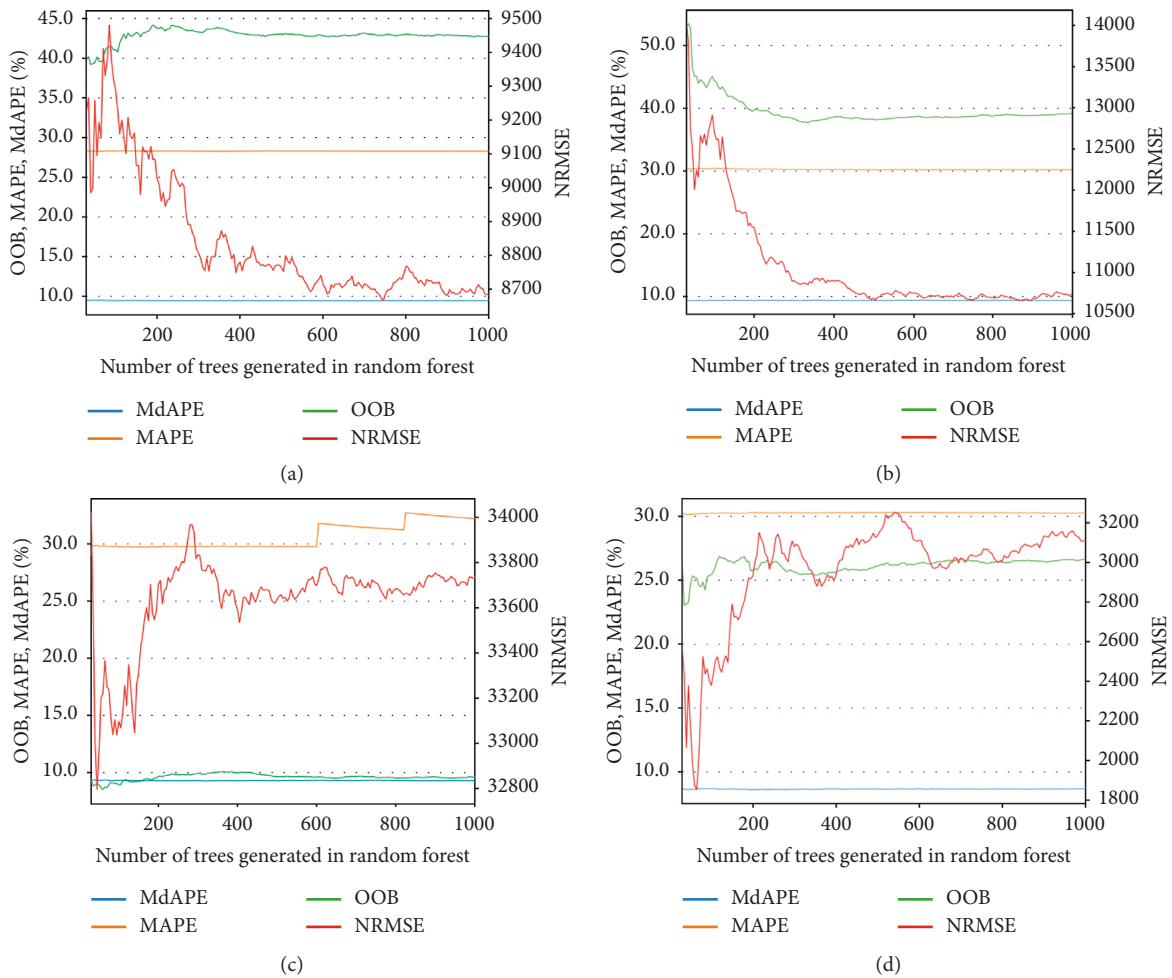


FIGURE 7: Continued.

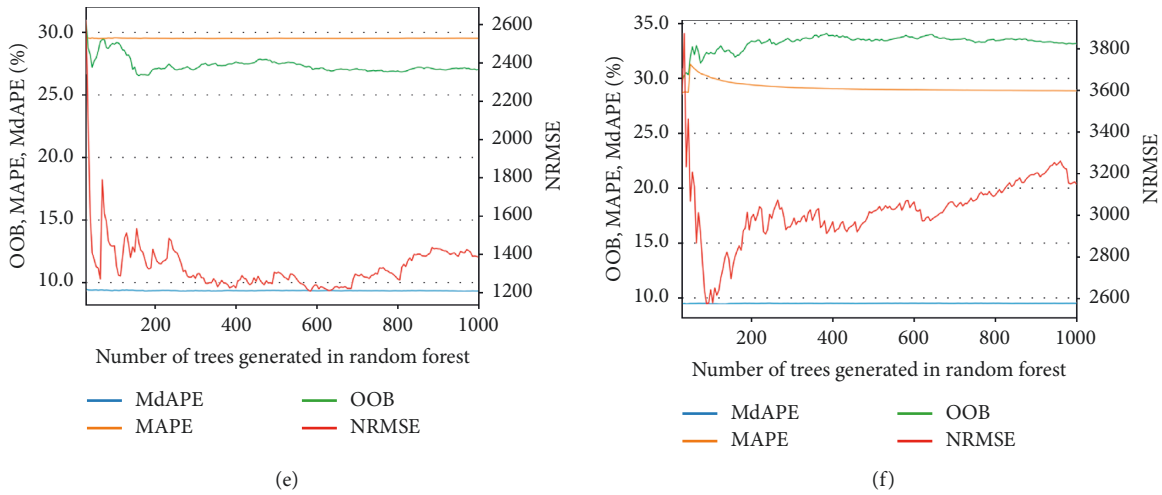


FIGURE 7: Relationship between trees in random forests and error metrics (MdAPE, MAPE, OOB, and NRMSE) for different ratios of training and testing subsets. (a) 65 : 35. (b) 70 : 30. (c) 75 : 25. (d) 80 : 20. (e) 85 : 15. (f) 90 : 10.

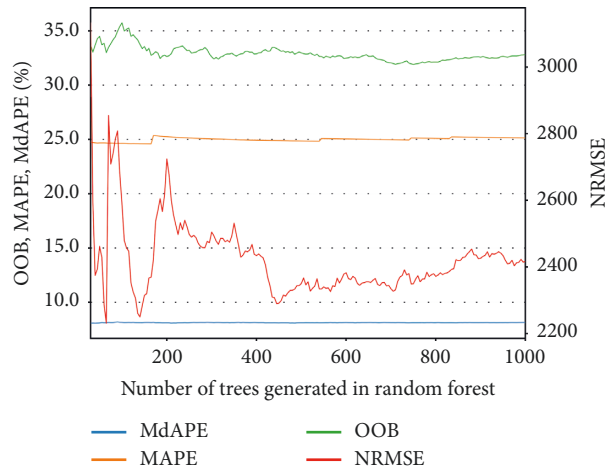


FIGURE 8: Relationship between trees in random forests and error metrics (MdAPE, MAPE, OOB, and NRMSE) using the 14 input variables plus 3 variables of the tender's resolution.

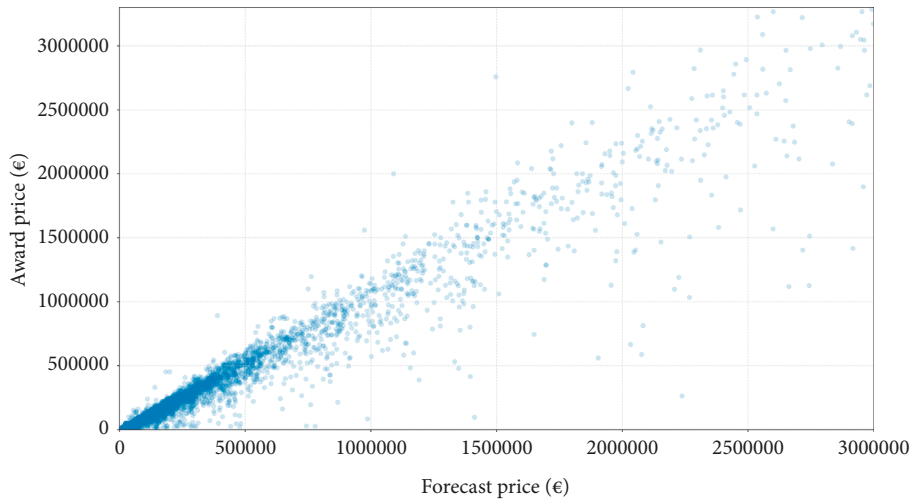


FIGURE 9: Scatter plot between forecast price and award price.

TABLE 6: Error metrics between award price and forecast price.

Error	Value	Difference with respect to Tender_Price
Absolute percentage error (APE)	See Figure 8	See Figure 8
Median absolute error (MdAE)	€7,575.45	+€620.45
Median absolute percentage error (MdAPE)	9.26%	-2.58%
Mean absolute error (MAE)	€67,241.34	+€70,537.3
Mean absolute percentage error (MAPE)	28.60%	-11.19%
Root mean square error (RMSE)	364,901,707,583	-101,086,707,913,131
Coefficient of determination (R^2)	0.92	—

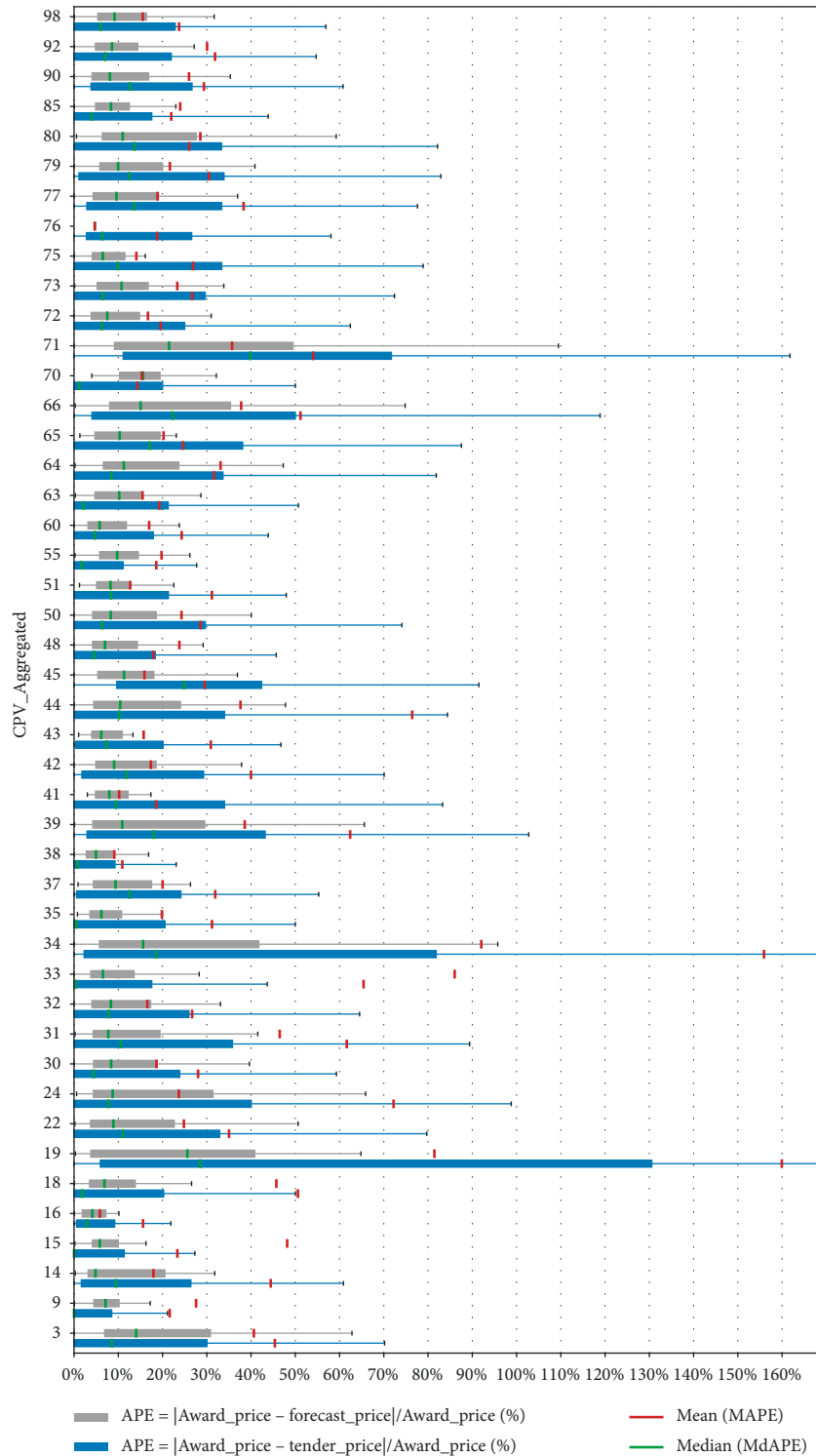


FIGURE 10: Boxplot of absolute percentage error (APE, grey colour) between award price and forecast price, grouped by CPV. The APE reference (blue colour) is the award price and tender price shown in Figure 5.

TABLE 7: European countries' dataset: quantitative description.

Topic	Description	Value
General values	Total number of tenders in the dataset	41,556
	Number of tenders by country: France (FR), Croatia (HR), Slovenia (SI), Bulgaria (BG), Germany (DE), Italy (IT), Hungary (HU), and Latvia (LV)	12,449 (FR); 7,910 (HR); 6,473 (SI); 6,096 (BG); 3,918 (DE); 3,782 (IT); 3,724 (HU); 1,736 (LV)
	Temporal range of tenders	2016/12/22–2017/12/29
	Total number of tendering organisations	6,163
	Total number of winning/award companies	19,100
Dataset's variables	Mean received offers by tender	5.02
	Input variables of tender's notice: Date, Name_Organisation, Postalzone, ISO_country_code, Main_activity, Type_code, CPV, CPV_Aggregated, Tender_Price, and Procedure_code	10 input variables
Prices (without taxes)	Output variables of tender's resolution: Award_Price	1 output variable
	Median tender price	€425,000.00
Number of tenders by CPV	Median award price	€394,951.26
	Tenders with CPV = 33: medical equipments, pharmaceuticals, and personal care products	10,927 (26.29%)
	Tenders with CPV = 15: food, beverages, tobacco, and related products	4,363 (10.50%)
	Tenders with CPV = 45: construction work	4,053 (9.75%)
	Tenders with CPV = 71: architectural, construction, engineering, and inspection services	1,973 (4.75%)
	Tenders with CPV = 34: transport equipment and auxiliary products to transportation	1,893 (4.56%)
Number of tenders by type code	Tenders with Type_code = 1: goods/supplies	24,593 (59.18%)
	Tenders with Type_code = 2: services	12,849 (30.92%)
	Tenders with Type_code = 3: works	4,114 (9.90%)

TABLE 8: European countries' dataset: errors between award price vs. tender price and award price vs forecast price and their differences.

Error	Award price vs. tender price	Award price vs. forecast price	Difference
Median absolute error (MdAE)	€4,514.50	€20,982.94	+€16,468.44
Median absolute percentage error (MdAPE)	4.17%	6.48%	+2.31%
Mean absolute percentage error (MAPE)	27.49%	23.57%	-3.92%
Normalised root mean square error (NRMSE)	99,018.04	2,816,245.06	+2,717,227.02
Coefficient of determination (R^2)	0.9680	0.7303	-0.2377

output parameter) ordered from highest to lowest are Tender_Price (0.870%), Received_Offers (0.035%), Duration (0.017%), Date (0.013%), Name_Organisation (0.012%), CIF_Winner (0.010%), CPV (0.009%), Postalzone (0.007%), Subtype_code (0.006%), CPV_Aggregated (0.005%), Winner_Province (0.004%), Type_code (0.004%), Procedure_code (0.003%), Postalzone_Municipality (0.002%), Postalzone_Province (0.001%), Postalzone_CCAA (0.001%), and Urgency_code (0.0001%). It is clear that the 3 output variables are important in the previous ranking.

4.2. Empirical Results and Analysis. RF has been trained with 80% of tenders (46,670). The remaining 20% (11,667) have been used as the test group. Figure 9 shows the scatter plot between forecast price and award price for the test group. As has already been mentioned, if the estimator were perfect, all points would have to be on the line at 45 degrees.

The prediction's errors are presented in Table 6. Furthermore, in the third column, it is compared with the error made by Tender_Price (see Table 4) to check if the proposed estimator is better or worse. It makes no sense to compare the absolute errors because the sizes of the datasets are different. It is best to compare the percentage errors, such as MdAPE and MAPE; they are significantly lower, MdAPE—2.58% and MAPE—11.19%.

Figure 10 shows the boxplot of APE (grey colour) between award price and forecast price grouped by CPV. It is also plotted the APE reference (blue colour) which has been presented previously in Figure 5. It is clearly visible how the APE of the estimator has boxplots with a smaller interquartile range (IQR). In general, MdAPE and MAPE are lower than the APE reference. In conclusion, the proposed estimator reduces significantly the error with respect to tender price (analysed in Section 3.3).

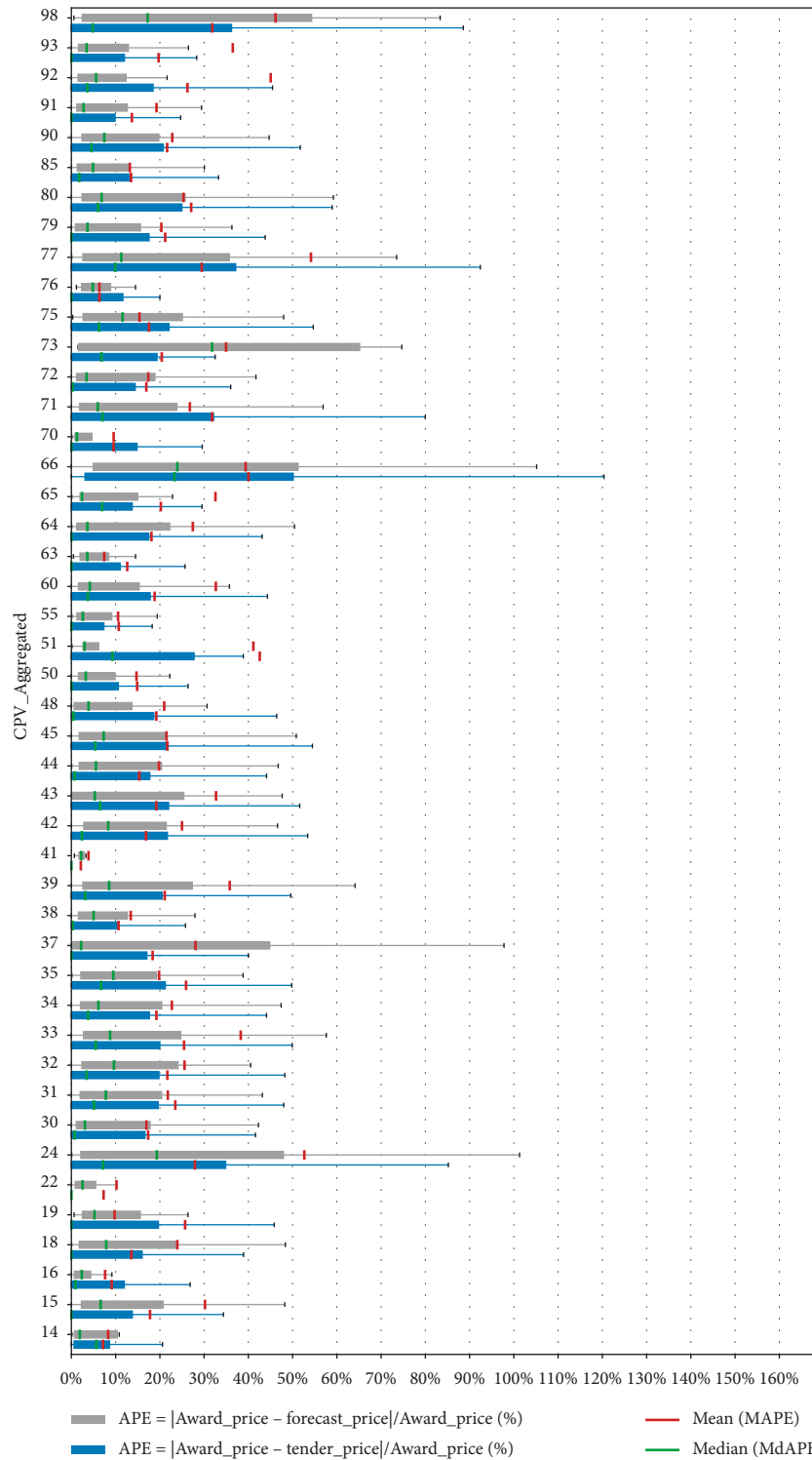


FIGURE 11: European countries’ boxplot: absolute percentage error (APE) between award price and tender price (blue colour) and award price and forecast price (grey colour), grouped by CPV.

The variable importances (RF output parameter) ordered from highest to lowest are Tender_Price (88.34%), Date (1.84%), Duration (1.83%), Name_Organisation (1.56%), Subtype_code (1.52%), CPV (1.10%), Postalzone (1.09%), Type_code (0.97%), Procedure_code (0.66%), CPV_Aggregated (0.49%), Postalzone_Municipality (0.24%),

Postalzone_Province (0.18%), Postalzone_CCAA (0.17%), and Urgency_code (0.03%).

4.3. Empirical Results and Analysis for Other Countries. In this section, a study is made with tenders from other countries, similar to the previous one for Spanish tenders. The

purpose is to evaluate the award price estimator with a different dataset using the same machine learning technique (random forest). The countries selected are from the European Union because they have almost the same characteristics associated with public procurement announcements: legislation, tender's regulation, public administrations, purchase procedures, etc. The raw data have been downloaded from the European Open Data Portal [21], in particular the tenders' database of the year 2017 (link in the Data Availability section). However, the quality of the data is not good: fields without data, errors in tender and award prices, the winning company does not have its tax identification number, tender and award prices have the same value, etc. It is an official dataset provided by the European Union, but it does not have as good a quality as the Spanish dataset. In the beginning, there were 706,104 tenders. After data preprocessing, there were only 41,556 tenders.

Table 7 shows the quantitative description of the dataset for the following 8 European countries: France, Germany, Italy, Croatia, Slovenia, Bulgaria, Hungary, and Latvia. They have been selected because they have the highest number of tenders after data preprocessing.

This dataset has been trained with 80% of the tenders (33,244). The remaining 20% (8,312) have been used as the test group. The random forest process is analogous to the Spanish one. The 10 input variables used in RF are Date, Name_Organisation, Postalzone, ISO_country_code, Main_activity, Type_code, CPV, CPV_Aggregated, Tender_Price, and Procedure_code. The variable to perform the regression is Award_Price, and the output generated by RF (prediction) will be called Forecast_Price.

The errors MdAE, MdAPE, MAPE, and NRMSE and R^2 are shown in Table 8. The second column shows award price vs. tender price (the reference), and the third column shows award price vs. forecast price (the estimator created with RF). MdAPE between award price and tender price is very low (4.17%) if it is compared to the Spanish MdAPE (11.84%, see Table 4). This means that award price is very close to tender price or, in other words, a lot of tenders have the same price for both and, consequently, without error. MAPE is also lower (27.49%) than the Spanish MAPE (39.79%). The estimator is better in MAPE (-3.92%) but it is worse in MdAPE (+2.31%) (see fourth column in Table 8).

Figure 11 shows the boxplot of APE (grey colour) between award price and forecast price grouped by CPV. The APE reference (blue colour) between award price and tender price is also plotted. It is not clearly visible how the APE of the estimator has boxplots with a smaller interquartile range (IQR). In general, MdAPE and MAPE are similar to the APE reference.

In conclusion, the estimator created for this dataset has similar error metrics with respect to tender price. Why do a lot of tender notices in the European countries have the same value of tender price and award price? Why not in the Spanish case? This could be due to the bad quality of the European dataset (tender's notices with mistakes) or, a less likely hypothesis, the fact that the Spanish public procurement agencies fail to estimate the tender price and the

European agencies never fail in anything. The proposed method can be useful and generalisable to other countries with a large dataset without mistakes.

5. Conclusions and Future Research

The European and Spanish public procurement legislation has been presented. A dataset of 58,337 Spanish public tenders from 2012 to 2018 has been analysed. The relations between the main fields of the public procurement notices have been studied mathematically. Error metrics between the tender price and the award price have been calculated (MdAPE = 11.84% and MAPE = 39.79%). An award price estimator, which reduces the previous errors (MdAPE = 9.26% and MAPE = 28.60%), has been proposed by using a machine learning algorithm (random forest). The estimator has 14 fields as input variables, of which the most important are the tender price, date, duration, public procurement agency name, subtype code, CPV classification, and postal zone.

A good award price estimator would be useful for companies and public procurement agencies. It would be useful for companies because it can be a key tool in their project management decision making: it would reduce the economic risks, thus winning tenders more easily. For public procurement agencies, it would be useful because, for example, in the Spanish dataset, the tender price could have been reduced by 2.24% (MdAPE reduction), equivalent to approximately 811 million euros. This is a significant error reduction that, consequently, would improve the accuracy of the budget for public procurement.

An analogous analysis has been made with 8 European countries (France, Germany, Italy, Croatia, Slovenia, Bulgaria, Hungary, and Latvia) to generalise the award price estimator to other real situations and check the results. The dataset used has 41,556 tenders, but the quality of the data is worse than the Spanish dataset. The new award price estimator obtains predictions with error metrics that are similar to those between the tender price and award price. The estimator is better in MAPE (-3.92%) but it is worse in MdAPE (+2.31%).

An accurate estimate is impossible to achieve because the market is theoretically open and free and, therefore, unpredictable. Furthermore, the award price is not always the final price paid by the public procurement agency because the contract may be modified during its execution. However, this article illustrates how a machine learning algorithm can be useful. Particularly, random forest predicts the award prices with less uncertainty, adapting to the real market. This market reality is gathered implicitly through the public procurement notices. Therefore, this estimator is interesting for the public procurement agencies and for the companies because their risk is reduced.

Thanks to the open data sources of public procurement, it is possible to avoid depending on government statistics offices such as the Spanish (INE [47]) or the European (Eurostat [23]). Therefore, there is independence, and there are resources to perform low-level analysis or cross data with

other databases or external services to extract more valuable information.

This article opens the doors to future research related to the analysis of massive data on public procurement, in particular:

- (i) It achieves a more accurate estimator by incorporating business data of the winning bidder: location, core business, annual turnover, number of employees, financial situation, etc. With the new data, the estimator has more input variables that could be relevant to predicting the award price.
- (ii) It compares other machine learning algorithms to estimate award prices, number of received offers, and other interesting fields.
- (iii) It performs data business analysis such as companies with a higher success rate in public procurement or the characterisation of the winning bidder: type of company, size, national origin or foreign, etc.

Data Availability

The processed data used to support the findings of this study are available from the corresponding author upon request. The raw data from Spain are available at the Ministry of Finance, Spain. Open data of Spanish tenders are hosted in http://www.hacienda.gob.es/es-ES/GobiernoAbierto/Datos%20Abiertos/Paginas/licitaciones_plataforma_contratacion.aspx. The raw data from other countries are available in the European Union Open Data Portal (Publications Office of the European Union) hosted in <https://data.europa.eu/euodp/en/data/dataset/ted-csv>.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the Plan of Science, Technology and Innovation of the Principality of Asturias (Ref: FC-GRUPIN-IDI/2018/000225).

References

- [1] European Commission, "European semester thematic factsheet public procurement," 2017, https://ec.europa.eu/info/sites/info/files/file_import/european-semester_thematic-factsheet_public-procurement_en_0.pdf.
- [2] The National Securities Market Commission (CNMV) from Spain, "Radiography of public procurement procedures in Spain," 2019, https://www.cnmv.es/sites/default/files/2314114_5.pdf.
- [3] N. Obwegeser and S. D. Müller, "Innovation and public procurement: terminology, concepts, and applications," *Technovation*, vol. 74-75, pp. 1-17, 2018.
- [4] J. Attard, F. Orlandi, S. Scerri, and S. Auer, "A systematic review of open government data initiatives," *Government Information Quarterly*, vol. 32, no. 4, pp. 399-418, 2015.
- [5] H. R. Varian, "Big data: new tricks for econometrics," *Journal of Economic Perspectives*, vol. 28, no. 2, pp. 3-28, 2014.
- [6] S. Mullainathan and J. Spiess, "Machine learning: an applied econometric approach," *Journal of Economic Perspectives*, vol. 31, no. 2, pp. 87-106, 2017.
- [7] J. M. Alvarez-Rodríguez, J. E. Labra-Gayo, and P. O. De Pablos, "New trends on e-procurement applying semantic technologies: current status and future challenges," *Computers in Industry*, vol. 65, no. 5, pp. 800-820, 2014.
- [8] M. Nečaský, J. Klímeck, J. Mynarz, T. Knap, V. Svátek, and J. Stárka, "Linked data support for filing public contracts," *Computers in Industry*, vol. 65, no. 5, pp. 862-877, 2014.
- [9] J. D. Twizeyimana and A. Andersson, "The public value of e-government—a literature review," *Government Information Quarterly*, vol. 36, no. 2, pp. 167-178, 2019.
- [10] M. A. Bergman and S. Lundberg, "Tender evaluation and supplier selection methods in public procurement," *Journal of Purchasing and Supply Management*, vol. 19, no. 2, pp. 73-83, 2013.
- [11] T. D. Fry, R. A. Leitch, P. R. Philipoom, and Y. Tian, "Empirical analysis of cost estimation accuracy in procurement auctions," *International Journal of Business and Management*, vol. 11, no. 3, p. 1, 2016.
- [12] H. Jung, G. Kosmopoulou, C. Lamarche, and R. Sicotte, "Strategic bidding and contract renegotiation," *International Economic Review*, vol. 60, no. 2, pp. 801-820, 2019.
- [13] K. Bloomfield, T. Williams, C. Bovis, and Y. Merali, "Systemic risk in major public contracts," *International Journal of Forecasting*, vol. 35, no. 2, pp. 667-676, 2019.
- [14] P. Ballesteros-Pérez, M. Skitmore, E. Pellicer, and J. H. Gutiérrez-Bahamondes, "Improving the estimation of probability of bidder participation in procurement auctions," *International Journal of Project Management*, vol. 34, no. 2, pp. 158-172, 2016.
- [15] T. Hanák and P. Muchová, "Impact of competition on prices in public sector procurement," *Procedia Computer Science*, vol. 64, pp. 729-735, 2015.
- [16] V. Títl and B. Geys, "Political donations and the allocation of public procurement contracts," *European Economic Review*, vol. 111, pp. 443-458, 2019.
- [17] S. Tadelis, "Public procurement design: lessons from the private sector," *International Journal of Industrial Organization*, vol. 30, no. 3, pp. 297-302, 2012.
- [18] T. Hanák and C. Serrat, "Analysis of construction auctions data in Slovak public procurement," *Advances in Civil Engineering*, vol. 2018, Article ID 9036340, 13 pages, 2018.
- [19] J.-M. Kim and H. Jung, "Predicting bid prices by using machine learning methods," *Applied Economics*, vol. 51, no. 19, pp. 2011-2018, 2019.
- [20] Publications Office of the European Union, *The Official Journal of the European Union*, Publications Office of the European Union, Brussels, Belgium, 2019, <https://eur-lex.europa.eu/oj/direct-access.html>.
- [21] Publications Office of the European Union, *European Union Open Data Portal*, Publications Office of the European Union, Brussels, Belgium, 2019, <http://data.europa.eu/euodp>.
- [22] Tenders Electronic Daily (TED), "Online version of the supplement to the official journal of the EU," 2019, <https://ted.europa.eu>.
- [23] European Commission, Eurostat, <https://ec.europa.eu/eurostat>.
- [24] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5-32, 2001.
- [25] A. Verikas, A. Gelzinis, and M. Bacauskiene, "Mining data with random forests: a survey and results of new tests," *Pattern Recognition*, vol. 44, no. 2, pp. 330-349, 2011.

- [26] M. Fernández-Delgado, M. S. Sirsat, E. Cernadas, S. Alawadi, S. Barro, and M. Febrero-Bande, "An extensive experimental survey of regression methods," *Neural Networks*, vol. 111, pp. 11–34, 2019.
- [27] M. R. Segal, *Machine Learning Benchmarks and Random Forest Regression*, UCSF: Center for Bioinformatics and Molecular Biostatistics, San Francisco, CA, USA, 2004, <https://escholarship.org/uc/item/35x3v9t4>.
- [28] G. Biau and E. Scornet, "A random forest guided tour," *Test*, vol. 25, no. 2, pp. 197–227, 2016.
- [29] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, Berlin, Germany, 2nd edition, 2008.
- [30] M. B. Kursa and W. R. Rudnicki, "Feature selection with the Boruta package," *Journal Of Statistical Software*, vol. 36, no. 11, pp. 1–13, 2010.
- [31] H. Deng and G. Runger, "Gene selection with guided regularized random forest," *Pattern Recognition*, vol. 46, no. 12, pp. 3483–3489, 2013.
- [32] T. Hothorn, K. Hornik, and A. Zeileis, "Unbiased recursive partitioning: a conditional inference framework," *Journal of Computational and Graphical Statistics*, vol. 15, no. 3, pp. 651–674, 2006.
- [33] N. Meinshausen, "Quantile regression forests," *Journal of Machine Learning Research*, vol. 7, pp. 983–999, 2006.
- [34] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine Learning*, vol. 63, no. 1, pp. 3–42, 2006.
- [35] Y. Cheng, X. Chen, X. Ding, and L. Zeng, "Optimizing location of car-sharing stations based on potential travel demand and present operation characteristics: the case of Chengdu," *Journal of Advanced Transportation*, vol. 2019, Article ID 7546303, 13 pages, 2019.
- [36] Q. Shang, D. Tan, S. Gao, and L. Feng, "A hybrid method for traffic incident duration prediction using BOA-optimized random forest combined with neighborhood components analysis," *Journal of Advanced Transportation*, vol. 2019, Article ID 4202735, 11 pages, 2019.
- [37] J. Xing and G. Zheng, "Stress field gradient analysis technique using lower-order C^0 elements," *Mathematical Problems in Engineering*, vol. 2015, Article ID 457046, 12 pages, 2015.
- [38] Z. Sun, H. Sun, and J. Zhang, "Multistep wind speed and wind power prediction based on a predictive deep belief network and an optimized random forest," *Mathematical Problems in Engineering*, vol. 2018, no. 4, Article ID 6231745, 15 pages, 2018.
- [39] Z. Liao, Y. Ju, and Q. Zou, "Prediction of G protein-coupled receptors with SVM-prot features and random forest," *Scientifica*, vol. 2016, Article ID 8309253, 10 pages, 2016.
- [40] L. Dong, X. Li, and G. Xie, "Nonlinear methodologies for identifying seismic event and nuclear explosion using random forest, support vector machine, and naive bayes classification," *Abstract and Applied Analysis*, vol. 2014, Article ID 459137, 8 pages, 2014.
- [41] T. Xiang, T. Li, M. Ye, and Z. Liu, "Random forest with adaptive local template for pedestrian detection," *Mathematical Problems in Engineering*, vol. 2015, Article ID 767423, 11 pages, 2015.
- [42] H. R. Zhang, F. Min, and X. He, "Aggregated recommendation through random forests," *The Scientific World Journal*, vol. 2014, Article ID 649596, 11 pages, 2014.
- [43] J. Ruysinck, J. van der Herten, R. Houthoof et al., "Random survival forests for predicting the bed occupancy in the intensive care unit," *Computational and Mathematical Methods in Medicine*, vol. 2016, Article ID 7087053, 7 pages, 2016.
- [44] A. A. Akinyelu and A. O. Adewumi, "Classification of phishing email using random forest machine learning technique," *Journal of Applied Mathematics*, vol. 2014, Article ID 425731, 6 pages, 2014.
- [45] E. Min, J. Long, Q. Liu, J. Cui, and W. Chen, "Tr-ids: anomaly-based intrusion detection through text-convolutional neural network and random forest," *Security and Communication Networks*, vol. 2018, Article ID 4943509, 9 pages, 2018.
- [46] X. Gao, J. Wen, and C. Zhang, "An improved random forest algorithm for predicting employee turnover," *Mathematical Problems in Engineering*, vol. 2019, Article ID 4140707, 12 pages, 2019.
- [47] National Statistics Institute (INE), Spain, <https://www.ine.es>.

Research Article

Chinese Currency Exchange Rates Forecasting with EMD-Based Neural Network

Jying-Nan Wang ^{1,2}, Jiangze Du ^{3,4}, Chonghui Jiang ^{3,4} and Kin-Keung Lai ⁵

¹College of International Finance and Trade, Zhejiang Yuexiu University of Foreign Languages, Shaoxing, Zhejiang, China

²Research Institute for Modern Economics and Management, Zhejiang Yuexiu University of Foreign Languages, Shaoxing, Zhejiang, China

³School of Finance, Jiangxi University of Finance and Economics, Nanchang, China

⁴Research Centre of Financial Management and Risk Prevention, Jiangxi University of Finance and Economics, Nanchang, China

⁵Department of Industrial and Manufacturing Systems Engineering, The University of Hong Kong, Pok Fu Lam, Hong Kong

Correspondence should be addressed to Jiangze Du; jiangze.du@hotmail.com

Received 27 June 2019; Revised 1 September 2019; Accepted 19 September 2019; Published 30 October 2019

Guest Editor: Benjamin M. Tabak

Copyright © 2019 Jying-Nan Wang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The Chinese currency, RMB, is developing as an international currency. Therefore, the effective strategy for trading RMB exchange rates would be attractive to international investors and policymakers. In this paper, we have constructed hybrid EMD-MLP models to forecast RMB exchange rates and developed a trading strategy based on these models. Empirical results show that the proposed hybrid EMD-MLP* model always performs best based on both NMSE and D_{stat} criteria when the forecasting period is greater than five days. Moreover, we compare the models' performance using different horizons and find that accuracy will increase with the growth of the forecasting horizons; however, the NMSE will become larger. Lastly, we adopt the best performing model to develop trading strategies with longer forecasting horizons when considering the number of profitable trading activities. If we consider a 0.3% transaction cost, the developed strategy will bring an annual return exceeding 10%, as well as enough trading opportunities.

1. Introduction

As the Chinese government continues the process of RMB internationalization, RMB currency trading becomes increasingly important in personal investment, corporate financial decision-making, governments' economic policies, and international trade and commerce. The RMB has therefore attracted an increasing attention from policymakers, investment institutions, and entrepreneurs worldwide. One reflection of this is the 2015 finding of the Society for Worldwide Interbank Financial Telecommunication (SWIFT) that the RMB has overtaken the Japanese yen to become the fourth ranking world payment currency. The first milestone in the process of RMB internationalization is the establishment of an RMB offshore center in Hong Kong in 2004. In December 2008, the

Chinese Premier announced the pilot program for cross border trade settlement in RMB, making the offshore RMB officially deliverable in Hong Kong. From July 19, 2010, the offshore market for RMB officially commenced. Recent literature [1–4] has found that RMB movements have an impact on regional currencies.

Due to the importance of the RMB, the purpose of this study is to create a reliable RMB forecasting model and provide a possible application for designing trading strategies. Two kinds of current rates for exchanging the US dollar (USD) into Chinese RMB will be considered. If the RMB is traded onshore (in mainland China), it is referred to as CNY, whereas if traded offshore (mainly in Hong Kong), it is designated as CNH. Since the onshore and offshore markets might respond differently to changes depending on financial markets, the CNY and CNH are not always at the

same price level. Craig et al. [5] and Funke et al. [6] study CNY-CNH pricing differentials in detail.

It is well documented that the exchange rate series is considered as nonlinear and nonstationary time series and interactively influenced by many factors, which makes the accurate forecasting of exchange rate rather challenging. During past decades, traditional econometric and statistical techniques, including autoregressive integrated moving average (ARIMA), cointegration analysis, vector autoregression (VAR), and error correction (ECM) models, have been widely used in foreign exchange rates forecasting. However, in the real-world financial markets, exchange rate series are nonlinear and rarely form purely linear combinations [7–10]. Thus, the above traditional models always provide unreliable forecasting if one continue applying these traditional econometric and statistical models. The main reason of this deficiency is that traditional econometric and statistical models are constructed based on linear assumptions, which will be unable to capture the nonlinear features hidden in the exchange rate series.

Considering the limitation of traditional econometric and statistical models, many nonlinear artificial intelligence models (AI), such as artificial neural networks (ANN) [11, 12], feedforward neural networks (FNN) [13, 14], support vector regression (SVR) [15–17], and genetic programming (GP) [18, 19], have been applied to investigate the forecasting ability of financial time series. Yu et al. [20] provide a complete review of foreign exchange rate forecasting with ANN and also introduce SVR and GP. However, many AI-based models also have their own shortcomings. For instances, ANN usually suffers from overfitting and local minima, while other models, including GP and SVR, are sensitive to parameter estimation. Recently, due to the complex characteristics of exchange rate series, several studies show that a single model is unable to capture all the features and make accurate forecasts [21].

To overcome above shortcomings, many researchers start to rely on the hybrid model to forecast exchange rate series accurately. The composite forecasting method allows us to obtain an approach to the dynamics underlying the data by combining the predictions obtained from different individual techniques. Álvarez-Díaz and Álvarez [22] attempt to exploit the nonlinear structure by constructing the genetic programming and neural networks composite method and apply this composite method to forecast Yen/US\$ and Pound Sterling/US\$ exchange rates. Other scholars also try to improve the forecasting accuracy of exchange rate series by applying hybrid forecasting models. For example, Zhao and Yang [23] and Wong et al. [24] use fuzzy clustering and ANN to forecast financial time series. Yu et al. [25] propose an online big data-driven oil consumption forecasting model based on Google trend by combining the relationship investigation and prediction improvement. Yu et al. [14] design a novel ensemble forecasting approach for complex time series by coupling sparse representation (SR) and feedforward neural network (FNN), i.e., the SR-based FNN approach.

The empirical mode decomposition (EMD) technique is first proposed by Huang et al. [26]. From the theoretical

views, EMD is suitable for time series data in terms of decomposing the original data into components, which could break the forecasting task down into simpler forecasting subtasks. These decompositions consist of a finite and often small number of intrinsic mode functions (IMFs) and one residual. From the perspective of financial practice, various economic activities may cause a different impact periodicity on the exchange rate; for example, the company's financial report is released quarterly, but government economic indicators are usually published once a year. EMD is potentially helpful in decomposing these effects and leads to improve forecasting works. There are several studies employing the EMD technique in hybrid models. For instance, Yu et al. [27] forecast crude oil prices with an EMD-based neural network model; Chen et al. [28] also combine EMD and the ANN approach to forecast tourism demand, and Lin et al. [29] propose a hybrid model using EMD and SVR for foreign exchange rate forecasting. In this empirical experiment, the original CNY or CNH price series, with characteristics of nonlinearity and nonstationarity, are divided into several independent subseries by the EMD technique and then partial or all IMFs and one residual are used to forecast. Tang et al. [30] combine the ensemble empirical mode decomposition (EEMD) with random vector functional link (RVFL) network and find the proposed EEMD-based RVFL network performs significantly better in terms of prediction accuracy than not only single algorithms such as RVFL network, extreme learning machine (ELM), kernel ridge regression, random forest, backpropagation neural network, least square support vector regression, and autoregressive integrated moving average, but also their respective EEMD-based ensemble variants. We summarize in Table 1 the main characteristics of these studies on individual and hybrid forecasting methods in the recent literature.

This study will focus on using ANN to forecast the RMB with different forecasting horizons of 1, 5, 10, 20, and 30 days. Nevertheless, Huang et al. [32] show that ANN performs better than the random walk while the forecasting horizon is less than five days, but for longer horizons, such as 10 and 30 days, the general performance of ANN is worse than the random walk model. In this study, a type of feedforward artificial neural network model called multi-layer perceptron (MLP) is adopted. Empirical studies consist not only of the pure MLP model but also a hybrid model with EMD to improve forecasting performance.

In this study, the hybrid forecasting model combining MLP and EMD is similar to the work of Yu et al. [27]. However, this study further considers the influence of IMFs that have different levels of frequency. In concrete terms, higher frequency IMFs could be regarded as noise components, when our forecasting horizon is longer. Based on this concept, two hybrid models are proposed and named EMD-MLP and EMD-MLP*. Comparing these to pure MLP, the former adjusts the original time series data by subtracting higher frequency IMFs and then uses MLP to compute the one-day ahead predictions. The EMD-MLP* model applies the “divide-and-conquer” principle to construct a novel forecasting methodology, in which partial or

TABLE 1: Literature on individual and hybrid forecasting models.

Study	Type	Algorithm	Data	Time range	Criteria
Aladag et al. [11]	Individual	ANN	Exchange rates (TL/EUR, LEU/EUR)	2005–2012	RMSE
Alvarez-Diaz and Alvarez [18]	Individual	GP	Exchange rates (various pairs)	1971–2000	R-square
Álvarez-Díaz and Álvarez [22]	Individual	GP, ANN	Exchange rates (GBP/USD, JPY/USD)	1973–2002	NMSE, SR
Álvarez Díaz [31]	Individual	GP	Exchange rates (GBP/USD, JPY/USD)	1973–2002	U-Theil value, SR
Huang et al. [32]	Individual	ANN	Exchange rates (USD/GBP, USD/JPY)	1997–2002	RMSE
Kajitani et al. [13]	Individual	ANN	Canadian lynx data	1821–1934	RMSE
Neely et al. [19]	Individual	GP	Exchange rates (various pairs)	1974–1995	Excess return
Nikolsko-Rzhevskyy and Prodan [33]	Individual	Markov switching	Exchange rates (various pairs)	1983–2008	MSE
Wong et al. [24]	Individual	Novel ANN	Simulated time series data	–	MAPE, NMSE
Yu et al. [16]	Individual	Least squares SVM	Crude oil	2008–2015	MAPE
Yu et al. [25]	Individual	Artificial intelligence	Global oil consumption	2004–2015	PCC, AUC, RMSE, MAPE
Zhang [10]	Individual	ANN	Simulated time series data	—	MSE, MAPE
Zhao and Yang [23]	Individual	CRPSO-based neuron model	Simulated time series data, electroencephalogram data	—	MSE
Cao [15]	Hybrid	SVM + SOM	Sunspot data, Santa Fe datasets A, C, and D, and the two building datasets	—	NMSE, RMSE, CV
Chen et al. [28]	Hybrid	EMD + ANN	Tourism demand	1971–2009	MAPE, RMSE, MAE
Khashei et al. [7]	Hybrid	ARIMA + ANNs + fuzzy	Exchange rates (USD/Iran rials), gold price	2005–2006	MAE, MSE
Lin et al. [29]	Hybrid	EMD + SVM	Exchange rates (various pairs)	2005–2009	MAPE, RMSE, MAE, DS, CP, CD
Tang et al. [30]	Hybrid	EEMD + RVFL	Crude oil	1986–2010	MAPE, RMSE, DS
Yu et al. [12]	Hybrid	ANN + WD	Patient visits	2010–2015	RMSE, MAPE
Yu et al. [27]	Hybrid	EMD + ANN	Crude oil	1986–2006	NMSE, DS
Yu et al. [17]	Hybrid	OPL + SVMQR	Ten publicly available datasets	—	Empirical risk, quantile property
Yu et al. [14]	Hybrid	Sparse representation + ANN	Crude oil	1986–2013	RMSE MAPE
Zhang [21]	Hybrid	ARIMA + ANN	Wolf's sunspot data, Canadian lynx data, and GBP/USD	—	MSE, MAE

ANN: artificial neural network; SVM: support vector machine; GP: genetic programming; CRPSO: cooperative random learning particle swarm optimization; RVFL: random vector functional link; OPL: orthogonal pinball loss; SVMQR: SVM quantile regression; WD: wavelet decomposition; SOM: self-organization feature map; RMSE: root mean squared error; NMSE: normalized mean squared error; MAE: mean absolute error; MAPE: mean absolute percentage error; CV: coefficient of variation; PCC: percentage correctly classified accuracy; AUC: area under the receiver operating curve; SR: success ratio; DS: directional symmetry; CP: correct uptrend; CD: correct downtrend.

all IMFs and one residual are used. According to the empirical evidence in this study, both types of hybrid models are superior to the pure MLP model, whether with 1-, 5-, 10-, 20-, or 30-day forecasting horizons. Moreover, the application of trading strategies is proposed. Although the transaction costs can make the profits practically disappear or become negative, as proved by Álvarez Díaz [31], this empirical analysis shows that, even considering a 0.3% transaction cost in each trade, the trading strategy based on EMD-MLP, on average, produces an annualized profit exceeding 10%.

In this paper, we will investigate RMB exchange rate forecasting and develop the relevant trading strategies based on the constructed models. The main contributions come

from three aspects. First, this study focuses on the RMB, including the onshore RMB exchange rate (CNY) and offshore RMB exchange rate (CNH). We will use daily data to forecast RMB exchange rates with different horizons based on three types of models, i.e., MLP, EMD-MLP, and EMD-MLP*. Second, we will consider not only the MLP model but also the hybrid EMD-MLP and EMD-MLP* models to improve forecasting performance. It should be noted that the EMD-MLP* is different from the methodology proposed by Yu et al. [27]. We regard some IMF components as noise factors and delete them to reduce the volatility of the RMB. Finally, we will choose the best forecasting models to construct the trading strategies by introducing different critical numbers and considering different transaction costs.

The remainder of the paper is organized in the following manner. Section 2 provides a brief description of ANN and EMD. The overall forecasting process and model notations of three kinds of forecasting models are also included in this part. In Section 3, two currency exchange rates of USD to RMB, CNY, and CNH, are used to test the effectiveness of the proposed methodology, and the selected models are applied in trading strategies. The conclusions drawn from this study are presented in Section 4.

2. Methodology

2.1. Artificial Neural Network (ANN). This study considers MLP based on an error backpropagation algorithm. Figure 1 shows a simple MLP structure with three input nodes ($X1$, $X2$, and $X3$). One hidden layer consists of four hidden nodes and one output node (Y). The nodes are organized in layers and are usually fully connected by weights, which indicate the effects of the corresponding nodes. In each node of the hidden and output layers, all data are firstly processed by the integration function (also called the summation function), which combines all incoming signals, and secondly processed by the activation function (also called the transfer function), which transforms the output of the node. In general, the amount of the hidden layer is less than 3, due to the converging restriction.

Particularly, the MLP model is fitted by the training data, and then the testing data and an out-of-sample dataset are used to verify its forecasting performance. Through supervised learning algorithms, the parameters (weights and node intercepts) are adjusted iteratively by a process of minimizing the forecasting error function. Formally, an MLP with an input layer with n nodes, one hidden layer consisting of J hidden nodes, and an output layer with one output node calculates the following function:

$$\begin{aligned} o(x) &= f\left(w_0 + \sum_{j=1}^J w_j \cdot f\left(w_{0j} + \sum_{i=1}^n w_{ij}x_i\right)\right) \\ &= f\left(w_0 + \sum_{j=1}^J w_j \cdot f(w_{0j} + \mathbf{w}_j^T \mathbf{x})\right), \end{aligned} \quad (1)$$

where $\mathbf{x} = (x_1, \dots, x_n)$ is the vector of all input variables, w_0 is the intercept of the output node, w_{0j} is the intercept of the j th hidden node, w_j denotes the weight corresponding to the node starting at the j th hidden node to the output node, and w_{ij} denotes the weight corresponding to the node starting at i th input node to the j th hidden node. Therefore, all hidden and output nodes calculate the function $f(g(\mathbf{z}))$, where $g(\cdot)$ denotes the integration function, which is defined as $g(\mathbf{z}) = w_0 + \mathbf{w}^T \mathbf{z}$, and $f(\cdot)$ denotes the activation function, which is often a bounded nondecreasing nonlinear and differentiable function. In this study, the logistic function ($f(u) = 1/(1 + e^{-u})$) is used as the activation function.

Given inputs \mathbf{x} and the current weights, which are initialized with random values from a standard normal distribution, the MLP produces an output $o(\mathbf{x})$. Then, an error function is defined. This study selects the mean square error as follows:

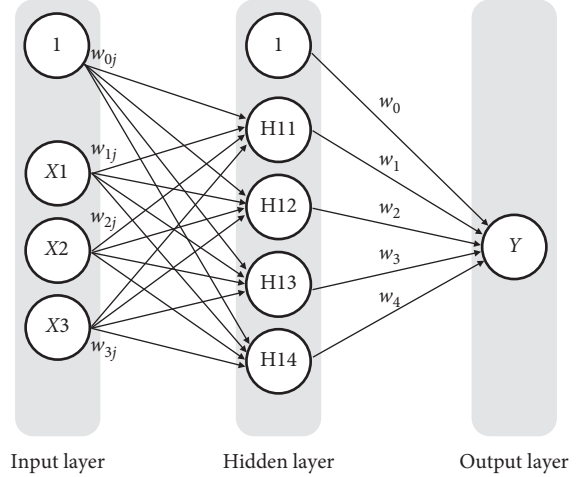


FIGURE 1: An artificial neural network structure with three input neurons ($X1$, $X2$, and $X3$), one hidden layer consisting of four hidden neurons ($H11$, $H12$, $H13$, and $H14$), and one output neuron (Y).

$$E = \frac{1}{N} \sum_{h=1}^N (o_h(\mathbf{x}) - y_h)^2, \quad (2)$$

where N is the number of data samples and y_h is the observed output. During the iterative training process, the above steps are repeated to adapt all weights until a pre-specified criterion is fulfilled. In order to find a local minimum of the error function, the resilient backpropagation algorithm modifies the weights in the opposite direction of partial derivatives. According to Riedmiller and Braun [34], the weights are adjusted by the following rule:

$$w_k^{(t+1)} = w_k^t - \eta_k^{(t)} \cdot \text{sign}\left(\frac{\partial E^{(t)}}{\partial w_k^{(t)}}\right), \quad (3)$$

where t and k index the iteration steps and the weights, respectively. In order to speed up convergence, the learning rate $\eta_k^{(t)}$ increases if the corresponding partial derivative keeps the same sign; otherwise, it will decrease.

2.2. Empirical Mode Decomposition (EMD). In this study, we further apply EMD to decompose the time series into several IMFs and remaining residues. These IMFs usually satisfy two conditions: the first is that the number of extrema and zero crossings must be equal or different by not more than one. Second, the mean value of the envelopes, which include both local maxima and minima, must be zero at all points.

Given the time series data $x(t)$, $t = 1, 2, \dots, T$, Huang et al. [26] propose a sifting process to decompose $x(t)$. The first step is to identify all local maxima and local minima of $x(t)$. Then, the upper and lower envelopes are defined by connecting all local extrema by a spline line. Next, for all points at the envelope, the mean value $m_1^1(t)$ from upper and lower envelopes is calculated. Then, it follows computation of the first IMF of $x(t)$:

$$h_1^1(t) = x(t) - m_1^1(t). \quad (4)$$

If $h_1^1(t)$ does not meet the above two conditions, this study then takes $h_1^1(t)$ as a new data series and repeats procedure (4). Thus, we calculate

$$h_2^1(t) = h_1^1(t) - m_2^1(t). \quad (5)$$

In this calculation, $m_2^1(t)$ is the mean value of the upper and lower envelopes of $h_1^1(t)$.

Repeating the same procedure until meeting both conditions, we get the first IMF component of $x(t)$, $c_1(t)$, that is

$$c_1(t) = h_q^1(t). \quad (6)$$

The stopping rule indicates that absolute values of the envelope mean must be less than the user-specified tolerance level. In this study, the tolerance level is denoted as the standard deviation of $x(t)$ times 0.01. Other interesting stopping rules can be found in the works of Huang et al. [26] and Huang and Wu [35].

After extracting the component $c_1(t)$ from $x(t)$, we denote another series $r_1(t) = x(t) - c_1(t)$, which contains all information except $c_1(t)$. Huang et al. [36] suggest a sifting stop criterion that is $r_n(t)$ becomes a monotonic function or cannot extract more IMFs. Finally, the time series $x(t)$ can be expressed as

$$x(t) = \sum_{i=1}^n c_i(t) + r_n(t), \quad (7)$$

where n is the number of IMFs and $r_n(t)$ is the final residue, which represents the central tendency of the data series $x(t)$. These IMFs are nearly orthogonal to each other and all have means of nearly zero. According to the above properties, it is possible to forecast all decompositions and summarize these estimations to predict $x(t)$.

2.3. Overall Forecasting Process and Model Notations. Considering the time series $p_t, t = 1, 2, \dots, T$, we would like to predict l -day ahead, which is denoted as \hat{p}_{t+l} . In this study, the input variables (past observations) include $p_{t-59}, p_{t-19}, p_{t-9}$, and p_{t-4} to p_t , which represent the past price levels of 60 days, 20 days, 10 days, and the last 5 days, respectively. The output is the l -day ahead prediction. Formally, it could be shown as follows:

$$\hat{p}_{t+l} = \varphi(p_t, \dots, p_{t-4}, p_{t-9}, p_{t-19}, p_{t-59}, \mathbf{w}), \quad (8)$$

where $\varphi(\cdot)$ is a function determined by neural network training and \mathbf{w} is a weight vector of all parameters of MLP.

Three kinds of forecasting models are adopted in this study. The first is the pure MLP model with one and two hidden layers. The second is an EMD-MLP model, which uses EMD to subtract some volatile IMFs from the original data series and then uses the new data series to calculate the final prediction by MLP technology. Figure 2 indicates the procedure of the EMD(-1)-MLP model, which means the first IMF component is ignored. The third kind of model is

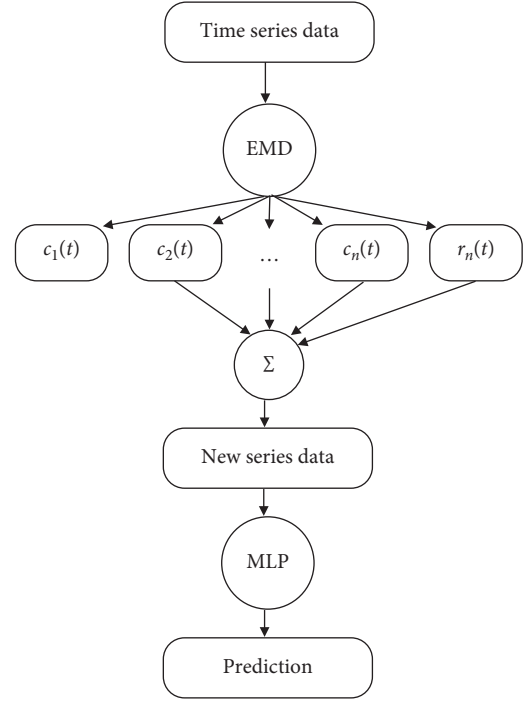


FIGURE 2: An example of the EMD(-1)-MLP model. We decompose a time series data by the EMD and generate n IMF components, $c_1(t), c_2(t), \dots, c_n(t)$, and one residue $r_n(t)$. We sum all decompositions except $c_1(t)$ to produce a new time series data. Then, the MLP is applied to compute the prediction.

named EMD-MLP*, which generally consists of the following four steps:

- (1) Decompose the original time series into IMF components and one residual component via EMD
- (2) Determine how many IMF components are used, which depends on the length of the forecasting period
- (3) For each chosen IMF and residual component, the MLP model is used as a forecasting tool to model these components and to make the corresponding prediction
- (4) Add all prediction results to one value, which can be seen as the final prediction result for the original time series.

As an example, Figure 3 represents the above procedure of the EMD(-1)-MLP* model. Time series data are decomposed via EMD and generates n IMF components, $c_1(t), c_2(t), \dots, c_n(t)$, and one residue $r_n(t)$. In addition to $c_1(t)$, we forecast each decomposition and then sum them as the prediction results for the original time series.

3. Empirical Experiments

3.1. Data. Two kinds of currency exchange rates of USD to RMB are considered in this study. If the RMB is traded onshore (in mainland China), it is referred to as CNY, and if traded offshore (mainly in Hong Kong), it is named CNH.

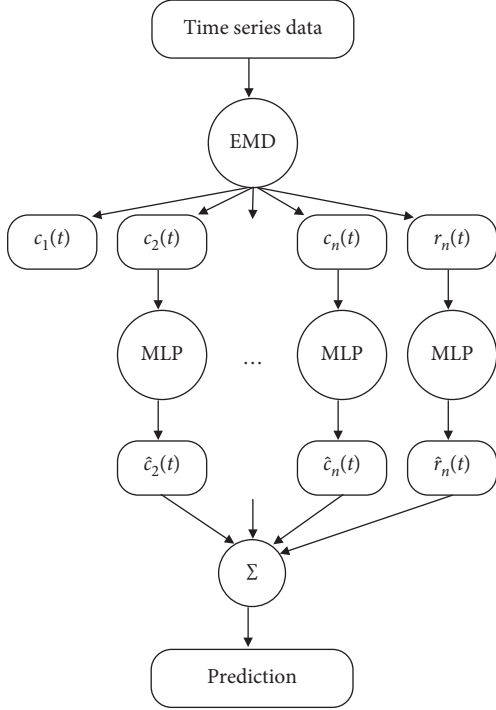


FIGURE 3: An example of the EMD(-1)-MLP* model. We decompose a time series data by the EMD and generate n IMF components, $c_1(t), c_2(t), \dots, c_n(t)$, and one residue $r_n(t)$. Besides $c_1(t)$, we forecast each decomposition and then sum them as the prediction.

Both time series data are downloaded from Bloomberg. For CNY, daily data from January 2, 2006, to December 21, 2015, with a total of 2584 observations are examined in this study. Since the CNH officially commenced on July 19, 2010, its sample period is from January 3, 2011, to December 21, 2015, with a total of 1304 observations. For training the neural network models, two-thirds of the observations are randomly assigned to the training dataset and the remainder is used as the testing dataset. In addition, it should be noted that the time series in price levels is used in the following analysis, rather than in return levels.

According to descriptions shown in Section 3.2, the EMD technique is used to decompose both CNY and CNH price series into several independent IMFs and one residual component. The decomposed results of CNY and CNH are represented in Figures 4 and 5. Comparing these two figures, the original data series and decompositions seem dissimilar, which is due to the different lengths of the sample periods. In fact, their IMFs have some similar characteristics. First, focusing on the sample period 2011–2015, their residual components have the same trend. Starting at about 6.5, they drop slightly to 6.2 and then rise slightly to 6.5. Second, the swing period of high-frequency IMFs is shorter and also the mean of absolute values of high-frequency IMFs is smaller when compared with low frequency. For example, in the series of CNY, the absolute values of c_1 , c_3 , and c_5 are 0.00298, 0.00596, and 0.02002, respectively. In our estimation model, it is not necessary to include all the high-frequency IMFs, as considering high-frequency IMFs may

reduce the forecasting accuracy when the forecasting period is long. Therefore, in the model EMD-MLP, the high-frequency IMF is removed to get the denoised time series. In the model EMD-MLP*, not only are all the decompositions chosen to forecast the exchange rate, but consideration is also given to using part of the decompositions to conduct forecasting, discarding the high-frequency IMF series.

3.2. Experimental Results. Two main criteria are considered in this empirical experiment: the normalized mean squared error (NMSE), and the directional statistic (D_{stat}), to evaluate the levels of prediction and directional forecasting, respectively. Typically, following [37], the NMSE is defined by

$$\text{NMSE} = \frac{1}{\sigma_{\Psi}^2} \frac{1}{N_t} \sum_{s \in \Psi} (\hat{p}_{s+l} - p_{s+l})^2, \quad (9)$$

where Ψ refers to the test dataset containing N_t observations, \hat{p}_{s+l} is the l -day ahead prediction, p_{s+l} is the actual value, and σ_{Ψ}^2 is the variance of p_{s+l} . Clearly, the NMSE is one of the most important criteria for measuring the validity of the forecasting model, but from a business perspective, improving the accuracy of directional predictions can support decision-making, so as to generate greater profits. Furthermore, the Diebold-Mariano (DM) test [38] is adopted to investigate whether adding EMD could improve the forecasting performance of the MLP model. Specifically, the null hypothesis is that the EMD-MLP (or EMD-MLP*) model has a lower forecast accuracy than the corresponding MLP model. For example, in the case of the EMD(-1)-MLP(5,3), the corresponding model infers the MLP(5,3) model.

Besides the NMSE evaluation, the directional statistic (D_{stat}) is applied to measure the ability to predict movement direction [27, 39], which can be expressed as

$$D_{\text{stat}} = \frac{1}{N_t} \sum_{s \in \Psi} a_s \times 100\%, \quad (10)$$

where $a_s = 1$ if $(\hat{p}_{s+l} - p_s)(p_{s+l} - p_s) \geq 0$ and $a_s = 0$ otherwise. Pesaran and Timmermann [40] provide the directional accuracy (DAC) test to examine predictive performance. In this case, the DAC test is used to determine whether D_{stat} is significantly larger than 0.5. We have also adopted the excess profitability test proposed by Anatolyev and Gerko [41] in every related empirical study. Since both tests produce the same outcome, we only report the results of DAC tests in the following empirical studies.

Table 2 compares the forecasting performance, in terms of the NMSE, for the MLP, EMD-MLP, and EMD-MLP* models. The NMSE is reported as the percentage for l -day ahead predictions where $l = 1, 5, 10, 20$, and 30. For each prediction period, the minimum NMSE is designated in bold font. Panel A of Table 2 displays the experimental results of the MLP models. It is clear that, with the increase in forecasting period, the NMSE becomes larger, suggesting reduced forecasting accuracy for a longer forecasting period. Also, the more complex MLP models with two hidden layers

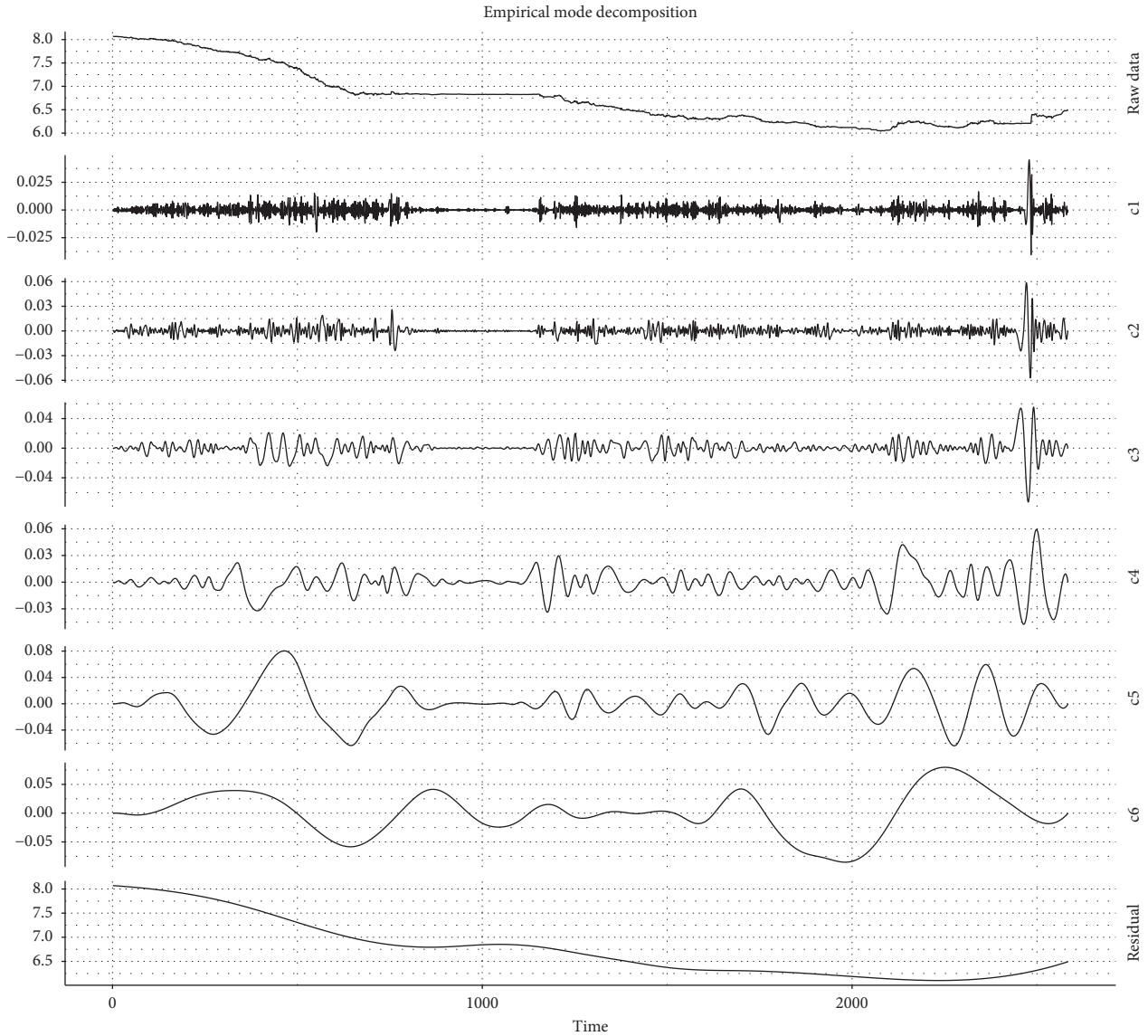


FIGURE 4: EMD decomposition of CNY from 2006 to 2015 is shown in this figure including the raw data series, 6 IMFs, and one residual.

do not offer better forecasting accuracy and are even worse for 1-day ahead prediction.

The empirical results of EMD-MLP are discussed in Panel B of Table 2. EMD(-1)-MLP treats the highest frequency IMF c_1 as noise and applies the MLP model after subtracting the c_1 series from the original time series. Empirical results show that the NMSE of 1-day ahead forecasting dropped dramatically; however, improvement of 5-day ahead forecasting with EMD(-1)-MLP models is limited. EMD(-2)-MLP models, which deduct the two highest frequency IMF series, perform better than EMD(-1)-MLP models in the 5-day ahead forecasting experiment. In the same way, the EMD(-3)-MLP models perform better when the forecasting period is longer than 10 days. The results of EMD-MLP* models are shown in Panel C of Table 2. Although the calculation process is more complicated, the forecasting performance of EMD-MLP* is always better than MLP or EMD-MLP models, according to the

criterion of NMSE. In addition, the random walk models with and without drifts (the random walk model without drifts predicts that all future values will equal the last observed value. Given a variable Y , the k -step-ahead forecast from period t of Y is $\hat{Y}_{t+k} = Y_t$. For the random walk model with drifts, the k -step-ahead forecast from period t of Y is $\hat{Y}_{t+k} = Y_t + k\hat{d}$, where \hat{d} is estimated by the average change of Y_t in the past 60 days) are used to examine the same data. The related results in Panel D of Table 2 indicate that, for the MLP-EMD and MLP-EMD* models, their performance is obviously better than the random walk models in most situations.

Table 3 is based on the forecasting performance of the above models with D_{stat} , which shows similar experimental results to the NMSE criterion. The table shows that if EMD(-2)-MLP(5,3)* is applied to forecasting the direction of the CNY series 30 days later, the hitting rate can be as high as 86.39%. In general, although the calculation of EMD-MLP*

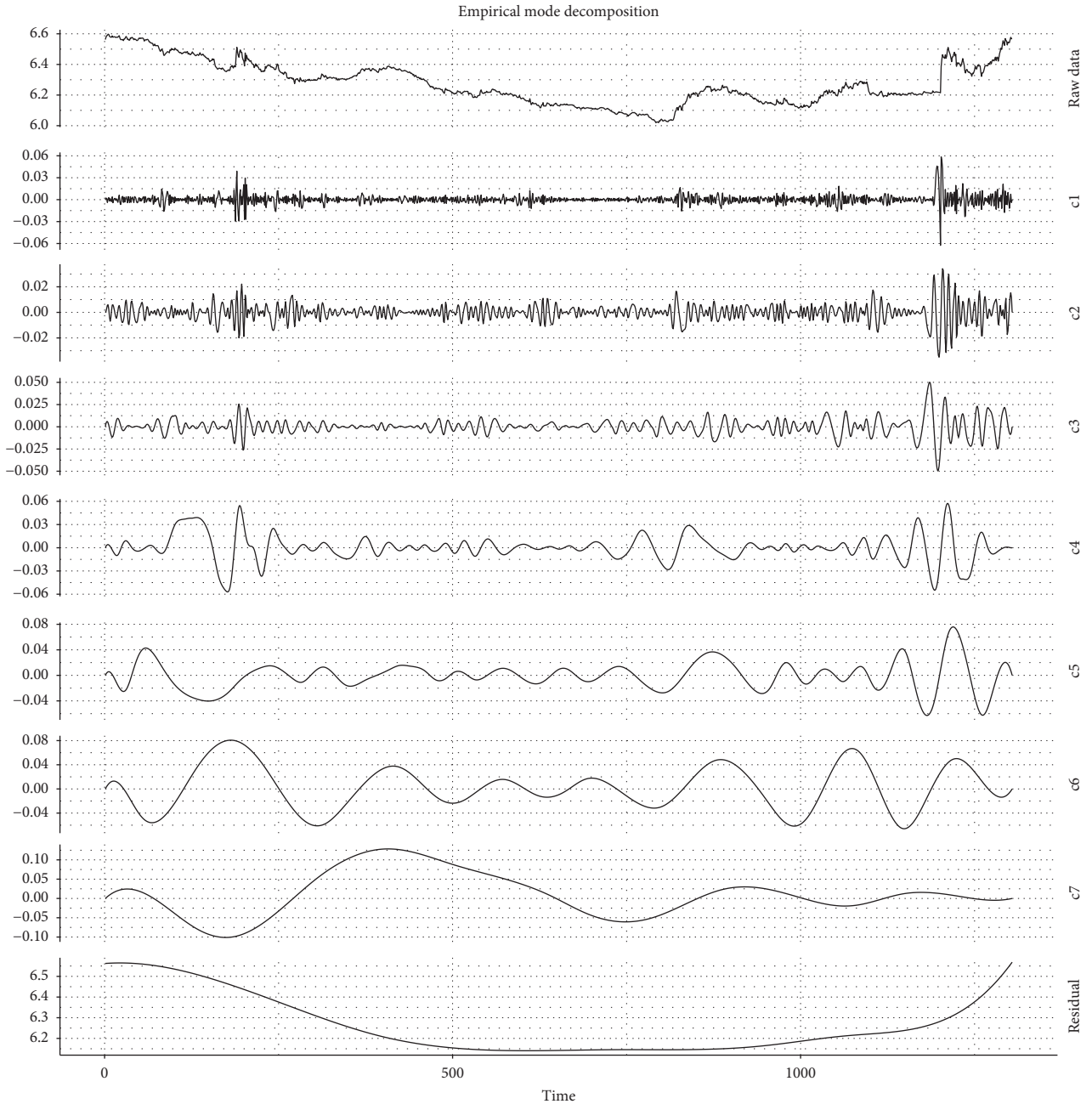


FIGURE 5: EMD decomposition of CNH from 2011 to 2015 is shown in this figure including the raw data series, 7 IMFs, and one residual.

models is more complicated, the forecasting ability appears to be the best of all the proposed models. The highest frequency IMF is therefore treated as noise and EMD(-1)-MLP* is used for the forecast. In this way, the performance of the selected model would be relatively better, based on both NMSE and D_{stat} criteria. Besides, it should be noted that performance of predicting over longer periods is worse than shorter ones intuitively. In terms of NMSE, the experimental results are consistent with above argument, i.e., the NMSE performance of predicting 20 and 30-days ahead is poor. But the D_{stat} performance is totally different; our results show that predictions over longer periods often have higher D_{stat} (in order to verify the robustness, we also used

three-fourths of the observations as a training dataset to reexamine the results of Tables 2 and 3. The related outcomes are shown in Tables 4 and 5. In addition, we considered a hyperbolic tangent function as the activation function and reported related results in Tables 6 and 7).

In addition to the above experiments, the dataset of CNY and CNH series from 2011 to 2015 is considered. We use the same method of dividing the data into the training set and testing set. According to the experimental results in Tables 2 and 3, we do not report MLP models and only adopt a number of EMD-MLP and EMD-MLP* models to conduct predictions. The statistical results of NMSE and D_{stat} are shown in Tables 8 and 9, respectively.

TABLE 2: NMSE comparisons for different models.

	1-day	5-day	10-day	20-day	30-day
Panel A: MLP model					
MLP(3)	0.0231	0.0874	0.2118	0.3698	0.5888
MLP(5)	0.0204	0.0854	0.2088	0.3939	0.5723
MLP(5,3)	0.0285	0.0943	0.2088	0.3623	0.4746
MLP(6,4)	0.0376	0.0896	0.2021	0.3613	0.5453
Panel B: EMD-MLP model					
EMD(-1)-MLP(3)	0.0145***	0.0823	0.2120	0.3894	0.5217*
EMD(-1)-MLP(5,3)	0.0143***	0.0821**	0.2111	0.3836	0.5447
EMD(-2)-MLP(3)	0.0207	0.0450***	0.1613**	0.3760	0.5258
EMD(-2)-MLP(5,3)	0.0236**	0.0505***	0.1583**	0.3667	0.5258
EMD(-3)-MLP(3)	0.0434	0.0436***	0.0739**	0.2162**	0.4355***
EMD(-3)-MLP(5,3)	0.0419	0.0456***	0.0694**	0.2144**	0.3892**
Panel C: EMD-MLP* model					
EMD(0)-MLP(3)*	0.0087***	0.0217***	0.0410***	0.0806***	0.1244***
EMD(0)-MLP(5,3)*	0.0088***	0.0222***	0.0404***	0.0695***	0.1048***
EMD(-1)-MLP(3)*	0.0075***	0.0200***	0.0385***	0.0812***	0.1136***
EMD(-1)-MLP(5,3)*	0.0083***	0.0314***	0.0368***	0.0681***	0.1051***
EMD(-2)-MLP(3)*	0.0172*	0.0214***	0.0444***	0.0760***	0.1152***
EMD(-2)-MLP(5,3)*	0.0190**	0.0256***	0.0423***	0.0759***	0.1034***
Panel D: random walk model					
No drift	0.0144	0.0861	0.2349	0.5323	0.9357
With drift	0.0148	0.0973	0.2834	0.6981	1.2954

Consider the CNY from January 2, 2006, to December 21, 2015, with a total of 2584 observations. This table compares the forecasting performance, in terms of the NMSE, for the MLP, EMD-MLP, and EMD-MLP* models. We report the NMSE as percentage for l -day ahead predictions where $l = 1, 5, 10, 20, \text{ and } 30$. The DM test [38] is used to compare the forecast accuracy of EMD-MLP (EMD-MLP*) model and the corresponding MLP model. ***, **, and * denote statistical significance at 1%, 5%, and 10%, respectively. For each length of prediction, we mark the minimum NMSE as bold.

TABLE 3: D_{stat} comparisons for different models.

	1-day	5-day	10-day	20-day	30-day
Panel A: MLP model					
MLP(3)	51.74	53.49	63.49***	69.65***	70.72***
MLP(5)	51.98	55.17	62.17***	68.44***	71.08***
MLP(5,3)	50.18	54.93	65.78***	69.29***	73.39***
MLP(6,4)	48.98	55.41	62.65***	69.89***	71.20***
Panel B: EMD-MLP model					
EMD(-1)-MLP(3)	61.94***	59.74***	66.99***	69.04***	72.54***
EMD(-1)-MLP(5,3)	62.42***	57.69***	65.54***	68.92***	71.32***
EMD(-2)-MLP(3)	62.30***	69.47***	69.64***	69.41***	72.42***
EMD(-2)-MLP(5,3)	60.02***	66.71***	70.24***	70.25***	72.54***
EMD(-3)-MLP(3)	57.50***	72.96***	78.67***	75.33***	73.63***
EMD(-3)-MLP(5,3)	56.78***	69.59***	80.48***	74.37***	73.27***
Panel C: EMD-MLP* model					
EMD(0)-MLP(3)*	69.99***	80.65***	81.45***	82.59***	86.39***
EMD(0)-MLP(5,3)*	68.31***	79.45***	80.36***	82.95***	85.78***
EMD(-1)-MLP(3)*	71.67***	81.01***	81.08***	81.86***	84.08***
EMD(-1)-MLP(5,3)*	70.35***	80.17***	82.89***	82.95***	86.39***
EMD(-2)-MLP(3)*	63.51***	80.89***	80.72***	82.10***	86.15***
EMD(-2)-MLP(5,3)*	63.51***	78.73***	82.41***	81.98***	86.39***

Consider CNY from January 2, 2006, to December 21, 2015, with a total of 2584 observations. This table compares the forecasting performance, in terms of the D_{stat} , for the MLP, EMD-MLP, and EMD-MLP* models. We report D_{stat} as percentage for l -day ahead predictions where $l = 1, 5, 10, 20, \text{ and } 30$. Moreover, we examine the ability of all models to predict the direction of change by the DAC test [40]. ***, **, and * denote statistical significance at 1%, 5%, and 10%, respectively. For each length of prediction, we mark the maximum D_{stat} as bold.

Tables 8 and 9 indicate the EMD-MLP* models perform better than the EMD-MLP models on both NMSE and D_{stat} criteria. Comparing the results of the CNY and CNH series, the NMSE of CNY is found to be less than that of CNH on

average, no matter how long the forecasting period. The main reason for this is the original volatility of CNY series is smaller than that of CNH series. For the D_{stat} test part, forecasting performance improves with the growth of

TABLE 4: Robust tests for NMSE comparisons with different subsamples.

	1-day	5-day	10-day	20-day	30-day
Panel A: MLP model					
MLP(3)	0.0175	0.0740	0.2010	0.3638	0.5954
MLP(5)	0.0201	0.0699	0.2082	0.3663	0.5158
MLP(5,3)	0.0212	0.0761	0.2117	0.3614	0.5561
MLP(6,4)	0.0221	0.0694	0.2066	0.3586	0.5746
Panel B: EMD-MLP model					
EMD(-1)-MLP(3)	0.0137***	0.0705	0.1444**	0.3827	0.5885
EMD(-1)-MLP(5,3)	0.0117***	0.0713	0.1982***	0.3251**	0.4948***
EMD(-2)-MLP(3)	0.0192	0.0371***	0.1434**	0.3568	0.5725*
EMD(-2)-MLP(5,3)	0.0222	0.0380***	0.1433**	0.3380*	0.5417
EMD(-3)-MLP(3)	0.0423	0.0468***	0.0691***	0.2003**	0.4676**
EMD(-3)-MLP(5,3)	0.0442	0.0410***	0.0647***	0.1965**	0.4551**
Panel C: EMD-MLP* model					
EMD(0)-MLP(3)*	0.0079***	0.0201***	0.0391***	0.0798***	0.1302***
EMD(0)-MLP(5,3)*	0.0084***	0.0213***	0.0461***	0.0614***	0.1160***
EMD(-1)-MLP(3)*	0.0083***	0.0192***	0.0413***	0.0717***	0.1298***
EMD(-1)-MLP(5,3)*	0.0091***	0.0220***	0.0425***	0.0655***	0.1140***
EMD(-2)-MLP(3)*	0.0191	0.0222***	0.0456***	0.0801***	0.1294***
EMD(-2)-MLP(5,3)*	0.0201	0.0240***	0.0381***	0.0660***	0.1050***

Consider the CNY from January 2, 2006, to December 21, 2015, with a total of 2584 observations. For training the neural network models, three-fourths of the observations are randomly assigned to the training dataset and the remainder is used as the testing dataset. This table compares the forecasting performance, in terms of the NMSE, for the MLP, EMD-MLP, and EMD-MLP* models. We report the NMSE as percentage for l -day ahead predictions where $l = 1, 5, 10, 20, \text{ and } 30$. The DM test [38] is used to compare the forecast accuracy of EMD-MLP (EMD-MLP*) model and the corresponding MLP model. ***, **, and * denote statistical significance at 1%, 5%, and 10%, respectively. For each length of prediction, we mark the minimum NMSE as bold.

TABLE 5: Robust tests for D_{stat} comparisons with different subsamples.

	1-day	5-day	10-day	20-day	30-day
Panel A: MLP model					
MLP(3)	49.76	56.98	63.43***	70.13***	71.15***
MLP(5)	45.96	56.98	66.61***	69.81***	72.12***
MLP(5,3)	53.88**	55.56	64.07***	69.49***	70.67***
MLP(6,4)	49.92	59.68**	63.12***	68.85***	71.47***
Panel B: EMD-MLP model					
EMD(-1)-MLP(3)	62.12***	57.14*	71.54***	69.17***	71.79***
EMD(-1)-MLP(5,3)	62.76***	56.98	67.25***	70.93***	74.04***
EMD(-2)-MLP(3)	64.03***	71.59***	72.18***	68.05***	70.99***
EMD(-2)-MLP(5,3)	62.92***	72.70***	72.02***	69.97***	72.12***
EMD(-3)-MLP(3)	55.94***	70.16***	79.01***	75.88***	74.36***
EMD(-3)-MLP(5,3)	55.63***	75.24***	80.60***	76.04***	73.08***
Panel C: EMD-MLP* model					
EMD(0)-MLP(3)*	69.89***	79.37***	81.08***	80.35***	86.86***
EMD(0)-MLP(5,3)*	72.11***	82.54***	80.60***	82.43***	86.38***
EMD(-1)-MLP(3)*	67.83***	81.90***	81.40***	81.31***	87.50***
EMD(-1)-MLP(5,3)*	68.46***	81.59***	82.19***	80.67***	87.18***
EMD(-2)-MLP(3)*	61.81***	80.63***	81.24***	80.35***	87.18***
EMD(-2)-MLP(5,3)*	63.23***	80.32***	82.83***	81.47***	86.70***

Consider the CNY from January 2, 2006, to December 21, 2015, with a total of 2584 observations. For training the neural network models, three-fourths of the observations are randomly assigned to the training dataset and the remainder is used as the testing dataset. This table compares the forecasting performance, in terms of D_{stat} , for the MLP, EMD-MLP, and EMD-MLP* models. We report D_{stat} as percentage for l -day ahead predictions where $l = 1, 5, 10, 20, \text{ and } 30$. Moreover, we examine the ability of all models to predict the direction of change by the DAC test [40]. ***, **, and * denote statistical significance at 1%, 5%, and 10%, respectively. For each length of prediction, we mark the maximum D_{stat} as bold.

forecasting periods. Also, no significant difference is found for the results of CNY and CNH series based on D_{stat} criteria. For forecasting over 20 days ahead, the hitting rate exceeds 89%. Even for forecasting 1-day ahead, D_{stat} can achieve more than 79%.

3.3. Applying the Trading Strategy. This part introduces an application for designing a trading strategy for CNY (since the cases of CNY and CNH, from 2011 to 2015, produce the similar results, we do not report the latter in this paper). According to the results in Tables 2 and 3, in terms of NMSE

TABLE 6: Robust tests for NMSE comparisons with different activation functions.

	1-day	5-day	10-day	20-day	30-day
Panel A: MLP model					
MLP(3)	0.0281	0.0877	0.2123	0.3729	0.5579
MLP(5)	0.0245	0.0820	0.2075	0.3729	0.5149
MLP(5,3)	0.0260	0.0859	0.2079	0.3491	0.4450
MLP(6,4)	0.0222	0.0887	0.1996	0.3436	0.4733
Panel B: EMD-MLP model					
EMD(-1)-MLP(3)	0.0120***	0.0830	0.2045	0.3717	0.5285
EMD(-1)-MLP(5,3)	0.0137***	0.0862	0.2097	0.3662	0.5282
EMD(-2)-MLP(3)	0.0204**	0.0354***	0.1466***	0.3654	0.5571
EMD(-2)-MLP(5,3)	0.0182**	0.0351***	0.1636**	0.3665	0.4772
EMD(-3)-MLP(3)	0.0417	0.0450***	0.0705***	0.2140**	0.5782
EMD(-3)-MLP(5,3)	0.0393	0.0441***	0.0711***	0.2198**	0.3996
Panel C: EMD-MLP* model					
EMD(0)-MLP(3)*	0.0077***	0.0226***	0.0491***	0.0820***	0.1246***
EMD(0)-MLP(5,3)*	0.0084***	0.0198***	0.0382***	0.0669***	0.1106***
EMD(-1)-MLP(3)*	0.0095***	0.0223***	0.0482***	0.0826***	0.1528***
EMD(-1)-MLP(5,3)*	0.0084***	0.0258***	0.0460***	0.0845***	0.1210***
EMD(-2)-MLP(3)*	0.0212**	0.0239***	0.0469***	0.0892***	0.1316***
EMD(-2)-MLP(5,3)*	0.0196**	0.0229***	0.0414***	0.0784***	0.1497***

Consider the CNY from January 2, 2006, to December 21, 2015, with a total of 2584 observations. In this table, a hyperbolic tangent function is used as the activation function. This table compares the forecasting performance, in terms of the NMSE, for the MLP, EMD-MLP, and EMD-MLP* models. We report the NMSE as percentage for l -day ahead predictions where $l = 1, 5, 10, 20, \text{ and } 30$. The DM test [38] is used to compare the forecast accuracy of EMD-MLP (EMD-MLP*) model and the corresponding MLP model. ***, **, and * denote statistical significance at 1%, 5%, and 10%, respectively. For each length of prediction, we mark the minimum NMSE as bold.

TABLE 7: Robust tests for D_{stat} comparisons with different activation functions.

	1-day	5-day	10-day	20-day	30-day
Panel A: MLP model					
MLP(3)	50.06	55.17	62.41***	68.56***	71.81***
MLP(5)	48.98	56.85*	63.25***	69.17***	72.54***
MLP(5,3)	48.14	54.81	65.30***	66.99***	72.17***
MLP(6,4)	49.94	56.25**	61.45***	68.68***	72.42***
Panel B: EMD-MLP model					
EMD(-1)-MLP(3)	69.99***	58.41***	67.95***	70.01***	71.45***
EMD(-1)-MLP(5,3)	67.95***	57.81***	64.46***	70.25***	73.39***
EMD(-2)-MLP(3)	64.59***	73.20***	70.36***	70.37***	71.08***
EMD(-2)-MLP(5,3)	63.51***	75.84***	69.40***	69.89***	74.00***
EMD(-3)-MLP(3)	57.26***	72.48***	78.80***	75.21***	68.89***
EMD(-3)-MLP(5,3)	57.38***	71.88***	80.00***	73.40***	73.39***
Panel C: EMD-MLP* model					
EMD(0)-MLP(3)*	71.07***	78.97***	80.96***	82.71***	86.76***
EMD(0)-MLP(5,3)*	72.03***	78.85***	80.36***	82.10***	87.48***
EMD(-1)-MLP(3)*	68.19***	77.52***	80.72***	81.62***	84.81***
EMD(-1)-MLP(5,3)*	70.11***	77.28***	80.12***	83.07***	86.03***
EMD(-2)-MLP(3)*	61.82***	78.97***	81.08***	80.77***	85.18***
EMD(-2)-MLP(5,3)*	64.71***	78.85***	78.80***	82.59***	83.35***

Consider the CNY from January 2, 2006, to December 21, 2015, with a total of 2584 observations. In this table, a hyperbolic tangent function is used as the activation function. This table compares the forecasting performance, in terms of D_{stat} , for the MLP, EMD-MLP, and EMD-MLP* models. We report D_{stat} as percentage for l -day ahead predictions where $l = 1, 5, 10, 20, \text{ and } 30$. Moreover, we examine the ability of all models to predict the direction of change by the DAC test [40]. ***, **, and * denote statistical significance at 1%, 5%, and 10%, respectively. For each length of prediction, we mark the maximum D_{stat} as bold.

and D_{stat} , the best model for predicting l -day ahead can be determined. The forecasting model can be trained based on past exchange rate series and produce l -day ahead predictions. According to the forecasting results, trading strategies are set as follows:

$$\begin{aligned} \text{long: if } \hat{p}_{s+l} &> p_s \times (1 + \tau), \\ \text{short: if } \hat{p}_{s+l} &< p_s \times (1 - \tau), \end{aligned} \quad (11)$$

where \hat{p}_{s+l} is the l -day ahead predictions at time s and τ is a critical number. We long (short) the CNY if l -day ahead

TABLE 8: NMSE comparisons for CNY and CNH.

	1-day	5-day	10-day	20-day	30-day
Panel A: CNY					
EMD(-1)-MLP(5,3)	0.2078*	0.7330***	4.9711	11.4496	14.8922
EMD(-2)-MLP(5,3)	0.3417	0.6292***	2.7752**	8.4836	8.5811
EMD(-3)-MLP(5,3)	0.6015	0.6098***	1.2525***	3.9844***	7.2250
EMD(0)-MLP(3)*	0.2171*	0.4204***	0.6304***	1.6314***	2.1985***
EMD(0)-MLP(5,3)*	0.1711**	0.4124***	0.6912***	1.5200***	1.6964***
EMD(-1)-MLP(3)*	0.1497**	0.3928***	0.7120***	1.8627***	2.1755***
EMD(-1)-MLP(5,3)*	0.1610*	0.4774***	0.5999***	1.2579***	1.5669***
EMD(-2)-MLP(3)*	0.2847	0.4485***	0.8156***	1.3609***	1.9762***
EMD(-2)-MLP(5,3)*	0.2960	0.4846***	0.8578***	1.2355***	1.4642***
Panel B: CNH					
EMD(-1)-MLP(5,3)	0.3814***	2.2893	6.4050	12.8439	20.9652
EMD(-2)-MLP(5,3)	0.4732**	1.1208**	4.7997*	12.2240	22.9451
EMD(-3)-MLP(5,3)	0.9658	0.9270***	2.3329***	7.6194**	19.4264
EMD(0)-MLP(3)*	0.2317***	0.8170***	1.2729***	2.5680**	3.5197***
EMD(0)-MLP(5,3)*	0.2600***	0.6797***	1.3611***	2.4448**	3.2915**
EMD(-1)-MLP(3)*	0.2699***	0.6471***	1.4162***	2.2991***	3.8647***
EMD(-1)-MLP(5,3)*	0.2379***	0.7165***	1.3509***	2.3486**	3.0931**
EMD(-2)-MLP(3)*	0.4297*	0.6346***	1.2586***	2.4497**	4.4241***
EMD(-2)-MLP(5,3)*	0.4173**	0.7528***	1.1915***	2.5860**	3.1497**

Consider both the CNY and CNH from January 3, 2011, to December 21, 2015, with a total of 1304 observations. This table compares the forecasting performance, in terms of the NMSE, for several forecasting models. We report the NMSE as percentage for l -day ahead predictions where $l = 1, 5, 10, 20$, and 30. The DM test [38] is used to compare the forecast accuracy of EMD-MLP (EMD-MLP*) model and the corresponding MLP model. ***, **, and * denote statistical significance at 1%, 5%, and 10%, respectively. For each length of prediction, we mark the minimum NMSE as bold.

TABLE 9: D_{stat} comparisons for CNY and CNH.

	1-day	5-day	10-day	20-day	30-day
Panel A: CNY					
EMD(-1)-MLP(5,3)	73.65***	75.00***	64.27***	60.90***	64.90***
EMD(-2)-MLP(5,3)	67.49***	77.97***	69.23***	68.67***	77.78***
EMD(-3)-MLP(5,3)	61.82***	78.71***	83.13***	77.44***	78.28***
EMD(0)-MLP(3)*	75.37***	79.70***	83.62***	86.22***	84.60***
EMD(0)-MLP(5,3)*	80.05***	82.43***	84.86***	89.47***	88.13***
EMD(-1)-MLP(3)*	75.12***	83.17***	84.37***	87.22***	84.09***
EMD(-1)-MLP(5,3)*	74.88***	83.42***	85.36***	88.72***	86.11***
EMD(-2)-MLP(3)*	68.47***	82.43***	82.88***	88.72***	83.59***
EMD(-2)-MLP(5,3)*	69.70***	82.67***	83.37***	86.97***	87.88***
Panel B: CNH					
EMD(-1)-MLP(5,3)	73.24***	62.10***	62.50***	62.38***	65.84***
EMD(-2)-MLP(5,3)	70.32***	76.28***	66.67***	63.37***	64.84***
EMD(-3)-MLP(5,3)	61.07***	79.46***	78.92***	75.00***	75.31***
EMD(0)-MLP(3)*	79.81***	82.64***	84.07***	84.65***	89.03***
EMD(0)-MLP(5,3)*	76.64***	83.13***	83.09***	85.40***	89.78***
EMD(-1)-MLP(3)*	76.64***	82.15***	84.80***	84.65***	84.79***
EMD(-1)-MLP(5,3)*	78.83***	83.37***	85.05***	84.65***	89.53***
EMD(-2)-MLP(3)*	71.53***	83.86***	84.31***	87.87***	85.29***
EMD(-2)-MLP(5,3)*	72.99***	84.60***	83.82***	83.42***	88.78***

Consider both the CNY and CNH from January 3, 2011, to December 21, 2015, with a total of 1304 observations. This table compares the forecasting performance, in terms of D_{stat} , for several forecasting models. We report D_{stat} as percentage for l -day ahead predictions where $l = 1, 5, 10, 20$, and 30. Moreover, we examine the ability of all models to predict the direction of change by the DAC test [40]. ***, **, and * denote statistical significance at 1%, 5%, and 10%, respectively. For each length of prediction, we mark the maximum D_{stat} as bold.

predictions are greater (smaller) than the present price multiplied by $1 + \tau$. So, unless $\hat{p}_{s+l} = p_s(1 + \tau)$, people always conduct a transaction at time s and hold it for l days. For instance, consider the case with $l = 10$ and $\tau = 0$. If $p_s = 5$ and we use the EMD-MLP model to find $\hat{p}_{s+l} = 5.2$, we

immediately long CNY and hold it for 10 days. On the contrary, if $p_s = 5$ and $\hat{p}_{s+l} = 4.9$, we short it.

The reason of setup τ comes from the following several aspects: first, when we set $\tau = 0$, the number of transactions must be very large. Since high frequency trading comes with

TABLE 10: Performance of trading strategy based on the best EMD-MLP.

		N_s	SD%	Return with transaction cost (%)			
				0.0	0.1	0.2	0.3
Panel A: $\tau = 0$							
1-day	EMD(-1)-MLP(3)*	812	1.46	13.59	-11.41	-36.41	-61.41
5-day	EMD(-1)-MLP(3)*	822	1.53	7.12	2.12	-2.88	-7.88
10-day	EMD(-1)-MLP(5,3)*	830	1.70	6.19	3.69	1.19	-1.31
20-day	EMD(-1)-MLP(5,3)*	823	1.76	4.98	3.73	2.48	1.23
30-day	EMD(-2)-MLP(5,3)*	823	1.73	4.59	3.76	2.92	2.09
Panel B: $\tau = 0.5\%$							
5-day	EMD(-1)-MLP(3)*	32	3.65	36.78	31.78	26.78	21.78
10-day	EMD(-1)-MLP(5,3)*	112	2.80	19.18	16.68	14.18	11.68
20-day	EMD(-1)-MLP(5,3)*	220	2.03	12.15	10.90	9.65	8.40
30-day	EMD(-2)-MLP(5,3)*	364	1.75	8.30	7.47	6.64	5.80
Panel C: $\tau = 1\%$							
5-day	EMD(-1)-MLP(3)*	5	4.93	85.02	80.02	75.02	70.02
10-day	EMD(-1)-MLP(5,3)*	17	4.72	40.20	37.70	35.20	32.70
20-day	EMD(-1)-MLP(5,3)*	71	2.29	18.42	17.17	15.92	14.67
30-day	EMD(-2)-MLP(5,3)*	131	1.72	13.08	12.24	11.41	10.58

In terms of NMSE and D_{stat} , we select the best models to forecasting l -day ahead predictions where $l = 1, 5, 10, 20$, and 30. The third column shows the sample size in each model and is denoted as N_s . By the trading strategy rule as (11) with $\tau = 0$, we generate N_s l -day returns. The fourth column reports the standard deviation of the annualized returns without transaction cost. The last four columns represent the averages of N_s annualized returns with different transaction costs.

extremely high transaction costs, the erosion of profits should not be ignored and it usually makes the trading strategy fail in reality. Therefore, letting τ be larger than zero can reduce the number of trades. More importantly, τ could be regarded as a confidence level indicator, which produces long and short trading thresholds, i.e., $p_s(1 + \tau)$ and $p_s(1 - \tau)$. Only when the l -day ahead prediction \hat{p}_{s+l} is larger (smaller) than the thresholds, we long (short) the CNY. Specifically, when we set $\tau = 1\%$, denoting that if the forecast price exceeds the present price by more than 1%, we buy it; on the contrary, if the forecast price exceeds the present price by less than 1%, we short it. To fit the practical situation, this empirical analysis also considers annual returns with transaction costs of 0.0%, 0.1%, 0.2%, and 0.3%, respectively.

Table 10 discusses the performance of the above trading strategy with $\tau = 0, 0.5\%, 1\%$, based on best models, selected according to Tables 2 and 3. It displays the best performance models with different forecasting periods. The returns shown in the last four columns have been adjusted to annual returns using different transaction costs. Firstly, it can be seen that the larger the l is, which means the forecasting period is longer, the larger the standard deviation will be. If there is no transaction cost, the return decreases with the increase of the forecasting period l . For instance, in Panel A, choosing a forecasting period of 1 day with $\tau = 0$ and zero transaction cost, the annual return will reach 13.59% by choosing the best performance model EMD(-1)-MLP(3)*. However, extending the forecasting period to 30 days, the return will become 4.59% with the best performance model.

However, if transaction costs are considered, it is clear that the return increases as the forecasting period becomes longer, which is opposite to the case where there are no transaction costs. The annual return suffers significant falls in short forecasting periods after considering the transaction

cost since the annual return of a short period is offset by the significant cost of high-frequency trading. For instance, with a 0.3% trading cost, the annual return of the best performance model for 30-day ahead forecasting is positive and is only 2.09%, but the annual return of EMD(-1)-MLP(3)*, which is the best performance for forecasting 1-day ahead, reaches as low as -61.41%. The annual return decreases with increasing transaction costs for every forecasting period.

Panels B and C, respectively, display trading strategy performance based on the best model when $\tau = 0.5\%$ and $\tau = 1\%$. The 1-day forecasting case is ignored in this experiment since the predicted value of the 1-day period does not exceed τ . The standard deviation in the fourth column decreases with the growth of forecasting periods. Another finding in these two panels is that the annual return decreases as the transaction cost becomes larger in the same forecasting period, which is similar to Panel A. However, with the same transaction cost, the annual return decreases with the growth of the forecasting period, which is contrary to the situation in Panel A. This may be the result of setting $\tau = 0.5\%$ or $\tau = 1\%$; the annual return will not be eroded by trading costs.

When we set $\tau = 0.5\%$ for 5-day ahead forecasting, although the average annual return with different trading costs is at least 21.78%, there are only 32 trading activities in a total of 832 trading days. However, in choosing long period forecasting, such as 20 days ahead, trading activity is 220 with a 8.4% average annual return. In Panel C, the trading activities are less than in Panel B. If a 0.3% trading cost with $\tau = 1\%$ is considered, the annual return of 30-day ahead forecasting will exceed 10%, and there are more than 130 trading activities. It can be seen that trading activities reduce with the growth of critical number τ . To sum up, applying EMD-MLP* to forecast the CNY could produce some useful trading strategies, even considering reasonable trading costs.

4. Conclusions

In this paper, RMB exchange rate forecasting is investigated and trading strategies are developed based on the models constructed. There are three main contributions in this study. First, since the RMB's influence has been growing in recent years, we focus on the RMB, including the onshore RMB exchange rate (CNY) and offshore RMB exchange rate (CNH). We use daily data to forecast RMB exchange rates with different horizons based on three types of models, i.e., MLP, EMD-MLP, and EMD-MLP*. Our empirical study verifies the feasibility of these models. Secondly, in order to develop reliable forecasting models, we consider not only the MLP model but also the hybrid EMD-MLP and EMD-MLP* models to improve forecasting performance. Among all the selected models, the EMD-MLP* performs best, in terms of both NMSE and D_{stat} criteria. It should be noted that the EMD-MLP* is different from the methodology proposed by Yu et al. [27]. We regard some IMF components as noise factors and delete them to reduce the volatility of the RMB. The empirical experiments verify that the above process could clearly improve forecasting performance. For example, Table 1 shows that the best models are EMD(-1)-MLP(3)*, EMD(-1)-MLP(5,3)*, and EMD(-2)-MLP(5,3)*. Finally, we choose the best forecasting models to construct the trading strategies by introducing different critical numbers and considering different transaction costs. As a result, we abandon the trading strategy of $\tau = 0$ since the annual returns are mostly negative. However, given $\tau = 0.5\%$ or 1% , all annual returns are more than 5% , even including the transaction cost. In particular, by setting $\tau = 1\%$ with 0.3% transaction cost, the annual return is more than 10% in different forecasting horizons. Thus, this trading strategy can help investors make decisions about the timing of RMB trading.

As the Chinese government continues to promoting RMB internationalization, RMB currency trading becomes increasingly important in personal investment, corporate financial decision-making, government's economic policies, and international trade and commerce. This study is trying to apply the neural network into financial forecasting and trading area. Based on the empirical analysis, the forecasting accuracy and trading performance of RMB exchange rate can be enhanced. Considering the performance of this proposed method, the study should be attractive not only to policymakers and investment institutions but also to individual investors who are interested in RMB currency or RMB-related products.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Authors' Contributions

All authors have contributed equally to this article.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (NSFC no. 71961007, 71801117, and 71561012). It was also supported in part by Humanities and Social Sciences Project in Jiangxi Province of China (JJ18207 and JD18094) and Science and Technology Project of Education Department in Jiangxi Province of China (GJJ180278, GJJ170327, and GJJ180245).

References

- [1] M. Fratzscher and A. Mehl, "China's dominance hypothesis and the emergence of a tri-polar global currency system," *The Economic Journal*, vol. 124, no. 581, pp. 1343–1370, 2014.
- [2] C. R. Henning, "Choice and coercion in East Asian exchange-rate regimes," *Power in a Changing World Economy, Routledge*, pp. 103–124, 2013.
- [3] C. Shu, D. He, and X. Cheng, "One currency, two markets: the renminbi's growing influence in Asia-Pacific," *China Economic Review*, vol. 33, pp. 163–178, 2015.
- [4] A. Subramanian and M. Kessler, "The renminbi bloc is here: Asia down, rest of the world to go?," *Journal of Globalization and Development*, vol. 4, no. 1, pp. 49–94, 2013.
- [5] R. Craig, C. Hua, P. Ng, and R. Yuen, "Chinese capital account liberalization and the internationalization of the renminbi," *IMF Working Paper*, 2013.
- [6] M. Funke, C. Shu, X. Cheng, and S. Eraslan, "Assessing the CNH-CNY pricing differential: role of fundamentals, contagion and policy," *Journal of International Money and Finance*, vol. 59, pp. 245–262, 2015.
- [7] M. Khashei, M. Bijari, and G. A. Raissi Ardali, "Improvement of auto-regressive integrated moving average models using fuzzy logic and artificial neural networks (ANNs)," *Neurocomputing*, vol. 72, no. 4–6, pp. 956–967, 2009.
- [8] Z.-X. Wang and Y.-N. Chen, "Nonlinear mechanism of RMB real effective exchange rate fluctuations since exchange reform: empirical research based on smooth transition autoregression model," *Financial Theory & Practice*, vol. 6, p. 4, 2016.
- [9] G. Zhang, B. Eddy Patuwo, and M. Y. Hu, "Forecasting with artificial neural networks: the state of the art," *International Journal of Forecasting*, vol. 14, no. 1, pp. 35–62, 1998.
- [10] G. P. Zhang, "An investigation of neural networks for linear time-series forecasting," *Computers & Operations Research*, vol. 28, no. 12, pp. 1183–1202, 2001.
- [11] C. H. Aladag and M. M. Marinescu, "TI/Euro and Leu/Euro exchange rates forecasting with artificial neural networks," *Journal of Social and Economic Statistics*, vol. 2, no. 2, pp. 1–6, 2013.
- [12] L. Yu, G. Hang, L. Tang, Y. Zhao, and K. Lai, "Forecasting patient visits to hospitals using a WD&ANN-based decomposition and ensemble model," *Eurasia Journal of Mathematics, Science and Technology Education*, vol. 13, no. 12, pp. 7615–7627, 2017a.
- [13] Y. Kajitani, K. W. Hipel, and A. I. McLeod, "Forecasting nonlinear time series with feed-forward neural networks: a case study of Canadian lynx data," *Journal of Forecasting*, vol. 24, no. 2, pp. 105–117, 2005.

- [14] L. Yu, Y. Zhao, and L. Tang, "Ensemble forecasting for complex time series using sparse representation and neural networks," *Journal of Forecasting*, vol. 36, no. 2, pp. 122–138, 2017.
- [15] L. Cao, "Support vector machines experts for time series forecasting," *Neurocomputing*, vol. 51, pp. 321–339, 2003.
- [16] L. Yu, H. Xu, and L. Tang, "LSSVR ensemble learning with uncertain parameters for crude oil price forecasting," *Applied Soft Computing*, vol. 56, pp. 692–701, 2017b.
- [17] L. Yu, Z. Yang, and L. Tang, "Quantile estimators with orthogonal pinball loss function," *Journal of Forecasting*, vol. 37, no. 3, pp. 401–417, 2018.
- [18] M. Álvarez-Díaz and A. Álvarez, "Forecasting exchange rates using genetic algorithms," *Applied Economics Letters*, vol. 10, no. 6, pp. 319–322, 2003.
- [19] C. Neely, P. Weller, and R. Dittmar, "Is technical analysis in the foreign exchange market profitable? A genetic programming approach," *The Journal of Financial and Quantitative Analysis*, vol. 32, no. 4, pp. 405–426, 1997.
- [20] L. Yu, S. Wang, and K. K. Lai, *Foreign-exchange-rate Forecasting with Artificial Neural Networks*, vol. 107, Springer Science & Business Media, Berlin, Germany, 2010.
- [21] G. P. Zhang, "Time series forecasting using a hybrid ARIMA and neural network model," *Neurocomputing*, vol. 50, pp. 159–175, 2003.
- [22] M. Álvarez-Díaz and A. Álvarez, "Genetic multi-model composite forecast for non-linear prediction of exchange rates," *Empirical Economics*, vol. 30, no. 3, pp. 643–663, 2005.
- [23] L. Zhao and Y. Yang, "PSO-based single multiplicative neuron model for time series prediction," *Expert Systems with Applications*, vol. 36, no. 2, pp. 2805–2812, 2009.
- [24] W. K. Wong, M. Xia, and W. C. Chu, "Adaptive neural network model for time-series forecasting," *European Journal of Operational Research*, vol. 207, no. 2, pp. 807–816, 2010.
- [25] L. Yu, Y. Zhao, L. Tang, and Z. Yang, "Online big data-driven oil consumption forecasting with google trends," *International Journal of Forecasting*, vol. 35, no. 1, pp. 213–223, 2019.
- [26] N. E. Huang, Z. Shen, S. R. Long et al., "The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis," *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, vol. 454, no. 1971, pp. 903–995, 1998.
- [27] L. Yu, S. Wang, and K. K. Lai, "Forecasting crude oil price with an EMD-based neural network ensemble learning paradigm," *Energy Economics*, vol. 30, no. 5, pp. 2623–2635, 2008.
- [28] C.-F. Chen, M.-C. Lai, and C.-C. Yeh, "Forecasting tourism demand based on empirical mode decomposition and neural network," *Knowledge-Based Systems*, vol. 26, pp. 281–287, 2012.
- [29] C.-S. Lin, S.-H. Chiu, and T.-Y. Lin, "Empirical mode decomposition-based least squares support vector regression for foreign exchange rate forecasting," *Economic Modelling*, vol. 29, no. 6, pp. 2583–2590, 2012.
- [30] L. Tang, Y. Wu, and L. Yu, "A non-iterative decomposition-ensemble learning paradigm using RVFL network for crude oil price forecasting," *Applied Soft Computing*, vol. 70, pp. 1097–1108, 2018.
- [31] M. Álvarez Díaz, "Speculative strategies in the foreign exchange market based on genetic programming predictions," *Applied Financial Economics*, vol. 20, no. 6, pp. 465–476, 2010.
- [32] W. Huang, K. Lai, Y. Nakamori, and S. Wang, "An empirical analysis of sampling interval for exchange rate forecasting with neural networks," *Journal of Systems Science and Complexity*, vol. 16, no. 2, pp. 165–176, 2003b.
- [33] A. Nikolsko-Rzhevskyy and R. Prodan, "Markov switching and exchange rate predictability," *International Journal of Forecasting*, vol. 28, no. 2, pp. 353–365, 2012.
- [34] M. Riedmiller and H. Braun, "A direct adaptive method for faster backpropagation learning: the RPROP algorithm," in *Proceedings of the IEEE International Conference on Neural Networks*, pp. 586–591, IEEE, San Francisco, CA, USA, March–April 1993.
- [35] N. E. Huang and Z. Wu, "A review on Hilbert-Huang transform: method and its applications to geophysical studies," *Reviews of Geophysics*, vol. 46, no. 2, 2008.
- [36] N. E. Huang, M.-L. C. Wu, S. R. Long et al., "A confidence limit for the empirical mode decomposition and Hilbert spectral analysis," *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, vol. 459, no. 2037, pp. 2317–2345, 2003a.
- [37] J. Yao and C. L. Tan, "A case study on using neural networks to perform technical forecasting of forex," *Neurocomputing*, vol. 34, no. 1–4, pp. 79–98, 2000.
- [38] F. X. Diebold and R. S. Mariano, "Comparing predictive accuracy," *Journal of Business & Economic Statistics*, vol. 20, no. 1, pp. 134–144, 2002.
- [39] L. Yu, S. Wang, and K. K. Lai, "A novel nonlinear ensemble forecasting model incorporating GLAR and ANN for foreign exchange rates," *Computers & Operations Research*, vol. 32, no. 10, pp. 2523–2541, 2005.
- [40] M. H. Pesaran and A. Timmermann, "A simple non-parametric test of predictive performance," *Journal of Business & Economic Statistics*, vol. 10, no. 4, pp. 461–465, 1992.
- [41] S. Anatolyev and A. Gerko, "A trading approach to testing for predictability," *Journal of Business & Economic Statistics*, vol. 23, no. 4, pp. 455–461, 2005.

Research Article

A Study of RMB Internationalization Path Based on Border Area Perspective

Po Sheng Ko,¹ Cheng Chung Wu ,² Ying Shih Mai,³ and Zhongrong Xu⁴

¹Department of Wealth and Taxation Management, National Kaohsiung University of Science and Technology, Kaohsiung, Taiwan

²Department of Finance Suqian College, Suqian, Jiangsu, China

³Department of International Business, National Kaohsiung University of Science and Technology, Kaohsiung, Taiwan

⁴Department of Marketing, Suqian College, Suqian, Jiangsu, China

Correspondence should be addressed to Cheng Chung Wu; wu_0110@yahoo.com.tw

Received 24 June 2019; Accepted 30 August 2019; Published 29 October 2019

Guest Editor: Benjamin M. Tabak

Copyright © 2019 Po Sheng Ko et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

At present, the enhancement of China's comprehensive national strength, stable currency policy, and the new round of opening-up strategy layout have provided opportunities for the financial development in border areas, especially for the RMB internationalization advancement in border areas, while the lagging financial system construction in current border areas also challenges its RMB internationalization. Based on Friedman's monetary demand theory (1970), the thesis has combined the periodical characteristics and selected the representative Yunnan province in the border areas as the object, taking into overall consideration the two paths of geography and function for RMB internationalization in border areas, so as to build the two-dimensional path theory framework of RMB internationalization from a border area perspective. In view of the above, a panel econometric regression model is built to estimate the factors influencing RMB internationalization path in Yunnan province. The results show that the process of opening up to the outside world in Yunnan province is its major driving force for RMB internationalization, while the great fluctuation of price level tends to hinder the process advancement and such possible impact of overall revenue on RMB internationalization development as the regional disparity.

1. Introduction

The ever-increasing comprehensive national strength of China, gradual expansion of RMB circulation scale, and steady promotion of RMB international status have accelerated RMB internationalization. As of the end of June 2015, RMB cross-border trade settlement has reached the cumulative amount of 14.6 trillion and the direct investment of RMB has come to the cumulative amount of 2.2 trillion, with RMB exchange reaching 3.2 trillion and global market share of RMB international payment accounting for 2.18%, rising up to no. 5. In addition to the increasing RMB internationalization scale, with the deepening of the major strategy of "One Belt and One Road," the border areas (China's border areas include nine provinces as Yunnan, Guangxi, Tibet, Xinjiang, Gansu, Inner Mongolia, Liaoning,

Jilin, and Heilongjiang) of China have been transformed from the open "end" to open "front." Various policies, especially financial policies, such as financial comprehensive reform pilot area construction in Yunnan and Guangxi border areas, regional open financial system construction in three provinces in Northeast China, and regional financial innovation district construction in Xinjiang Horgos, have promoted the continuous acceleration of RMB regional internationalization in border areas. Relying on its regional, resource, and humanitarian advantages, Yunnan province has taken the lead in its distinctive RMB internationalization path as compared with other border provinces and regions. Yunnan province established "estuary mode" for border trade bank settlement as early as 1994 and set a precedent of China's RMB cash exit and entry dispatching business for border trade settlement in 2003; then in 2004, it received the

approval of enjoying tax refund policy in small-scale border trade RMB settlement with neighboring countries and tried out RMB settlement verification system for general trade in 2005, conducting pilot RMB settlement for goods trade with ASEAN in 2008. In 2010, Yunnan province launched pilot cross-border RMB settlement and obtained the approval of financial comprehensive reform pilot area construction in 2013. After over twenty years of development, the promotion of RMB internationalization development in Yunnan province has gradually converted from single drive by border trade to multicore drive by various cross-border businesses, with financial organization system construction expanded to multiple levels and financial infrastructure facilities constantly improved.

At present, China's border areas have become the meeting point of various policy plannings which is conducive to the promotion of RMB internationalization in border areas, while all provinces in border areas have lagged behind for a long time, whether in economic strength or social development, as compared with other provinces in China, and such gap exerts great influence in RMB internationalization process in the area. According to statistics, the cross-border RMB business settlement amount in China's border areas was lower than the national average level from 2011 to 2014, with the gap of about 2.54%; meanwhile, the growth rate of cross-border RMB business settlement in China's border areas also lagged behind the average level, with an annual 70.40% on average left behind. In further analysis, from 2011 to 2014, the cross-border RMB business settlement amount in Yunnan province had been growing at an average annual rate of 45.75%, ranking no. 5 in China's border areas, as one of the lowest-ranked locations.

Since RMB internationalization process in China's border areas lags behind the national average level in both scale and growth rate, during the important period of opportunities for openness in China's border areas currently, it is worth deeply discussing how to speed up the development process of RMB internationalization in border areas so as to promote the overall improvement of RMB internationalization level and the balanced development within the region. The thesis is firstly based on literature of RMB internationalization path and monetary demand theory and builds the theoretical framework of RMB internationalization analysis in border areas, then selects the representative Yunnan province for empirical analysis, and finally draws the conclusion and puts forward the RMB internationalization development path in border areas for reference.

2. Literature Review

2.1. Evolution Mode of RMB Internationalization Path

2.1.1. Monetary Function Path of RMB Internationalization. Monetary function path of RMB internationalization refers to the analysis of the phased evolution for such four functions of international currency as trade settlement, investment currency, credit financing, and reserve currency. As to trade settlement path of RMB internationalization,

some scholars at home and abroad think that the trade settlement scale determines RMB internationalization path at the present stage and RMB settlement scale for goods trade is higher than that for service trade and trade of other current accounts in terms of structure [1, 2]. As regards investment currency and credit financing paths for RMB internationalization, Frankel [3] considers that the realization of RMB internationalization investment and financing paths in China at present mainly rely on the offshore market for capital flow; in particular, Hong Kong offshore market plays an important role. Lyrtzakakis [4] offers an examination of the main economic and political determinants of Renminbi internationalization, both at the domestic and international levels. The analysis suggests that domestic economic determinants such as capital account openness, financial market development, and interest rate/exchange rate liberalization must be seen as necessary conditions for internationalization. Various domestic political actors determine the success of internationalization indirectly, by affecting the implementation of the necessary economic reforms and conditions. It is further argued that even though several domestic economic and political conditions are necessary for internationalization, they are not a priori sufficient, as international economic and political factors also play an important role in the internationalization process. Through this analysis, the paper ultimately illustrates that the issue of RMB internationalization needs to be understood and analyzed within a political economy context rather than within a purely economic one.

2.1.2. Regional Evolution Path of RMB Internationalization.

Most scholars at home and abroad think that the regional evolution path of RMB internationalization should be as East Asianization and Southeast Asianization → Asianization → internationalization. As far as "East Asianization and Southeast Asianization" are concerned, Yang [5] carries out regression analysis in use of the currency data for the main countries in ASEAN and RMB data from 2000 to 2010 through establishing econometrics model and found that the status of dollar had gradually declined in "currency anchor" of main countries in ASEAN since the appreciation of RMB in 2005, with the position of RMB in ASEAN rising continuously and the "currency anchor" of main countries in ASEAN adjusted to RMB, Japanese Yen, and Euro. In terms of "Asianization," Peng et al. [6] apply SYRADF panel unit root test with Fourier transform to conduct empirical analysis of the economy convergence for the major 13 countries in Asia and further study the possibility for RMB to become the regional key currency. The results show that Japanese Yen remains the currency occupying an important position in Asia and RMB presents increasing impact in Asia region with great potential for becoming the regional key currency. According to Park [7], there are two options that could be taken in the following regional approach. One is creating an ASEAN + New 3 (the Chinese Mainland, Hong Kong, and Taiwan) RMB bloc, and the other is liberalizing China's financial industries and

internationalizing the RMB by playing a leading role in East Asia's economic integration within the framework of ASEAN + 3. This paper concludes that the latter is a more realistic and effective approach for China.

2.2. Research on RMB Internationalization in China's Border Areas. Tang et al. [8] conduct research on RMB regional internationalization path in Guangxi, propose cross-border settlement of RMB and mutual financial pilot points, and establish offshore RMB investment return project library and the development path for provincial RMB "asset pool." The research group of Xinjiang Financial Society [9] from the perspective of domestic and foreign regional layout puts forward the path in which Kazakhstan is chosen as the first country for promoting RMB regionalization beyond the border and Sino-Kazakhstan Economic and Trade Cooperation Center in Horgos is regarded as the pilot point for RMB regionalization in China at the same time. In view of such problems as lagged financial services in Inner Mongolia and Mongolia and lagged adjustment of dollar and RMB fund position in RMB account settlement, Wang [10] raises the path of expanding RMB settlement business in border trade and strengthening financial cooperation with surrounding countries.

At present, the literature about RMB regional internationalization in Yunnan province mainly includes the feasibility of implementation, obstacles, suggestions for development, and existing risks. Liu [11] considers that the feasibility of implementation for RMB regional internationalization in Yunnan includes the following: the first is the high degree of acceptance, large stock, and wide distribution in neighboring countries; the second is the early cross-border settlement, obvious geographic advantages, and numerous participating industries and countries; and the third is the early small currency conversion and distinctive characteristics. Sun [12] thinks that there still exist such problems in promoting RMB regional internationalization in Yunnan as impeded clearing system, unsound backflow channels, cumbersome process for case cross-border dispatching, and currency conversion to be further improved. Based on the above situation, he puts forward the suggestions of strengthening the financial cooperation with neighboring countries, researching and developing RMB cash backflow mechanism, and encouraging provincial financial institutions to go out. Ding et al. [13] conducts empirical analysis of the factors influencing RMB settlement in cross-border trade through establishing econometrics model, including such factors as local total output, local currency supply, and national identity of trade target country. The study by Wu and Tang [14] explores what factors influence RMB internationalization in the process of the Belt and Road. It firstly makes a summary of the important influencing factors, then sets up a semi-logarithmic model to quantitatively analyze these factors, and finally puts forward suggestions for the steady development of RMB internationalization. Through literature review, it is found that since the Belt and Road initiative was put forward, RMB has been used more extensively in the

surrounding countries and regions. Through quantitative research, it is found that with China's GDP as a share of world GDP, the scale of imports and exports of goods and services and economic freedom are all positively related to the internationalization of RMB; the inflation rate and the volatility of real effective exchange rate are negatively related to it.

2.3. Summary of Literature Review. The research on the theory of China's RMB internationalization path in existing literature can be divided into two angles: on one hand, RMB internationalization mainly concerns trade settlement, supplemented by offshore market, with currency swap as the complement; on the other hand, RMB internationalization follows "East Asianization and Southeast Asianization → Asianization" and provides reference for establishment of the theoretical framework for RMB internationalization path based on China's border area perspective in the thesis. In addition, research in the literature on the status quo of promoting RMB internationalization process in China's border area also lays a certain foundation for the research in the thesis. However, the existing literature presents mismatching of research situations between RMB internationalization path at national level and RMB internationalization path at border area level, with relative lagging of research for border areas. Therefore, it is in urgent need of building the analysis framework of RMB internationalization path in border area in combination with the practical conditions.

These new data sources are particularly relevant in the absence of reliable data on economic outcomes, such as tracking and setting poverty targets in developing countries [15]. Jean et al. [16] trained neural networks to predict local economic outcomes based on satellite data from five African countries. Machine learning can also draw economic forecasts from large-scale network data; for example, Blumenstock et al. [17] use mobile data to measure wealth, enabling them to quantify poverty in Rwanda at the individual level. Image recognition can certainly be used outside of satellite data, and localized predictions of economic outcomes are relevant outside developing countries.

3. Theoretical Framework

3.1. Research Hypothesis of RMB Internationalization Path from the Border Area Perspective. Effectively, internationalization of yuan has been pursued along two interrelated tracks [18]. One track focuses on cultivating use of the currency in foreign trade. At the official level, swap agreements have been arranged with an increasing number of foreign central banks, facilitating expanded use of the RMB as a means of payment [19]. By mid-2016, some three dozen agreements had been signed totaling more than RMB 3.3 trillion (C\$480 billion). At the private level, regulations have been gradually eased to permit more import and export transactions to be settled in yuan, bypassing traditional invoicing currencies like the dollar. The second track focuses on use of the RMB in international finance as a store of value.

Currently, the research theory of evolution mode for RMB internationalization path mainly focuses on the national level and seldom establishes the research analysis framework for RMB internationalization path reflecting geographical features in terms of a certain region. For this reason, the thesis has integrated Figure 1 and proposed the research hypothesis of RMB internationalization path in conformity with the characteristics of border areas:

The first is the regional path hypothesis of RMB internationalization from the border area perspective. Learning from the regional evolution path theory of “surrounding usage → regionalization → internationalization” at the national level for RMB internationalization, in consideration of the differences between national path and border area path for RMB internationalization, it is proposed that the regional evolution path for border areas should follow the hypothesis of border-oriented first and then internationalization, so as to reflect the location characteristics of the border areas in connection with the surrounding countries.

The second is the functional path hypothesis of RMB internationalization from the border area perspective. Learning from the functional evolution path theory of “trade settlement → investment currency → credit financing → reserve currency” at the national level for RMB internationalization, in consideration of the differences between national path and border area path for RMB internationalization, the hypothesis mainly focusing on such three functions as trade settlement, investment currency, and credit financing has been put forward from the border area perspective.

3.2. Theoretical Model of RMB Internationalization Path from the Border Area Perspective. Friedman [20] presents money demand function from the perspective of a country’s monetary demand as follows:

$$\frac{M}{P} = f\left(y, w, r_m, r_b, r_e, \frac{1}{P} \frac{dP}{dt}; u\right). \quad (1)$$

In the function, y refers to actual income, w means the proportion of nonhuman wealth in total wealth (namely, the proportion of income from property in total income), r_m represents the expected nominal yield of currency, r_b means the expected nominal yield of term bond, r_e refers to the expected nominal yield of stock, $(1/P)(dP/dt)$ means the expected change rate of commodity price (namely, reflecting inflation), and u stands for non-revenue factor.

According to money demand function, the main ideas of Friedman’s monetary demand theory can be summarized as the following aspects: firstly, in the economic society, money is a kind of asset and a form of holding wealth by people; the second is about the total amount of wealth held by people in various forms, and the difference in individual total wealth tends to influence his/her demand for money; the third is about the anticipated returns of wealth in all forms. In the economic society, people hold the total wealth in different forms; fourthly, the function can not only indicate the monetary demand of the ultimate wealth owner, but also

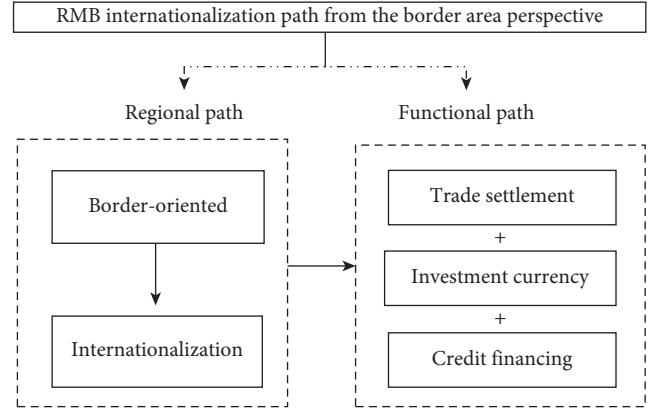


FIGURE 1: Two-dimensional research hypothesis of RMB internationalization path from the border area perspective.

represent the aggregate demand for a country’s money from the whole international community. Based on the above ideas, the thesis draws on the money demand function (formula (1)) established by Friedman [20], in comprehensive consideration of the two-dimensional hypotheses in accordance with regional and functional paths in border areas, and builds the following theoretical model of RMB internationalization path for border areas:

$$M_i = (y_i)^{\beta_1} \cdot (X_i)^{\beta_2} \cdot (R_i)^{\beta_3} \cdot \left(\frac{1}{P_i} \frac{dP_i}{dt}\right)^{\beta_4} \cdot u_i. \quad (2)$$

In the model, M_i refers to the RMB demand in the border country or border area i ; y_i means the overall revenue in the border area i ; X_i indicates the opening-up scale of the border area i , mainly reflecting the monetary function; R_i shows the wealth conditions in the border area i ; $(1/P_i)(dP_i/dt)$ represents the price change in the border area i ; u_i refers to the impact of other factors on the RMB demand in the border area i ; β_1 , β_2 , β_3 , and β_4 , respectively, represent the influence coefficient of various factors on the RMB demand in the border area i .

According to the above theoretical model, the main points of RMB internationalization path theory from the border area perspective proposed in the thesis include the following aspects: firstly, the demand for RMB in the border areas is affected by the opening-up scale of the region, namely, the functional path of foreign trade, investment currency, and credit financing; secondly, the demand for RMB in the border areas is influenced by the overall income, wealth of residents and enterprises, and price fluctuation in the region; thirdly, the regional characteristics of the border area and nonborder area are differentiated by the subscript i in the model.

3.3. Indicator System and Research Object

3.3.1. Indicator System and Research Hypothesis. In accordance with the selection of influencing factors for RMB internationalization path in border areas, the indicator system of influencing factors for RMB internationalization path in border areas has been established, as shown in Table 1.

TABLE 1: Indicator system of influencing factors for RMB internationalization path in border areas.

First class indicators		Second class indicators		Correlation	Variable type
Name	Code	Name	Code		
Demand for RMB	M	Cross-border RMB business settlement amount	kj	—	Explained variable
Overall revenue	Y	Local GDP	gdp	Positive correlation	
Opening-up scale	X	Total volume of foreign trade	trad	Positive correlation	Explanatory variable
		Foreign direct investment amount	fdi	Positive correlation	
Wealth conditions	R	Balance of deposits and loans in domestic and foreign currency for financial institutions	fina	Positive correlation	
Price change	$(1/P)(dP/dt)$	Consumer price index	cpi	Negative correlation	

The research hypotheses in five aspects have been put forward in view of the above indicator system:

H1: the local GDP in border areas presents the positive correlation with the demand for RMB in border areas

H2: the foreign trade volume in border areas presents the positive correlation with the demand for RMB in border areas

H3: the foreign investment amount in border areas presents the positive correlation with the demand for RMB in border areas

H4: the balance of deposits and loans in domestic and foreign currency for financial institutions in border areas presents the positive correlation with the demand for RMB in border areas

H5: consumer price index in border areas presents the negative correlation with the demand for RMB in border areas

3.3.2. Research Object. The research object of the thesis mainly refers to Yunnan province in China. Yunnan province covers a total area of about 390 thousand square kilometers and accounts for 4.11% of the country's territory, ranking no. 8 among the provincial administrative regions nationwide in terms of the area. The boundary line of Yunnan province is 4060 kilometers, bordering on such three countries as Burma, Laos and Vietnam. As to the concrete division, on one hand, Yunnan province was approved by China to carry out cross-border RMB settlement pilot work in 2011, and at the end of 2015, only the data for the past five years could be collected; for the econometric model with five explanatory variables, the regression results will be of significance with the minimum samples of freedom at 15. For this purpose, the thesis takes 16 cities and prefectures in Yunnan province into consideration from the cross-section angle to build the panel econometric model; in this way, the sample size increases to 80, meeting the requirements of regression. On the other hand, in order to show the innovativeness of the thesis and reveal the regional heterogeneity, Yunnan province is divided into such three regions as southern Yunnan, central Yunnan, and northern Yunnan, so as to reflect the differences of RMB internationalization path in different regions; see Table 2 for the detailed division.

3.4. Econometric Model and Data Specification

3.4.1. Econometric Model. Take the logarithm at both sides of the theoretical model (formula (2)) equation and establish the following four panel data econometric models of influencing factors for RMB internationalization in Yunnan province, in combination with the specific secondary indicator of the influencing factors for RMB internationalization path in the selected border area:

$$\ln kj_{it} = c + \alpha_1 \cdot \ln gdp_{it} + \alpha_2 \ln trad_{it} + \alpha_3 \ln fdi_{it} + \alpha_4 \ln fina_{it} + \alpha_5 \ln cpi_{it} + \mu_{it}, \quad (3)$$

$$\ln kj_{1t} = c_1 + \beta_1 \cdot \ln gdp_{1t} + \beta_2 \ln trad_{1t} + \beta_3 \ln fdi_{1t} + \beta_4 \ln fina_{1t} + \beta_5 \ln cpi_{1t} + \mu_{1t}, \quad (4)$$

$$\ln kj_{2t} = c_2 + \gamma_1 \cdot \ln gdp_{2t} + \gamma_2 \ln trad_{2t} + \gamma_3 \ln fdi_{2t} + \gamma_4 \ln fina_{2t} + \gamma_5 \ln cpi_{2t} + \mu_{2t}, \quad (5)$$

$$\ln kj_{3t} = c_3 + \delta_1 \cdot \ln gdp_{3t} + \delta_2 \ln trad_{3t} + \delta_3 \ln fdi_{3t} + \delta_4 \ln fina_{3t} + \delta_5 \ln cpi_{3t} + \mu_{3t}. \quad (6)$$

In the above four equations, equation (3) refers to the overall panel data econometric model for all the 16 cities and prefectures in Yunnan province; equations (4)–(6) indicate the panel data econometric models for the three regions of southern Yunnan, central Yunnan, and northern Yunnan in Yunnan province; c represents the constant term of the model; μ represents the random disturbance term of the model; α , β , γ , and σ show the estimated parameters before all explanatory variables.

3.4.2. Data Specification. Combine the specific indicator category in the above influencing factor indicator system for RMB internationalization path of Yunnan province and search the department concerned and relevant statistical yearbook, so as to collect and sort the panel data of influencing factor indicator from 2011 to 2015. The data for various indicators come from Kunming central subbranch of the People's Bank of China, statistical yearbook of Yunnan province, China's regional economic statistical yearbook, and Yunnan survey yearbook. On the basis of various indicator data collected and sorted, with the purpose of reducing the impact of absolute data dimension on regression results, as well as the fluctuation range of the data, take the

TABLE 2: Division of all regions in Yunnan province.

Region	Include cities and prefectures	Division explanation
Southern Yunnan (8 cities and prefectures)	Baoshan city, Pu'er city, Lincang city, Honghe prefecture, Wenshan prefecture, Xishuangbanna prefecture, Dehong prefecture, Nujiang prefecture	All the cities and prefectures in this region directly neighbor the surrounding countries, as the border cities and prefectures
Central Yunnan (4 cities and prefectures)	Kunming city, Yuxi city, Qujing city, Dali prefecture	The cities and prefectures in this region have relatively developed economic and social status
Northern Yunnan (4 cities and prefectures)	Chuxiong prefecture, Lijiang city, Diqing prefecture, Zhaotong city	The cities and prefectures in this region are adjacent to other provinces in China

logarithm of the five absolute indicators, namely, cross-border RMB settlement, local GDP, total volume of foreign trade, foreign direct investment amount, and the balance of deposits and loans for financial institutions.

4. Empirical Analysis

The collected panel data of various regression variables are composed of five time dimensions and sixteen cross-section dimensions, namely, "short panel" data, with the sample size 80. See Table 3 for the detailed description of such statistical properties as the mean value, standard deviation, maximum, and minimum for all variable samples.

Applied to F -test and Hausman test for the panel data models established in the thesis and determine the form of the four panel data models as entity-fixed effect models. As to the entity effect models, considering that the cross section data of samples are more than the "short panel" traits of time series, it is advisable to adopt the generalized least squares (GLS) method to evaluate and obtain the estimated results of overall model and divisional models.

4.1. Significance of Coefficient for Overall and Divisional Models. The results of the overall model 3 processed with GLS estimation are shown in Table 4. The results show that the significant influencing factors for cross-border RMB settlement in Yunnan province are the total volume of foreign trade, foreign direct investment amount, and consumer price index. According to the concrete analysis, one percent point rise of the total volume of foreign trade in Yunnan province will make the cross-border RMB settlement scale in Yunnan province increase by 0.5514% on average; one percent point rise of foreign direct investment amount in Yunnan province will make the cross-border RMB settlement scale in Yunnan province increase by 0.2031% on average; one percent point rise of consumer price index in Yunnan province will make the cross-border RMB settlement scale in Yunnan province decline by 0.0507% on average, while such two factors as local GDP and the balance of deposits and loans in domestic and foreign currency for financial institutions in Yunnan province have not exerted remarkable influence on the cross-border RMB settlement in Yunnan province yet. From the analysis of statistical properties, adjusted R -squared (R^2) is relatively high, F statistical value is high, and DW statistics are close to 2, indicating the sound fitting degree of overall model,

significant overall coefficient, and no first-order autocorrelation for stochastic error term.

The results of the divisional model 4, 5, and 6 processed with GLS estimation are shown in Table 5. The results show that the factors influencing the regional RMB settlement vary for different regions. In southern Yunnan, the significant factors influencing cross-border RMB settlement scale are the total volume of foreign trade and foreign direct investment amount; to be more specific, one percent point rise of the total volume of foreign trade in southern Yunnan will make the cross-border RMB settlement scale increase by 1.5331% on average; one percent point rise of foreign direct investment amount in southern Yunnan will make the cross-border RMB settlement scale increase by 0.5748% on average. In central Yunnan, the significant factors influencing cross-border RMB settlement scale are the local GDP, total volume of foreign trade, foreign direct investment amount, and consumer price index; to be more specific, one percent point rise of the local GDP in central Yunnan will make the cross-border RMB settlement scale increase by 2.1006% on average; one percent point rise of the total volume of foreign trade in central Yunnan will make the cross-border RMB settlement scale increase by 0.2019% on average; one percent point rise of foreign direct investment amount in central Yunnan will make the cross-border RMB settlement scale increase by 0.2024% on average; one percent point rise of consumer price index in central Yunnan will make the cross-border RMB settlement scale decline by 0.0783% on average. As to northern Yunnan, the significant factors influencing cross-border RMB settlement scale are the total volume of foreign trade and consumer price index; to be more specific, one percent point rise of the total volume of foreign trade in northern Yunnan will make the cross-border RMB settlement scale increase by 2.7323% on average; one percent point rise of consumer price index in northern Yunnan will make the cross-border RMB settlement scale decline by 0.441% on average. Judging from the statistical properties of model 3.2 to 3.4, the adjusted R -squared (R^2) of the three models is relatively high, showing the sound fitting degree, and high F statistical value indicates the overall significance of all model coefficients, without first-order autocorrelation for stochastic error term of all models.

4.2. Deviation Degree of Intercept Term for Overall and Divisional Models. All the four panel models are entity effect models, mainly featuring the existence of individual influence

TABLE 3: Statistical representation of regression variable samples.

Variable	Mean value	Standard deviation	Minimum	Maximum	Sample size
ln kj	9.9674	1.4862	0.0000	15.7755	80
ln gdp	5.8895	0.0294	3.8720	8.1360	80
ln trad	9.7986	0.2055	5.5595	14.3401	80
ln fdi	7.6029	0.8676	0.0000	12.0996	80
ln fina	6.7282	0.0386	4.7449	9.8709	80
ln cpi	4.6534	0.0053	4.6269	4.7362	80

TABLE 4: Econometric regression results of overall model 3.

Variable indicator and code	Variable parameter	Estimated value of variable parameter
Constant term	c	3.5741
Local GDP (ln gdp)	α_1	0.9607
Total volume of foreign trade (ln trad)	α_2	0.5514***
Foreign direct investment amount (ln fdi)	α_3	0.2031***
Balance of deposits and loans in domestic and foreign currency for financial institutions (ln fina)	α_4	-0.1319
Consumer price index (ln cpi)	α_5	-0.0507*
Adjusted R -squared, R^2		0.9737
F statistical value		147.3611
DW statistics		1.4161

* and *** represent the 10% and 1% significance levels for t statistics of all parameters, respectively.

TABLE 5: Econometric regression results of divisional model 4 to 6.

Variable indicator and code	Southern Yunnan (model 4)		Central Yunnan (model 5)		Northern Yunnan (model 6)	
	Variable parameter	Estimated value of variable parameter	Variable parameter	Estimated value of variable parameter	Variable parameter	Estimated value of variable parameter
Constant term	c_1	-5.1907	c_2	3.3347**	c_3	41.657**
Local GDP (ln gdp)	β_1	1.5843	γ_1	2.1006***	σ_1	-6.3848
Total volume of foreign trade (ln trad)	β_2	1.5331***	γ_2	0.2019***	σ_2	2.7323***
Foreign direct investment amount (ln fdi)	β_3	0.5748***	γ_3	0.2124**	σ_3	0.0099
Balance of deposits and loans in domestic and foreign currency for financial institutions (ln fina)	β_4	-2.4729	γ_4	0.2849	σ_4	3.9652
Consumer price index (ln cpi)	β_5	0.0275	γ_5	-0.0783***	σ_5	-0.441***
Adjusted R -squared, R^2		0.6828		0.998163		0.850353
F statistical value		17.78842		1291.449		14.49568
DW statistics		1.415892		2.20737		2.53379

** and *** represent the 5% and 1% significance levels for t statistics of all parameters, respectively.

among the 16 cities and prefectures in Yunnan province, without structural changes, and the individual influence is mainly embodied in the differences of intercept term of models for all cities and prefectures. It is noteworthy that the individual influence obtained by means of Eviews8.0 regression reflects the deviation of all cross section members from the overall average state. In other words, the differences of the intercept term for all cities and prefectures estimated from the overall and divisional models in the thesis are reflected in the deviation of the spontaneous cross-border RMB settlement (spontaneous cross-border RMB settlement refers to the existing cross-border RMB settlement value in all cities and prefectures when all the five factors influencing cross-border RMB settlement are zero at the same time; the

indicator can reflect the found) of all regions from the provincial average spontaneous cross-border RMB settlement. The estimated deviation value for all regions is shown in Tables 6 and 7. The analysis of the deviation degree of intercept term in different regions indicates that the following cities and prefectures have higher spontaneous cross-border RMB settlement scale than the provincial average spontaneous cross-border RMB settlement in terms of overall model: Dehong Prefecture, Honghe Prefecture, Lincang City, Xishuangbanna Prefecture, Pu'er City, Kunming City, Yuxi City, and Dali Prefecture, and the rest cities and prefectures have the spontaneous cross-border RMB settlement scale close to or smaller than the provincial average scale. As to divisional models, the regions with the spontaneous cross-

TABLE 6: Estimated results of spontaneous cross-border RMB settlement deviation (c_i^*) for all regions in overall model.

Region i		c_i^*
Dehong prefecture	DH	3.3764
Honghe prefecture	HH	1.1860
Lincang city	LC	1.1045
Xishuangbanna prefecture	BN	0.8273
Pu'er city	PE	0.6407
Kunming city	KM	0.5432
Yuxi city	YX	0.4631
Dali prefecture	DL	0.3214
Baoshan city	BS	0.0003
Lijiang city	LJ	-0.0503
Chuxiong prefecture	CX	-0.8801
Nujiang prefecture	NJ	-1.0530
Wenshan prefecture	WS	-1.0720
Zhaotong city	ZT	-1.6371
Diqing prefecture	DQ	-1.8310
Qujing city	QJ	-1.9394

TABLE 7: Estimated results of spontaneous cross-border RMB settlement deviation (c_i^*) for all regions in divisional models.

Region i		c_i^*
Southern Yunnan (model 4)		
Dehong prefecture	DH	0.1714
Honghe prefecture	HH	0.1720
Pu'er city	PE	0.0462
Xishuangbanna prefecture	BN	-0.2731
Lincang city	LC	0.1175
Baoshan city	BS	-0.1394
Wenshan prefecture	WS	-0.1522
Nujiang prefecture	NJ	0.0576
Central Yunnan (model 5)		
Kunming city	KM	1.1264
Yuxi city	YX	0.7139
Dali prefecture	DL	0.7090
Qujing city	QJ	-2.5493
Northern Yunnan (model 6)		
Chuxiong prefecture	CX	0.3814
Lijiang city	LJ	-3.6449
Diqing prefecture	DQ	-1.6196
Zhaotong city	ZT	4.8830

border RMB settlement scale higher than the provincial average spontaneous cross-border RMB settlement scale include Dehong Prefecture, Honghe Prefecture, Lincang City, and Pu'er City in southern Yunnan; Kunming City, Yuxi City, and Dali Prefecture in central Yunnan; and Zhaotong City and Chuxiong Prefecture in northern Yunnan.

4.3. Robustness Test of Overall and Divisional Models. The cointegration test of model 3 to 6 in Table 8 shows that the probability p value of ADF statistics for all models is less than 0.05; that is to say, at the significance level of 5%, all models refuse to accept the original hypothesis without cointegrated model, which indicates that all panel models are cointegrated, and all regression coefficients are effective, with explanatory power.

TABLE 8: Kao test results of overall model and divisional models.

	Kao residual cointegration test		
	ADF	t statistic	Prob.
	Overall model (3)	Residual variance	-5.116029
	HAC variance	2.073562	—
	Kao residual cointegration test		
Southern Yunnan model (4)	ADF	t statistic	Prob.
	Residual variance	-2.452871	0.0071
	HAC variance	1.475345	—
	Kao residual cointegration test		
Central Yunnan model (5)	ADF	t statistic	Prob.
	Residual variance	-4.66203	0.0000
	HAC variance	0.036618	—
	Kao residual cointegration test		
Northern Yunnan model (6)	ADF	t statistic	Prob.
	Residual variance	-2.252246	0.0122
	HAC variance	3.807812	—

5. Conclusions and Suggestions

The research features innovative research perspective on one hand, based on the regional heterogeneity analysis, combining the characteristics of cross-border RMB development at the local level and integrating the two paths of monetary function and regional evolution for RMB internationalization into one analysis framework according to the money demand function established by Friedman [20], so as to set up the theoretical model of RMB internationalization path in border areas; on the other hand, the research has the innovative research method, trying to find out the influencing factors for the regional internationalization of RMB in Yunnan province and the influencing degree from such aspects as demand for RMB, overall revenue, opening-up scale, wealth conditions, and price change according to the theoretical model of RMB internationalization path in border areas by establishing panel regression model, in order to provide the empirical support for the design of RMB internationalization path in border areas. On the basis of the analysis from theoretical and empirical levels, integrating the research hypotheses in five aspects, the research results of the thesis are drawn as follows:

- (1) Hypothesis H1: the local GDP in border areas presents the positive correlation with the demand for RMB in border areas. The analysis in the thesis indicates that the local GDP in Yunnan province presents no obvious correlation with the demand for RMB in Yunnan province from the 16 cities and prefectures in the province as a whole, while the local GDP of central Yunnan region presents positive correlation with the demand for RMB in the region

from the regional perspective, consistent with the hypothesis.

- (2) Hypothesis H2: the foreign trade volume in border areas presents the positive correlation with the demand for RMB in border areas. The analysis in the thesis shows that the foreign trade volume in Yunnan province presents obvious positive correlation with the demand for RMB in Yunnan province, either from the overall or regional perspective, consistent with the hypothesis.
- (3) Hypothesis H3: the foreign investment amount in border areas presents the positive correlation with the demand for RMB in border areas. The analysis in the thesis indicates that the foreign investment amount of Yunnan province presents obvious positive correlation with the demand for RMB in Yunnan province, from the 16 cities and prefectures in the province as a whole, consistent with the hypothesis; in terms of regional aspect, the conditions in southern Yunnan and central Yunnan are consistent with the hypothesis, while the conditions in northern Yunnan shows no significant correlation.
- (4) Hypothesis H4: the balance of deposits and loans in domestic and foreign currency for financial institutions in border areas presents the positive correlation with the demand for RMB in border areas. The analysis in the thesis shows that the balance of deposits and loans in domestic and foreign currency for financial institutions in Yunnan province presents no significant correlation with the demand for RMB in Yunnan province.
- (5) Hypothesis H5: consumer price index in border areas presents the negative correlation with the demand for RMB in border areas. The analysis in the thesis indicates that the consumer price index in Yunnan province presents obvious negative correlation with the demand for RMB in Yunnan province, from the 16 cities and prefectures in the province as a whole, consistent with the hypothesis; in terms of regional aspect, the conditions in central Yunnan and northern Yunnan are consistent with the hypothesis, while the conditions in southern Yunnan shows no significant correlation.

On the basis of the above research results, the research conclusions in the following aspects for the thesis can be drawn: firstly, the factor of opening-up scale is the main driving factor promoting RMB internationalization in Yunnan province; secondly, the factor of price change hinders RMB internationalization process in Yunnan province to a certain extent; thirdly, the factor of overall revenue promotes the development of RMB internationalization only in some regions of Yunnan province; fourthly, the factor of wealth conditions has not exerted or released its driving potential for RMB internationalization process in Yunnan province.

The following four aspects of policy suggestions are put forward based on the conclusions from the thesis.

The first is to accelerate the innovation of financial products and expand regionalization of cross-border RMB investment and financing. Efforts should be made to expand RMB capital export business; focus on promotion of cross-border two-way RMB capital pool business, cross-border RMB central collection business and cross-border RMB loan business under current account, cross-border RMB settlement and RMB international investment, and loan fund business under personal current account; rely on the major investment projects abroad; actively seek and expand market; and give play to other monetary functions of cross-border RMB other than the trade settlement of cross-border RMB; furthermore, it is suggested to speed up the innovation of RMB overseas investment products, increase the product categories for overseas RMB valuation, and expand regionalization of cross-border RMB investment and financing.

The second is to improve the foreign trade environment and stabilize the status of trade settlement for cross-border RMB; try to promote the industrial upgrading, adjust the trade structure, improve the trade conditions, and boost the currency options of import and export enterprises in trade; implement the differentiated encouraging policy for general trade and bulk commodity trade in use of RMB settlement; establish the trade cooperation platform and consolidate the trade and investment scale with the neighboring countries through such effective ways as simplifying customs clearance formalities for goods trade (especially in the less developed area on the border), accelerating approval and registration progress for cross-border investment and reducing trade and investment costs.

The third is to promote cross-border financial cooperation and unbar the development channel of cross-border RMB business; concentrate efforts on promoting cross-border financial cooperation, deepening foreign exchange, strengthening the communication and cooperation with the central banks and commercial banks of neighboring countries, and set up the official communication and collaboration mechanism as soon as possible; further expand the cooperation with foreign institutions and improve the cross-border RMB clearing and settlement system with the surrounding countries and regions; expand the RMB backflow channels by means of overseas RMB loan, issuance of RMB bonds abroad, and permitting purchase of domestic enterprise equity by foreign legal persons and individuals.

The fourth is to promote financial infrastructure upgrades and optimize the support mechanism for cross-border RMB process; set up the research, development, and test center for financial engineering on the border and actively explore the construction of such financial infrastructures as financial private network, financial cloud computing center, integrated service platform with electronic payment, sharing platform for enterprise credit information, mobile financial public service platform, and financial information interaction platform; give great impetus to service facilitation, improve market participation, simplify formalities, reduce the examination and approval, lower the trade cost for enterprises, and enhance the service efficiency and optimize the construction of support mechanism for cross-border RMB process.

Data Availability

WIND database were used to support this study.

Conflicts of Interest

We declares that there is no conflict of interest regarding the publication of this paper.

References

- [1] B. Ermon, "China's challenge to the international monetary system: incremental steps and long-term prospects for internationalization of the renminbi," *Pacific Forum CSIS Issues and Insights*, vol. 9, no. 2, pp. 1–17, 2009.
- [2] F. Zhen, "The internationalization of renminbi: development, prospects and orientation," *Economic Theory and Business Management*, vol. 33, no. 5, pp. 22–31, 2014.
- [3] J. Frankel, "Internationalization of the RMB and historical precedents," *Journal of Economic Integration*, vol. 27, no. 3, pp. 329–365, 2012.
- [4] D. Lyratzakis, "The determinants of RMB internationalization: the political economy of a currency's rise," *American Journal of Chinese Studies*, vol. 21, no. 2, pp. 163–184, 2014.
- [5] R. H. Yang, "Effect of RMB circulation in surrounding countries and adjustment of currency anchor in ASEAN," *Journal of International Trade*, vol. 339, no. 3, pp. 61–68, 2011.
- [6] H. F. Peng, X. Y. Tan, W. B. Chen, and Y. L. Li, "Asian monetary cooperation and RMB regionalization process: an empirical research based on panel SURADF test with a fourier function," *World Economy Studies*, no. 1, pp. 36–47, 2015.
- [7] Y. C. Park, "RMB internationalization and its implications for financial and monetary cooperation in East Asia," *China & World Economy*, vol. 18, no. 2, pp. 1–21, 2010.
- [8] W. L. Tang, L. L. Qin, H. Sun, and L. Q. Huang, "The orientation of financial policy under the background of RMB regionalization in Guangxi," *Journal of Guangxi Financial Research*, no. 11, pp. 40–45, 2009.
- [9] Xinjiang Financial Society, "Research on the regionalization of RMB and the expansion of border trade settlement," *Xinjiang Finance*, vol. 29, no. 11, pp. 4–19, 2007.
- [10] X. Wang, "Local currency settlement of border trade and the development of RMB regionalization," *Heilongjiang Finance*, vol. 28, no. 10, pp. 29–30, 2007.
- [11] G. Liu, "A study of RMB internationalization," *Journal of the Party School of the Central Committee of the C.P.C.*, no. 6, pp. 55–59, 2012.
- [12] L. Sun, "Feasibility study on promoting RMB regionalization in Yunnan province," *Heilongjiang's Foreign Economic and Trade*, vol. 25, no. 8, pp. 63–66, 2011.
- [13] W. L. Ding, L. L. Yang, and F. J. Lin, "Research on the influencing factors of RMB settlement in cross-border trade: an analysis based on Yunnan data," *Guizhou Social Sciences*, vol. 296, no. 8, pp. 80–87, 2014.
- [14] T. T. Wu and R. Tang, "Research on the influencing factors of RMB internationalization in the process of the Belt and Road initiative," in *Proceedings of the 2018 2nd International Conference on Management, Education and Social Science (ICMESS 2018)*, Qingdao, China, June 2018.
- [15] J. E. Blumenstock, "Fighting poverty with data," *Science*, vol. 353, no. 6301, pp. 753–754, 2016.
- [16] N. Jean, M. Burke, M. Xie, W. M. Davis, D. B. Lobell, and S. Ermon, "Combining satellite imagery and machine learning to predict poverty," *Science*, vol. 353, no. 6301, pp. 790–794, 2016.
- [17] J. Blumenstock, G. Cadamuro, and R. On, "Predicting poverty and wealth from mobile phone metadata," *Science*, vol. 350, no. 6264, pp. 1073–1076, 2015.
- [18] P. Subacchi, *The People's Money: How China is Building a Global Currency*, Columbia University Press, New York, NY, USA, 2017.
- [19] S. Liao and D. McDowell, "No reservations: international order and demand for the renminbi as a reserve currency," *International Studies Quarterly*, vol. 60, no. 2, pp. 272–293, 2016.
- [20] M. Friedman, "A theoretical framework for monetary analysis," *Journal of Political Economy*, vol. 78, no. 2, pp. 193–238, 1970.

Research Article

A Hybrid Approach Using Oversampling Technique and Cost-Sensitive Learning for Bankruptcy Prediction

Tuong Le ¹, Minh Thanh Vo,² Bay Vo ³, Mi Young Lee ¹ and Sung Wook Baik ¹

¹Digital Contents Research Institute, Sejong University, Seoul, Republic of Korea

²Institute of Research and Development, Duy Tan University, Da Nang 550000, Vietnam

³Faculty of Information Technology, Ho Chi Minh City University of Technology (HUTECH), Ho Chi Minh, Vietnam

Correspondence should be addressed to Sung Wook Baik; sbaik@sejong.ac.kr

Received 30 April 2019; Accepted 11 July 2019; Published 5 August 2019

Guest Editor: Thiago C. Silva

Copyright © 2019 Tuong Le et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The diagnosis of bankruptcy companies becomes extremely important for business owners, banks, governments, securities investors, and economic stakeholders to optimize the profitability as well as to minimize risks of investments. Many studies have been developed for bankruptcy prediction utilizing different machine learning approaches on various datasets around the world. Due to the class imbalance problem occurring in the bankruptcy datasets, several special techniques would be used to improve the prediction performance. Oversampling technique and cost-sensitive learning framework are two common methods for dealing with class imbalance problem. Using oversampling techniques and cost-sensitive learning framework independently also improves predictability. However, for datasets with very small balancing ratios, combining two above techniques will produce the better results. Therefore, this study develops a hybrid approach using oversampling technique and cost-sensitive learning, namely, HAOC for bankruptcy prediction on the Korean Bankruptcy dataset. The first module of HAOC is oversampling module with an optimal balancing ratio found in the first experiment that will give the best overall performance for the validation set. Then, the second module uses the cost-sensitive learning model, namely, CBoost algorithm to bankruptcy prediction. The experimental results show that HAOC will give the best performance value for bankruptcy prediction compared with the existing approaches.

1. Introduction

Machine learning and data mining [1–9], which is the process of learning in order to look for patterns in observations or data and make better decisions in the future based on the training samples, is widely used in various fields such as cybernetics [10–14], engineering [15–18], bioinformatics [19], medical informatics [20], economics [21–27], etc. Especially in economics, there are many issues for optimizing profits in the business such as customer lifetime value modeling (CLVM), churn customer modeling (CCM), dynamic pricing, customer segmentation, recommendation systems, etc. CLVM [23] is one of the most important models for eCommerce business. These models can identify, understand, and retain the most valuable customers in your business. With the obtained results from these models, the business managers may make a better business strategy to optimize profitability. CCM [24] can help the companies determine

their customers who will stop using their services. The outputs of these models, the customer list, are important inputs of an algorithmic retention strategy because they help optimize discount offers, marketing campaigns, and other targeted marketing initiatives. Dynamic pricing models [25] are for flexibly pricing products based on several factors such as the level of interest of the target customer, demand of the market at the time of purchase, and whether the customer has engaged with a marketing campaign. Meanwhile, customer segmentation models [26, 27] group customers into personas based on specific variations among them using several clustering and classification algorithms. Recommendation systems are another major way by which machine learning proves its business value. Recommendation systems sift through large quantities of data to predict how likely any given customer is to purchase an item or enjoy a piece of content and then suggest those things to the user. The result is a customer experience that encourages better engagement

and reduces churn. Bank lending and systemic risk [28, 29] is another issue in economics sector that attracted a lot of attention. This model will find empirical evidence against diversification as a mean to reduce systemic risk.

Bankruptcy prediction is also a hot topic in the field of business attracted by many scientists on computer science as well as economics around the world. In computer science domain, bankruptcy prediction, a predictive machine learning model, is to analyze the financial statement of a firm to make predictions for its fate in the future. Based on the obtained results from this task, investors and managers will devise appropriate strategies for companies that are going bankrupt. Many studies have been developed in recent years to predict the firm bankruptcy using various approaches [30–32]. In 2015, Kim et al. [30] introduced an efficient boosting algorithm, namely, GMBost, using geometric mean for dealing with the problem of imbalanced data occurring in bankruptcy datasets. This algorithm calculates the error of majority class and the error of minority class separately. Then geometric mean value of these values will be determined to calculate the weight values for the next phase. Next, a novel approach [31] utilizing eXtreme Gradient Boosting (XGB) with synthetic features was proposed for bankruptcy prediction. In this study, the synthetic features proposed are automatically generated by random selection of two existing features and random selection of the arithmetical operation which help to improve the prediction performance. Recently, Barboza et al. [32] performed and evaluated several existing classification models including SVC (linear and RBF kernels), artificial neural networks (ANN), logistic regression, boosting, Random Forest, and Bagging, for forecasting bankruptcy companies. The authors use a balanced bankruptcy dataset that includes 449 bankruptcy firms and 449 non-bankruptcy firms from 1985 to 2005 for training the above classifiers. The trained models will be evaluated by an imbalanced bankruptcy dataset collected between 2006 and 2013 that consists of 133 bankruptcy cases and 13,300 non-bankruptcy cases. The experimental results in this study indicate that three classifiers including boosting, bagging, and random forest provide better results for bankruptcy prediction.

In many datasets on various domains, class distribution is commonly imbalanced called by class imbalance problem. The minority class in these datasets consists of a small number of data points while the majority class has a very large number of data points. Specifically, the number of bankruptcies is extremely small compared to the normal companies in bankruptcy datasets. The traditional classification models have a big bias towards majority class in such datasets. It is the cause of reduced performance of the above models. Therefore, many methods are given to deal with class imbalance problem which are grouped into the four following categories [33]. (1) *Algorithm level approaches* adapt existing classifiers to bias the learning toward the minority class [34, 35] without changing training data. (2) *Data level approaches* change the class distribution by resampling the data space [36, 37] to improve the predictive performance. There are three subcategories in this group including undersampling, oversampling, and hybrids techniques. Undersampling techniques balance the data distribution by removing the real

data samples in majority class while oversampling techniques add the synthetic data samples to minority class. Meanwhile, hybrids techniques combine both undersampling and oversampling techniques. (3) *Cost-sensitive learning framework* is the hybrid methods that combine data and algorithm level approaches. These frameworks add costs to data samples (data level) and modify the learning process to accept costs (algorithm level) [38, 39]. The classifier in this group is biased toward the minority class by assuming higher misclassification costs for this class and seeking to minimize the total cost errors of both classes. (4) *Ensemble-based methods* usually consist of a combination of an ensemble learning algorithm and one of the techniques above, specifically, data level and cost-sensitive ones [40]. By combining data level approach to the ensemble learning algorithm, the new hybrid method usually preprocesses the data before training each classifier, whereas cost-sensitive ensembles, instead of modifying the base classifier in order to accept costs in the learning process, guide the cost minimization via the ensemble learning algorithm. The above four methods are used depending on the datasets to improve performance.

In 2018, Le et al. [41] first introduced the Korean Bankruptcy dataset denoted by KRBDS. In this study, the authors presented the oversampling based (OSB) framework that utilizes the oversampling techniques, a technique belonging to data level approach, for dealing with the class imbalance problem to predict the bankruptcy. This framework found that SMOTE-ENN is the best oversampling technique for KRBDS. Then, Le et al. [42] proposed a cluster-based boosting (CBoost) algorithm for dealing with the class imbalance problem. CBoost approach is considered as a cost-sensitive learning framework for dealing with the class imbalance problem. The framework, namely, RFCI, based on CBoost algorithm achieves the best AUC (The area under the receiver operating characteristics curve) with a shorter processing time compared with the first framework and several methods for bankruptcy prediction. In this study, we propose a hybrid approach, namely, HAOC, that combines the oversampling technique and cost-sensitive learning framework together for bankruptcy prediction. Our proposed approach firstly uses SMOTE-ENN to adjust class distribution of KRBDS with specific balancing ratio. Then, HAOC will use CBoost algorithm to predict the bankruptcy. The first experiment was conducted to find the best normalization technique among StandardScaler, MinMaxScaler, and RobustScaler for KRBDS. The second experiment is to find the optimal balancing ratio for oversampling phase. The comparison between HAOC with the existing approaches will be evaluated in the third experiment.

The rest of this manuscript is structured as follows. Section 2 first summarized the experimental dataset, namely, KRBDS, an oversampling technique, namely, SMOTE-ENN, and the CBoost algorithm. As the main contribution of this study, Section 2 introduces the hybrid approach for bankruptcy prediction, namely, HAOC. Two experiments were conducted to find the optimal balancing ratio and to show the effectiveness of proposed approach for bankruptcy prediction. Finally, the conclusions as well as several future

TABLE 1: The statistical information of KRBDS.

Feature	Description	Max	Min	Mean	Standard Deviation	Median	P25	P75
F1	Current assets	2.2×10^{11}	0	2.2×10^7	9.2×10^8	2.2×10^6	8.0×10^5	6.5×10^6
F2	Fixed assets, or fixed capital property	9.5×10^{10}	0	2.9×10^7	6.5×10^8	1.4×10^6	2.9×10^5	6.8×10^6
F3	Total assets	2.5×10^{11}	0	6.2×10^7	1.7×10^9	4.5×10^6	1.5×10^6	1.5×10^7
F4	Current liabilities within one year	2.1×10^{11}	-1.2×10^6	1.8×10^7	8.9×10^8	1.1×10^6	2.9×10^5	5.2×10^6
F5	Non-current liabilities.	6.5×10^{11}	-7.7×10^5	2.2×10^7	2.5×10^9	4.2×10^5	1.2×10^4	2.2×10^6
F6	Total liabilities	6.5×10^{11}	-2.1×10^5	4.9×10^7	2.9×10^9	2.1×10^6	5.5×10^5	8.3×10^6
F7	Capital	1.6×10^{10}	-2.9×10^7	5.1×10^6	1.2×10^8	4.0×10^5	1.5×10^5	1.0×10^6
F8	Earned surplus	4.8×10^{10}	-6.4×10^{11}	1.4×10^6	2.5×10^9	8.3×10^5	1.2×10^5	3.2×10^6
F9	Total capital	5.5×10^{10}	-6.3×10^{11}	1.3×10^7	2.5×10^9	1.7×10^6	5.4×10^5	5.5×10^6
F10	Total capital after liabilities	2.5×10^{11}	-4.3×10^4	6.2×10^7	1.7×10^9	4.5×10^6	1.4×10^6	1.5×10^7
F11	Sales revenue	6.0×10^{10}	-1.4×10^9	3.6×10^7	5.2×10^8	5.1×10^6	1.8×10^6	1.5×10^7
F12	Cost of sales	5.4×10^{10}	-4.7×10^6	2.7×10^7	4.2×10^8	3.4×10^6	8.6×10^5	1.1×10^7
F13	Net profit	2.5×10^{10}	-2.6×10^{10}	7.3×10^6	1.6×10^8	1.1×10^6	4.2×10^5	3.1×10^6
F14	Sales and administrative expenses	1.3×10^{10}	-5.2×10^6	5.5×10^6	9.6×10^7	8.8×10^5	3.4×10^5	2.4×10^6
F15	Operating profit that refers to the profits earned through business operations	2.5×10^{10}	-2.6×10^{10}	1.9×10^6	1.1×10^8	1.9×10^5	3.6×10^4	6.5×10^5
F16	Non-operating income	1.0×10^{10}	-4.4×10^5	1.6×10^6	5.1×10^7	4.3×10^4	8.1×10^3	2.2×10^5
F17	Non-operating expenses	3.0×10^9	-5.5×10^5	1.6×10^6	2.8×10^7	6.6×10^4	1.2×10^4	3.2×10^5
F18	Income and loss before income taxes	2.8×10^{10}	-2.3×10^{10}	2.0×10^6	1.2×10^8	1.6×10^5	3.3×10^4	5.8×10^5
F19	Net income	2.8×10^{10}	-2.3×10^{10}	1.5×10^6	1.2×10^8	1.4×10^5	2.9×10^4	5.0×10^5

research issues related to bankruptcy prediction are given in Section 4.

2. Materials and Methods

This section firstly introduces the experimental dataset, namely, KRBDS. Then, we summarize the oversampling technique named SMOTE-ENN and the cost-sensitive learning framework named CBoost algorithm. Finally, the proposed approach, namely, HAOC, will be introduced.

2.1. The Experimental Dataset. KRBDS was first introduced by Le et al. [41] that was provided by a Korean financial company. From the financial statements released by Korean companies from 2016 to 2017, nineteen financial features that have frequently been used in the previous bankruptcy prediction studies including assets, liabilities, capital, profit, etc. were extracted. Assets are any resources owned by the business such as buildings, equipment, and stocks while a liability is defined as any type of borrowing from persons or banks for improving their business. In addition, capital is any economic resource used by entrepreneurs and businesses to buy what they need to make their products or to provide their services. Meanwhile, profit is a financial benefit that is realized when the amount of revenue gained from a business activity exceeds the expenses, costs, and taxes needed to

sustain the activity. These values are extremely important in finance to consider the company's performance, especially bankruptcy prediction. These features and some statistical information including maximum, minimum, and mean are shown and described in Table 1.

There are 307 bankrupted firms and 120,048 normal firms in KRBDS which has the balancing ratio of 0.0026. This ratio is extreme small for the normal classifier to predict bankruptcy correctly. Therefore, we need to develop several specific techniques to improve the performance.

2.2. Oversampling Technique with MOTE-ENN. Resampling technique belonging to data level approaches for dealing with class imbalance problem is the most common approach by adjusting the class distribution. Resampling technique consists of three subcategories including oversampling techniques, undersampling techniques, and hybrids techniques as illustrated in Figure 1. Undersampling technique balances the data distribution by removing the real data samples in majority class while oversampling technique accomplishes that purpose by adding the synthetic data samples to minority class. Meanwhile, hybrids methods combine both undersampling and oversampling techniques.

The advantage of these techniques is to balance the class distribution for improving the predictive performance. However, there is no absolute advantage of one resampling method

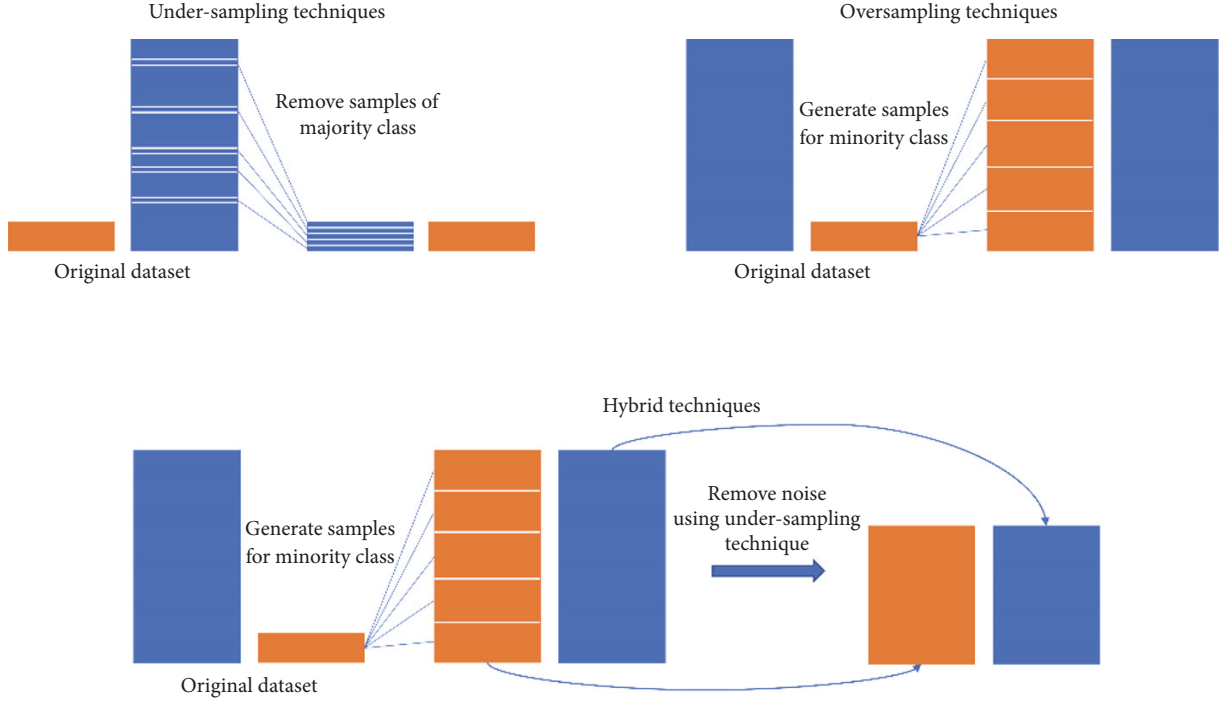


FIGURE 1: The illustration of oversampling, undersampling, and hybrids techniques.

over another. Application of these techniques depends on the use case it applies to and the dataset itself. Meanwhile, the disadvantage of undersampling techniques is that they can remove potentially useful data samples that could be important for the induction process. When the number of samples in the minority class is too small compared to that of samples in the majority class like KRBDS, undersampling techniques became ineffective. In this case, many samples in majority class are deleted. In addition, the main disadvantage with oversampling is that, by making exact copies of existing examples, it makes overfitting likely. A second disadvantage of oversampling is that it increases the number of training examples. Thus, the systems increase training time and the amount of memory required to hold the training set.

In 2018, Le et al. [41] conducted the oversampling framework that presents the empirical evaluation of oversampling techniques for bankruptcy prediction on KRBDS. Several oversampling techniques such as Synthetic Minority Over-sampling Technique (SMOTE) [36], Borderline-SMOTE [44], Adaptive Synthetic (ADASYN) sampling approach [45], SMOTE-ENN [46], and SMOTE-Tomek [46] were used to improve the bankruptcy prediction performance. The experiments conducted in this study found that SMOTE-ENN is the best oversampling technique for KRBDS. This approach is summarized as follows.

The SMOTE algorithm was first proposed by Chawla et al. [36] in 2002 that generates synthetic minority samples based on the feature similarities between the original minority samples. Firstly, SMOTE determines the k -nearest neighbors (NNs) which is denoted by $\mathcal{K}_{\mathbf{x}_i}$ for each minority sample $\mathbf{x}_i \in \chi_{min}$.

Figure 2(a) demonstrates the three NNs of \mathbf{x}_i that connect with \mathbf{x}_i by a line. To generate a synthetic data sample (\mathbf{x}_{new}) for \mathbf{x}_i , SMOTE randomly selects an element $\hat{\mathbf{x}}_i$ in $\mathcal{K}_{\mathbf{x}_i}$ and $\hat{\mathbf{x}}_i$ in χ_{min} . The feature vector of \mathbf{x}_{new} is the sum of the feature vectors of \mathbf{x}_i and the value, which can be obtained by multiplying the vector difference between \mathbf{x}_i and $\hat{\mathbf{x}}_i$ with a random value δ from 0 to 1 ($\delta \in [0, 1]$), as the following equation:

$$\mathbf{x}_{new} = \mathbf{x}_i + (\hat{\mathbf{x}}_i - \mathbf{x}_i) \times \delta \quad (1)$$

where $\hat{\mathbf{x}}_i$ is an element in $\mathcal{K}_{\mathbf{x}_i}$; $\hat{\mathbf{x}}_i \in \chi_{min}$.

According to (1), the synthetic sample is a point along the line segment joining \mathbf{x}_i and the randomly selected $\hat{\mathbf{x}}_i \in \mathcal{K}_{\mathbf{x}_i}$. Figure 2(b) shows a toy example of the SMOTE algorithm. The new sample \mathbf{x}_{new} is in the line between \mathbf{x}_i and $\hat{\mathbf{x}}_i$.

Then, SMOTE-ENN will apply the neighborhood cleaning rule based on the edited nearest neighbor (ENN) [46] to clean unwanted overlapping between classes, which removes samples that differ from two samples in the three nearest neighbors. Figure 3 shows the example of an ENN. Generally, SMOTE-ENN also uses SMOTE for the oversampling step and then uses ENN to remove the overlapping examples as shown in Figure 4.

2.3. Cluster-Based Boosting Algorithm. Recently, Le et al. [42] proposed CBoost algorithm that is based on the cost-sensitive learning framework for dealing with the class imbalance problem occurring in bankruptcy datasets effectively. CBoost algorithm first clusters the majority class in the bankruptcy datasets, i.e., the non-bankruptcy firms, by applied k -mean clustering with $k = 45$ which is considered as the best k value

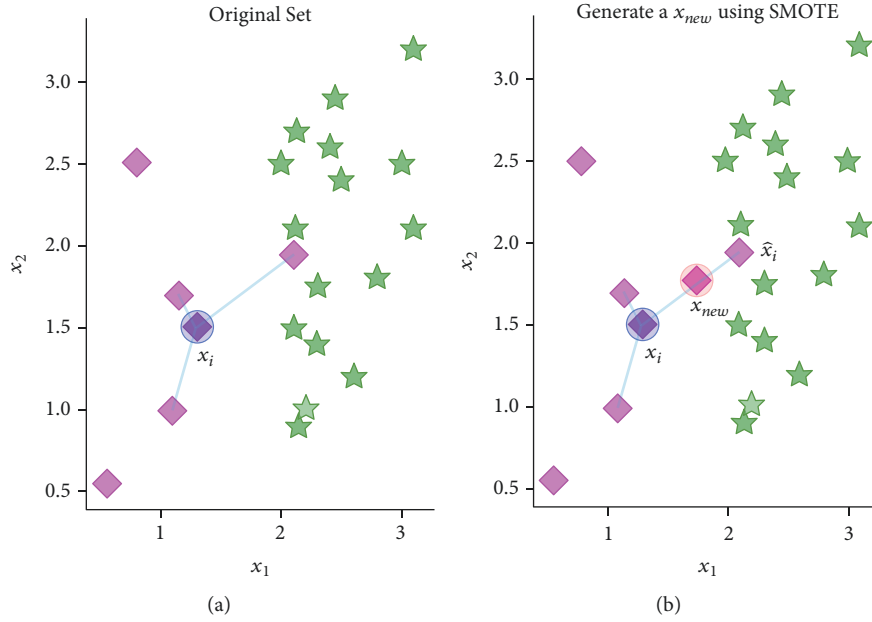


FIGURE 2: A toy example of the three-nearest neighbors for the x_i (a); and generate x_{new} using SMOTE (b).

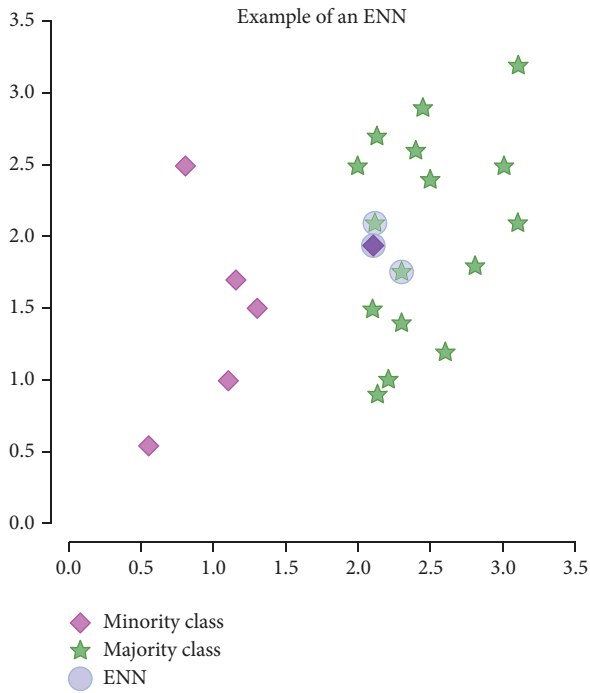


FIGURE 3: Example of an ENN.

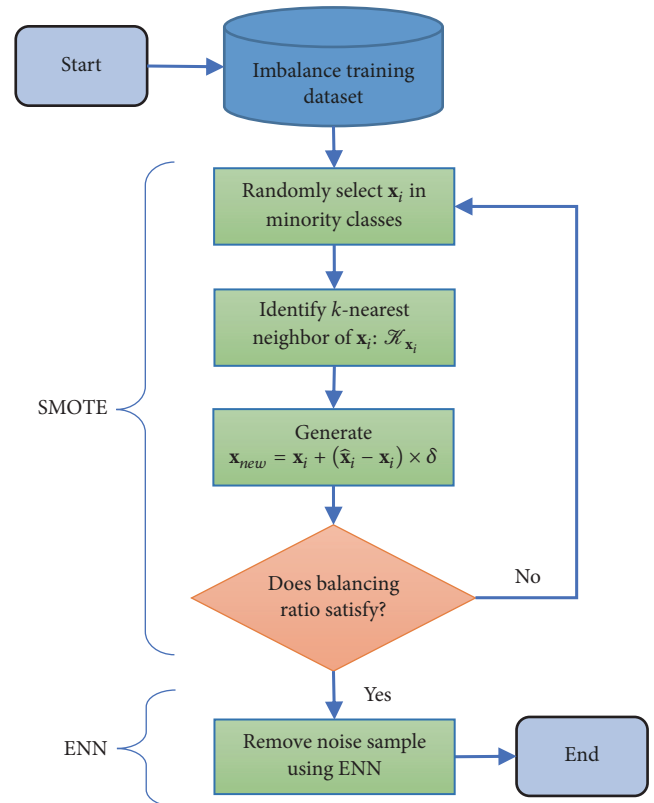


FIGURE 4: The flowchart of SMOTE-ENN algorithm.

based on the experimental results in [42]. Then, for each sample belonging to the majority class the algorithm will determine the distance from this sample to the nearest center point. Let d_{max} be the maximum value of the distances of data samples in class of bankruptcy firms. CBoost algorithm then assigns the values of each data sample in the minority

class equal to d_{max} . Then, CBoost algorithm will determine the initial weights denoted by W_1 as follows:

$$W_1(i) = \ln\left(\frac{1}{d(x_i)}\right) \quad (2)$$

where $d(\mathbf{x}_i)$ refers to the distance between data point \mathbf{x}_i and the nearest center point for the majority class and $d(\mathbf{x}_i) = d_{max}$ for the minority class. Equation (2) makes it so that the samples in the majority class closed the center points and the samples in the minority class will have higher weight values compared to the further samples in majority class. CBoost will then normalize these values by the following equation:

$$W_1(i) = \frac{W_1(i)}{\sum_{i=1}^m W_1(i)} \quad (3)$$

where m is the total number of data points in the training set. This step will ensure that

$$\sum W_1(i) = 1 \quad (4)$$

The initial weight W_1 helps the weak classifier classify more accurately the samples in the majority class close to the center points as well as the samples in the minority class. Therefore, it will improve the overall performance for class imbalance problem like bankruptcy dataset.

For each iteration, CBoost identifies the weak learner denoted by $h_t(\mathbf{x})$ that produces the lowest classification error denoted by ϵ_t , calculates the weight for this classifier denoted by α_t , and determines the next weight W_{t+1} for the next iteration as follows.

$$h_t = \underset{h_j \in \mathcal{H}}{\operatorname{argmin}} \epsilon_j = \sum_{i=1}^m W_t(i) [y_i \neq h_j(\mathbf{x}_i)]$$

$$\alpha_t = \eta \log \left(\frac{1 - \epsilon_t}{\epsilon_t} \right) \quad (5)$$

$$W_{t+1}(i) = \frac{W_t(i) \exp(-\alpha_t y_i h_t(\mathbf{x}_i))}{Z_t}$$

where Z_t is normalization factor. Finally, the algorithm will combine all weak learners to make the final classifier H as follows.

$$H(\mathbf{x}) = \operatorname{sign} \left(\sum_{t=1}^T \alpha_t h_t(\mathbf{x}) \right) \quad (6)$$

where $h_t(\mathbf{x})$ is the weak learner at the iteration t -th and α_t is its weight.

In short, CBoost is a greedy algorithm that finds one weak learner at an iteration, optimizes the weight of this learner, and updates the weighted distribution for the next iteration. The algorithm combines all weak learners as in (5) to create the final classifier. The flowchart of CBoost algorithm is shown in Figure 5.

2.4. The Hybrid Approach for Bankruptcy Prediction on KRBDS. The balancing ratio of KRBDS is very small which leads to a reduction in performance of oversampling and cost-sensitive learning independently. Therefore, this study proposes a hybrid approach that combines oversampling technique and cost-sensitive learning (HAOC) for bankruptcy prediction on KRBDS to improve the overall performance.

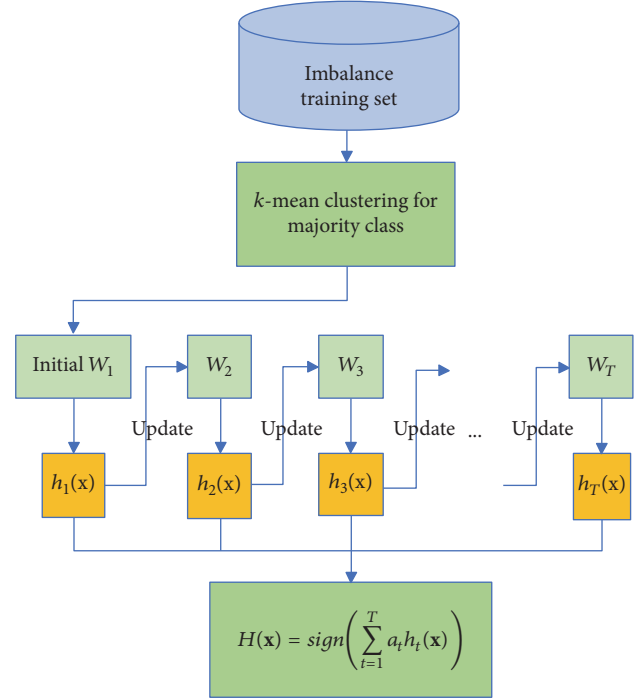


FIGURE 5: The flowchart of CBoost algorithm.

The flowchart of HAOC is presented in Figure 6. KRBDS is first normalized by using a normalization module that uses the best normalization technique in the first experiment (Data preprocessing). Next, the fivefold cross-validation module will be used to split the KRBDS into five parts, in which four parts were used for training and the remaining part was used for testing alternately.

The training set will be put into the found optimal balancing ratio module. This module will divide the training set into two subsets: the training set and validation set. Using these sets, this module tries various balancing ratios for SMOTE-ENN and will find the optimal balancing ratio for the KRBDS which will be presented in the first experiment. The training set will be balanced by SMOTE-ENN with the best balancing ratio that was found in the previous step. After this phase, the resample training set will be utilized to train the CBoost algorithm for bankruptcy prediction later. The testing set will be used to evaluate the proposed approach.

3. Experimental Results

3.1. Experiment Setup. The experimental methods were implemented in Python 2.7 environment and performed on a computer with Intel Core i7-2600 CPU (3.40 GHz \times 2 cores), 8 GB RAM that runs with Ubuntu 16.04 LTS. In addition, SMOTEENN was implemented by the imbalanced-learn package [47] and Bagging, AdaBoost, Random Forest, and MLP were in Scikit-learn package [48]. The imbalanced-learn package is an open-source Python toolbox which consists of several methods for dealing with the problem of class imbalance while Scikit-learn package is a free software machine learning library for the Python programming language.

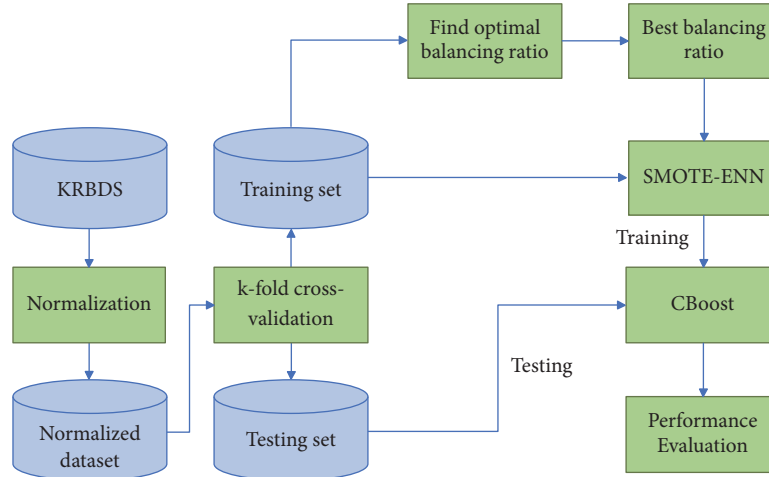


FIGURE 6: The flowchart of HAOC.

To show the effectiveness of the proposed approach, we compare the performance among the state-of-the-art methods and HAOC for bankruptcy prediction on KRBDS. The first four approaches are Bagging (BG), AdaBoost (AB), Random Forest (RF), and Multilayer Perceptron (MLP) which were recommended by Barboza et al. [32]. These approaches were used to predict bankruptcy directly; i.e., there is no resample approach applied to adjust the class distribution. The 5th to 8th approaches combine undersampling method based on clustering technique [43] with BG, AB, RF, and MLP classifiers. The 9th-12th approaches are oversampling method using SMOTE-ENN (with balancing ratio = 1) combined with BG, AB, RF, and MLP classifiers to predict bankruptcy. The 13th approach is RFCI introduced by Le et al. [42] and the 14th approach is the proposed approach (HAOC). Moreover, the study employs the fivefold cross-validation in 10 times with different configurations of folds for each run to get the average performance.

Next, we use GridSearchCV in Scikit-learn package [48] to tune several parameters of Bagging, AdaBoost, Random Forest, and MLP. We tuned the *n_estimators* (150) and *max_samples* (0.2) for Bagging, *learning_rate* (0.1) for AdaBoost, *max_depth* (5) for Random Forest, and *max_iter* (150), *learning_rate_init* (0.01), and *hidden_layer_sizes* (50, 5) for MLP.

3.2. Evaluation Metrics. This study uses two evaluation metrics including AUC (Area under the ROC Curve) and G-mean (Geometric Mean) to compare the performance among the experimental methods. A ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. This curve plots True Positive Rate (TPR) and False Positive Rate (FPR) computed as follows.

$$\begin{aligned} TPR &= \frac{TP}{TP + FN} \\ FPR &= \frac{FP}{FP + TN} \end{aligned} \quad (7)$$

where *TP*, *FN*, *FP*, and *TN* are true positives, false negatives, false positives, and true negatives, respectively. Lowering the classification threshold classifies more items as positive, thus increasing both false positives and true positives. AUC (Area under the ROC Curve) provides an aggregate measure of performance across all possible classification thresholds. If an algorithm has a larger AUC than that of another algorithm, this algorithm is better.

From ROC, Youden index which is the vertical distance between the 45-degree line and the point on the ROC curve was used to determine the optimal cut-off threshold. The Youden index is determined as follows.

$$J = sensitivity + specificity - 1 \quad (8)$$

The optimal cut-off threshold corresponds to the point with the maximum value of *J*. From that threshold, sensitivity and specificity, respectively, will be determined. G-mean is the root of the product of classwise sensitivity. This measure tries to maximize the accuracy on each of the classes while keeping these accuracies balanced. For binary classification G-mean is the squared root of the product of the sensitivity and specificity. Similar to the AUC, the algorithm with a larger G-mean is better.

3.3. Data Preprocessing. In this section, we apply some normalization techniques including StandardScaler, MinMaxScaler, and RobustScaler to the original features. StandardScaler normalizes the original features to create standardized features by removing the mean and scaling to unit variance. MinMaxScaler transforms the features by scaling each feature to a given range while RobustScaler scales the features using statistics that are robust to outliers. HAOC is then used to predict the bankruptcy from the normalized features. The performance results in Table 2 show that the StandardScaler is the best normalization technique for KRBDS. Therefore, we apply the StandardScaler for the next experiments. Please note that the settings of StandardScaler were found only using training data and then we used these settings for the training and testing data.

TABLE 2: Performance results of HAOC using several normalization techniques for KRBDS.

No	Normalization technique	Normalization formula	AUC
1	None	None	50.0±0.0
2	StandardScaler	$\mathbf{x}' = \frac{\mathbf{x} - \mathbf{x}_{mean}}{\mathbf{x}_{stdev}}$	87.1±0.6
3	MinMaxScaler	$\mathbf{x}' = \frac{\mathbf{x} - \mathbf{x}_{min}}{\mathbf{x}_{max} - \mathbf{x}_{min}}$	73.0±4.0
4	RobustScaler	$\mathbf{x}' = \frac{\mathbf{x} - \mathbf{x}_{Q_1}}{\mathbf{x}_{Q_3} - \mathbf{x}_{Q_1}}$	50.0±0.0

TABLE 3: The overall results of all experimental approaches for KRBDS.

No	Method	Resample approach	Classifier	AUC	G-mean	Average Rank	p -value
1	BG	None	Bagging	78.8±0.4	70.8±0.8	9.0	3.9×10^{-5}
2	AB	None	AdaBoost	84.9±0.8	78.2±0.6	7.0	0.0023
3	RF	None	Random Forest	86.2±0.6	79.9±0.6	4.7	0.069
4	MLP	None	MLP	86.7±0.8	80.1±1.0	2.6	0.487
5	USC-BG	Under-sampling method based on clustering technique (USC) [43]	Bagging	65.1±1.6	53.6±4.9	11.2	1.2×10^{-7}
6	USC-AB		AdaBoost	59.7±3.0	56.3±5.0	12.9	5.6×10^{-10}
7	USC-RF		Random Forest	64.7±1.0	62.6±1.9	11.9	1.5×10^{-8}
8	USC-MLP		MLP	46.9±2.7	36.5±3.7	14.0	1.1×10^{-11}
9	OSE-BG	Oversampling method using SMOTE-ENN (OSE) [41]	Bagging	83.9±0.3	77.4±0.3	7.8	5.1×10^{-4}
10	OSE-AB		AdaBoost	85.4±0.7	78.5±0.4	6.2	0.009
11	OSE-RF		Random Forest	86.6±0.7	80.2±1.0	3.3	0.285
12	OSE-MLP		MLP	72.8±2.1	69.8±1.8	10.0	3.3×10^{-6}
13	RFCI [42]	Under-sampling method using IHT concept	CBoost	86.6±0.7	79.1±3.5	3.1	0.336
14	HAOC	Oversampling method using SMOTE-ENN (with balancing ratio = 0.08)	CBoost	87.1±0.6	81.1±0.8	1.3	-

3.4. *Finding the Optimal Balancing Ratio.* This section is conducted to find the optimal balancing ratio of HAOC for KRBDS. Using different balancing ratios from 0.003 to 1 for oversampling module, we obtain the AUCs for the valuation sets shown as Figure 7 in five folds. According to the results, we found that the balancing ratio at 0.08 gives the best average AUC for validation sets. Therefore, we use this value for our proposed approach in the final experiment.

3.5. *Performance Results.* Figure 8 shows the box plot in terms of AUC of the experimental approaches for KRBDS in five folds. We can easily see found that CUS_BG, CUS_AB, CUS_RF, CUS_MLP, and OSE_MLP did not achieve good results. The remaining approaches get more positive results.

Figure 9 presents the box plot in term of G-mean of all the experimental approaches which indicate that AB, RF, NLP, OSE_RF, RFCI, and HAOC are the best methods in terms of G-mean.

Table 3 presents the average AUCs and G-mean of these approaches with standard deviation. According to these

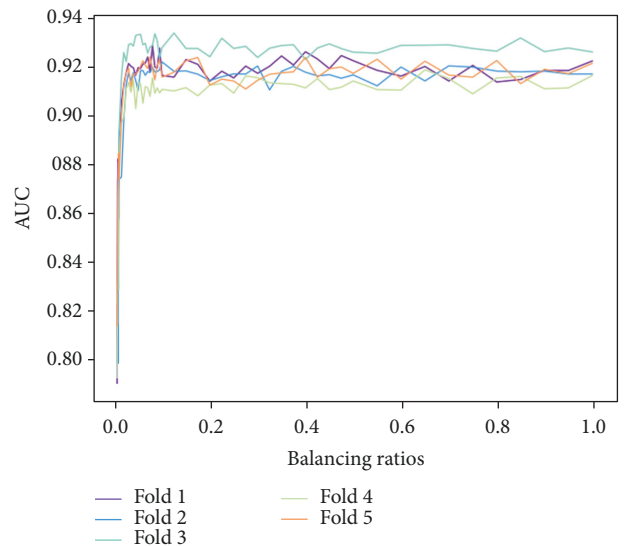


FIGURE 7: Performance of HAOC in terms of AUC for validation sets in five folds.

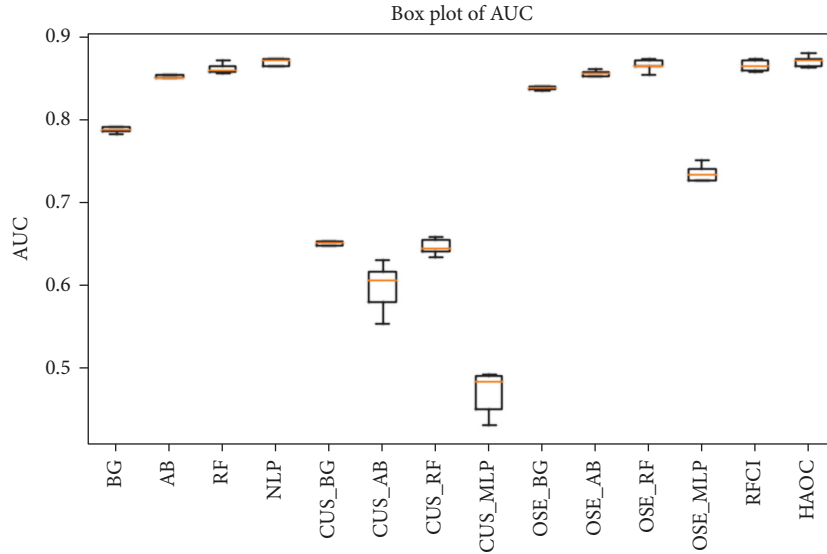


FIGURE 8: The box plot in terms of AUC of experimental approaches for KRBDS in five folds.

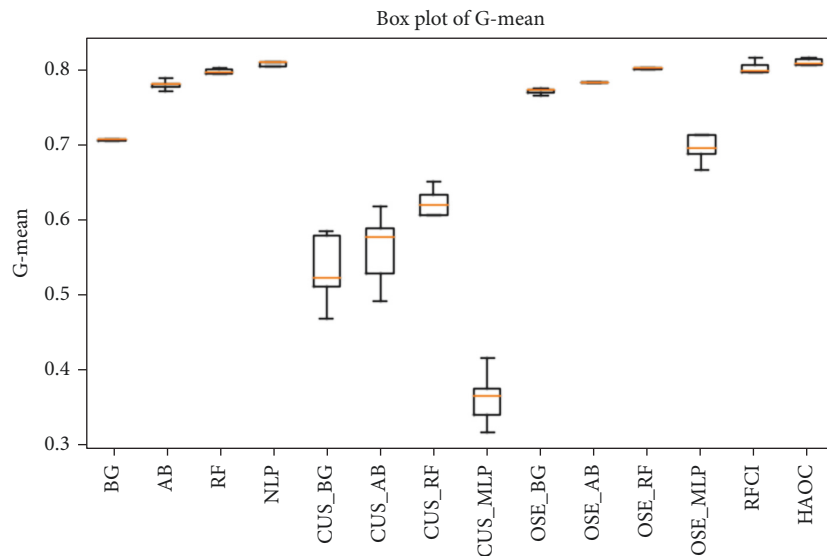


FIGURE 9: The box plot in terms of G-mean of experimental approaches for KRBDS in five folds.

results, Bagging without resample approach gives poor results at 78.8 in AUC, respectively. Meanwhile, AdaBoost, Random Forest, and MLP show acceptable results at 84.9, 86.2, and 86.7 in AUC. In addition, the undersampling method based on clustering technique (UCS) [43] is responsible for reducing the performance of classification algorithms including Bagging, MLP, RF, and AB. Therefore, UCS is not suitable for KRBDS when its balancing ratio is very small. The 9th–12th approaches, OSE-BG, OSE-AB, OSE-RF, and OSE-MLP, give the overall AUC at 83.9, 85.4, 86.6, and 72.8, respectively. Meanwhile, RFCI [42] that uses the cost-sensitive learning algorithm, namely, CBoost, achieved 86.6 in AUC. Our proposed method outperforms the other approaches when achieving the overall AUC at 87.1. Moreover, Table 3 also reports the G-mean of all experimental approaches.

According to these results, HAOC achieves the best value of G-mean while OSE-RF obtains the second value. Besides, RFCI, MLP, RF, and OSE-RF also have good results. In general, the proposed approach has the best values which balance between AUC and G-mean for KRBDS.

In addition, we employ the MULTIPLETEST package [49] for conducting multiple comparisons involving all possible pairwise experimental methods whose results are also presented in Table 3. The average rank of the proposed method is 1.3 which is the best rank in terms of AUC. Also, it can be noted that the results of our proposal do not have statistical differences against those results obtained by Random Forest, MLP, OSE-RF, and RFCI when the p -values are greater than 0.05. In addition, the p -values (≤ 0.05) show that the differences in the results of HAOC

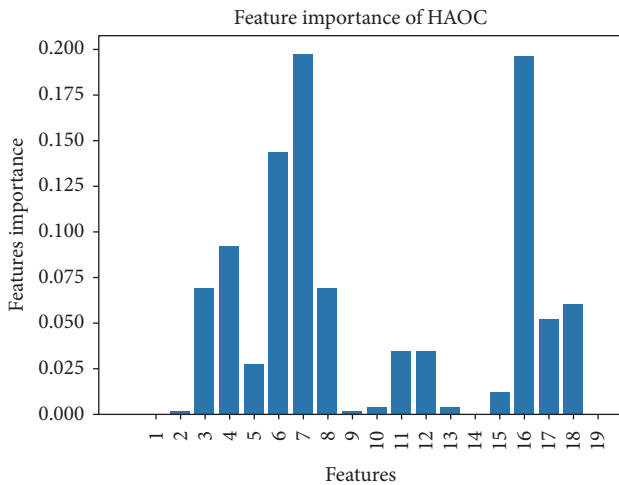


FIGURE 10: Feature importance of HAOC approach.

against the remaining tested classifiers are statistically significant.

Finally, Figure 10 presents the feature importance of HAOC approach on KRBDS. We can easily see that F3 (total assets), F4 (current liabilities within one year), F6 (total liabilities), F7 (capital), F8 (earned surplus), and F16 (nonoperating income) are the most important features. On the contrary, F1 (current assets), F2 (fixed assets, or fixed capital property), F9 (total capital), F10 (total capital after liabilities), F13 (net profit), F14 (sales and administrative expenses), and F19 (net income) are unimportant features and therefore they can be removed in the proposed model.

4. Conclusions

This study proposed a hybrid approach using oversampling technique and cost-sensitive learning framework for bankruptcy prediction on the Korean Bankruptcy dataset. In the first phase, the training set will be balanced by an oversampling module that utilizes the SMOTE-ENN algorithm with an optimal balancing ratio. Then, the second module uses the cost-sensitive learning framework, namely, CBoost, for bankruptcy prediction. Two experiments were conducted in this study to show the effectiveness of the proposed approach. The first experiment is to find the optimal balancing ratio that will give the best overall performance for bankruptcy prediction on the training set. Using the optimal balancing ratio that was found in the first experiment, we evaluate the performance in terms of AUC and G-mean between our proposed approach and the existing approaches. The results indicate that HAOC outperforms the existing approaches for bankruptcy prediction on KRBDS.

In the future, we will focus on how to find the optimal feature selection methods using evolutionary algorithms. In addition, several advanced methods for forecasting bankruptcy from multiple information sources to improve performance will be studied.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgments

This research was supported by the MSIP (Ministry of Science, ICT & Future Planning), Korea, under the National Program for Excellence in SW (2015-0-00938) supervised by the IITP (Institute for Information & communications Technology Planning & Evaluation).

References

- [1] T. H. Cupertino, M. Guimarães Carneiro, Q. Zheng, J. Zhang, and L. Zhao, "A scheme for high level data classification using random walk and network measures," *Expert Systems with Applications*, vol. 92, pp. 289–303, 2018.
- [2] T. C. Silva and L. Zhao, *Machine Learning in Complex Networks*, Springer, 2016.
- [3] T. Le, B. Vo, P. Fournier-Viger, M. Y. Lee, and S. W. Baik, "SPPC: a new tree structure for mining erasable patterns in data streams," *Applied Intelligence*, vol. 49, no. 2, pp. 478–495, 2019.
- [4] T. Le, B. Vo, and S. W. Baik, "Efficient algorithms for mining top-rank- k erasable patterns using pruning strategies and the subsume concept," *Engineering Applications of Artificial Intelligence*, vol. 68, pp. 1–9, 2018.
- [5] T. Le, A. Nguyen, B. Huynh, B. Vo, and W. Pedrycz, "Mining constrained inter-sequence patterns: a novel approach to cope with item constraints," *Applied Intelligence*, vol. 48, no. 5, pp. 1327–1343, 2018.
- [6] T. Kieu, B. Vo, T. Le, Z. Deng, and B. Le, "Mining top-k co-occurrence items with sequential pattern," *Expert Systems with Applications*, vol. 85, pp. 123–133, 2017.
- [7] B. Vo, T. Le, F. Coenen, and T.-P. Hong, "Mining frequent itemsets using the n-list and subsume concepts," *International Journal of Machine Learning and Cybernetics*, vol. 7, no. 2, pp. 253–265, 2016.
- [8] B. Vo, T. Le, G. Nguyen, and T. Hong, "Efficient algorithms for mining erasable closed patterns from product datasets," *IEEE Access*, vol. 5, pp. 3111–3120, 2017.
- [9] G. Nguyen, T. Le, B. Vo, and B. Le, "EIFDD: An efficient approach for erasable itemset mining of very dense datasets," *Applied Intelligence*, vol. 43, no. 1, pp. 85–94, 2015.
- [10] B. L. R. Stojkoska and K. V. Trivodaliev, "A review of Internet of things for smart home: challenges and solutions," *Journal of Cleaner Production*, vol. 140, pp. 1454–1464, 2017.
- [11] M. Wollschlaeger, T. Sauter, and J. Jasperneite, "The future of industrial communication: Automation networks in the era of the internet of things and industry 4.0," *IEEE Industrial Electronics Magazine*, vol. 11, no. 1, pp. 17–27, 2017.
- [12] N. P. Nguyen and S. K. Hong, "Sliding mode Thau observer for actuator fault diagnosis of quadcopter UAVs," *Applied Sciences*, vol. 8, no. 10, article 1893, 2018.

- [13] N. P. Nguyen and S. K. Hong, "Fault-Tolerant control of quadcopter uavs using robust adaptive sliding mode approach," *Energies*, vol. 12, no. 1, article 95, 2019.
- [14] N. Nguyen and S. Hong, "Fault diagnosis and fault-tolerant control scheme for quadcopter UAVs with a total loss of actuator," *Energies*, vol. 12, no. 6, article 1139, 2019.
- [15] T. N. Nguyen, S. Lee, H. Nguyen-Xuan, and J. Lee, "A novel analysis-prediction approach for geometrically nonlinear problems using group method of data handling," *Computer Methods Applied Mechanics and Engineering*, vol. 354, pp. 506–526, 2019.
- [16] T. N. Nguyen, C. H. Thai, A. Luu, H. Nguyen-Xuan, and J. Lee, "NURBS-based postbuckling analysis of functionally graded carbon nanotube-reinforced composite shells," *Computer Methods Applied Mechanics and Engineering*, vol. 347, pp. 983–1003, 2019.
- [17] T. N. Nguyen, C. H. Thai, H. Nguyen-Xuan, and J. Lee, "NURBS-based analyses of functionally graded carbon nanotube-reinforced composite shells," *Composite Structures*, vol. 203, pp. 349–360, 2018.
- [18] T. N. Nguyen, C. H. Thai, H. Nguyen-Xuan, and J. Lee, "Geometrically nonlinear analysis of functionally graded material plates using an improved moving Kriging meshfree method based on a refined plate theory," *Composite Structures*, vol. 193, pp. 268–280, 2018.
- [19] D. Le and V. Pham, "HGPEC: a Cytoscape app for prediction of novel disease-gene and disease-disease associations and evidence collection based on a random walk on heterogeneous network," *BMC Systems Biology*, vol. 11, no. 1, article 61, 2017.
- [20] D. J. Hemanth, J. Anitha, and L. H. Son, "Brain signal based human emotion analysis by circular back propagation and deep kohonen neural networks," *Computers and Electrical Engineering*, vol. 68, pp. 170–180, 2018.
- [21] D. M. Fazio, T. C. Silva, B. M. Tabak, and D. O. Cajueiro, "Inflation targeting and financial stability: Does the quality of institutions matter?" *Economic Modelling*, vol. 71, pp. 1–15, 2018.
- [22] T. Le, B. Vo, H. Fujita, N. Nguyen, and S. W. Baik, "A fast and accurate approach for bankruptcy forecasting using squared logistics loss with GPU-based extreme gradient boosting," *Information Sciences*, vol. 494, pp. 294–310, 2019.
- [23] A. Vanderveld, A. Pandey, A. Han, and R. Parekh, "An engagement-based customer lifetime value system for E-commerce," in *Proceedings of the 22nd ACM SIGKDD International Conference*, pp. 293–302, San Francisco, Calif, USA, August 2016.
- [24] B. Zhu, B. Baesens, and S. K. vanden Broucke, "An empirical comparison of techniques for the class imbalance problem in churn prediction," *Information Sciences*, vol. 408, pp. 84–99, 2017.
- [25] D. A. Chekired, L. Khoukhi, and H. T. Mouftah, "Decentralized cloud-SDN architecture in smart grid: a dynamic pricing model," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 3, pp. 1220–1231, 2018.
- [26] X. Chen, Y. Fang, M. Yang, F. Nie, Z. Zhao, and J. Z. Huang, "PurTreeClust: a clustering algorithm for customer segmentation from massive customer transaction data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 3, pp. 559–572, 2018.
- [27] H. V. Long, L. H. Son, M. Khari et al., "A new approach for construction of geodemographic segmentation model and prediction analysis," *Computational Intelligence and Neuroscience*, vol. 2019, Article ID 9252837, 10 pages, 2019.
- [28] T. C. Silva, M. D. S. Alexandre, and B. M. Tabak, "Bank lending and systemic risk: A financial-real sector network approach with feedback," *Journal of Financial Stability*, vol. 38, pp. 98–118, 2018.
- [29] B. M. Tabak, T. C. Silva, and A. Sensoy, "Financial Networks," *Complexity*, vol. 2018, Article ID 7802590, 2 pages, 2018.
- [30] M. Kim, D. Kang, and H. B. Kim, "Geometric mean based boosting algorithm with over-sampling to resolve data imbalance problem for bankruptcy prediction," *Expert Systems with Applications*, vol. 42, no. 3, pp. 1074–1082, 2015.
- [31] M. Zięba, S. K. Tomczak, and J. M. Tomczak, "Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction," *Expert Systems with Applications*, vol. 58, pp. 93–101, 2016.
- [32] F. Barboza, H. Kimura, and E. Altman, "Machine learning models and bankruptcy prediction," *Expert Systems with Applications*, vol. 83, pp. 405–417, 2017.
- [33] A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk, and F. Herrera, *Learning from Imbalanced Data Sets*, Springer, 2018.
- [34] Y. Lin, Y. Lee, and G. Wahba, "Support vector machines for classification in nonstandard situations," *Machine Learning*, vol. 46, no. 1-3, pp. 191–202, 2002.
- [35] B. Liu, Y. Ma, and C. Wong, "Improving an association rule-based classifier," *PKDD*, pp. 293–317, 2000.
- [36] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.
- [37] T. Le and S. W. Baik, "A robust framework for self-care problem identification for children with disability," *Symmetry*, vol. 11, no. 1, article 89, 2019.
- [38] N. V. Chawla, D. A. Cieslak, L. O. Hall, and A. Joshi, "Automatically countering imbalance and its empirical relationship to cost," *Data Mining and Knowledge Discovery*, vol. 17, no. 2, pp. 225–252, 2008.
- [39] C. Ling, V. Sheng, and Q. Yang, "Test strategies for cost-sensitive decision trees," *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, no. 8, pp. 1055–1067, 2006.
- [40] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, "A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 42, no. 4, pp. 463–484, 2012.
- [41] T. Le, M. Y. Lee, J. R. Park, and S. W. Baik, "Oversampling techniques for bankruptcy prediction: Novel features from a transaction dataset," *Symmetry*, vol. 10, no. 4, article 79, 2018.
- [42] T. Le, L. H. Son, M. T. Vo, M. Y. Lee, and S. W. Baik, "A cluster-based boosting algorithm for bankruptcy prediction in a highly imbalanced dataset," *Symmetry*, vol. 10, no. 7, article 250, 2018.
- [43] W. Lin, C. Tsai, Y. Hu, and J. Jhang, "Clustering-based under-sampling in class-imbalanced data," *Information Sciences*, vol. 409–410, pp. 17–26, 2017.
- [44] H. Han, W.-Y. Wang, and B.-H. Mao, "Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning," in *Proceedings of the International Conference on Intelligent Computing (ICIC '05)*, vol. 3644 of *Lecture Notes in Computer Science*, pp. 878–887, August 2005.
- [45] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: adaptive synthetic sampling approach for imbalanced learning," in *Proceedings of the International Joint Conference on Neural Networks (IJCNN '08)*, pp. 1322–1328, June 2008.

- [46] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A study of the behaviour of several methods for balancing machine learning training data," *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, pp. 20–29, 2004.
- [47] G. Lemaitre, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning," *Journal of Machine Learning Research*, vol. 18, no. 17, pp. 1–5, 2017.
- [48] F. Pedregosa, G. Varoquaux, A. Gramfort et al., "Scikit-learn: machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [49] S. García and F. Herrera, "An extension on statistical comparisons of classifiers over multiple data sets for all pairwise comparisons," *Journal of Machine Learning Research*, vol. 9, pp. 2677–2694, 2008.

Research Article

A Differential Evolution-Oriented Pruning Neural Network Model for Bankruptcy Prediction

Yajiao Tang,^{1,2} Junkai Ji ,³ Yulin Zhu,¹ Shangce Gao ,² Zheng Tang,² and Yuki Todo ⁴

¹College of Economics, Central South University of Forestry and Technology, Changsha 410004, China

²Faculty of Engineering, University of Toyama, Toyama 930-8555, Japan

³College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China

⁴School of Electrical and Computer Engineering, Kanazawa University, Kanazawa-shi 920-1192, Japan

Correspondence should be addressed to Shangce Gao; gaosc@eng.u-toyama.ac.jp and Yuki Todo; yktodo@ec.t.kanazawa-u.ac.jp

Received 7 June 2019; Accepted 14 July 2019; Published 4 August 2019

Guest Editor: Thiago C. Silva

Copyright © 2019 Yajiao Tang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Financial bankruptcy prediction is crucial for financial institutions in assessing the financial health of companies and individuals. Such work is necessary for financial institutions to establish effective prediction models to make appropriate lending decisions. In recent decades, various bankruptcy prediction models have been developed for academics and practitioners to predict the likelihood that a loan customer will go bankrupt. Among them, Artificial Neural Networks (ANNs) have been widely and effectively applied in bankruptcy prediction. Inspired by the mechanism of biological neurons, we propose an evolutionary pruning neural network (EPNN) model to conduct financial bankruptcy analysis. The EPNN possesses a dynamic dendritic structure that is trained by a global optimization learning algorithm: the Adaptive Differential Evolution algorithm with Optional External Archive (JADE). The EPNN can reduce the computational complexity by removing the superfluous and ineffective synapses and dendrites in the structure and is simultaneously able to achieve a competitive classification accuracy. After simplifying the structure, the EPNN can be entirely replaced by a logic circuit containing the comparators and the logic NOT, AND, and OR gates. This mechanism makes it feasible to apply the EPNN to bankruptcy analysis in hardware implementations. To verify the effectiveness of the EPNN, we adopt two benchmark datasets in our experiments. The experimental results reveal that the EPNN outperforms the Multilayer Perceptron (MLP) model and our previously developed preliminary pruning neural network (PNN) model in terms of accuracy, convergence speed, and Area Under the Receiver Operating Characteristics (ROC) curve (AUC). In addition, the EPNN also provides competitive and satisfactory classification performances in contrast with other commonly used classification methods.

1. Introduction

The overwhelming 2007/2008 financial crisis led to the bankruptcy of many large-scale financial institutions and made some subject to takeover by their government. Thus, bankruptcy risk management has become an important field of study worldwide. Bankruptcy by a company denotes a situation in which the operating cash flow of the company and its negative net assets cannot be balanced. This always results in the practical weakening of the profitability of a company. The purpose of bankruptcy prediction is to evaluate the present and future financial status of a company from the perspective of its long-term operation in the market.

Various quantitative statistical approaches have been adopted to improve bankruptcy forecasting models. Discriminant analysis is adopted to classify observations between good and bad payers [1], and logistic regression is adapted to determine the default probability of the borrowers [2]. However, it is argued that these popular models are inaccurate [3]. Hence, several machine learning tools are explored to assess bankruptcy risk using computer technology. Because bankruptcy risk analysis is similar to pattern recognition tasks, most methods can be adapted to classify the credit-worthiness of potential clients of financial institutions [4, 5]. Among them, ANNs achieve outstanding performances in applications such as predicting financial crises [6], scoring credit [7], and building up credit analysis models [8]. The

adoption of ANNs in bankruptcy prediction has been studied since the 1990s [9, 10]. Prior studies revealed that ANNs are powerful for use in pattern recognition because of their nonlinear and nonparametric adaptive-learning properties [11]. This imbues ANNs with obvious advantages over conventional statistical algorithms and inductive learning methods, especially in comparison with discriminant analysis and logistic regression [12]. Hence, researchers have put a major emphasis on the application of ANNs in finance and accounting.

ANNs are flexible and nonparametric modelling tools capable of performing any complicated function mapping with arbitrarily required accuracies [13–15]. Among the diverse types of ANNs, MLP is one of the simplest and most widely applied models, in which the hidden layer determines the mapping relationships between input and output layers and the relationships between neurons stored as the weights of the connecting links [16]. The MLP's learning algorithm implements a gradient search to minimize the squared error between the realized and desired outputs. This type of three-layer MLP is a commonly adopted ANN structure for binary classification problems such as bankruptcy prediction [11]. Although the characteristics of ANN ensembles, such as efficiency, robustness, and adaptability, make them a valuable tool for classification, decision support, financial analysis, and credit scoring, it should be noted that some researchers have shown that the ensembles of multiple neural network classifiers are not always superior to a single best neural network classifier [17]. Hence, we focus on applying a single neural network model to bankruptcy prediction.

In biological neuron models, a dendritic computation mechanism can provide a concrete explanation concerning the positioning of the synaptic inputs at the proper connections. This means that redundant synapses and dendrites are left in the neural network initially, while the useless ones are quickly deleted, with the remaining being strengthened. Ultimately, this process creates an enhanced neural network function form. Inspired by these histological theories, Koch et al. notes that interactions between excitatory and inhibitory inputs have apparent nonlinearity. Once inhibitory inputs and excitatory inputs are located on the same path to the soma, the inhibitory inputs can specifically eliminate the excitatory inputs. However, issues, such as whether the excitatory or inhibitory synapse should be kept, where it should locate, and which dendritic branch should be strengthened, are unaddressed in this model [18]. Later, Koch et al. noted that the interactions among synapses and the responses at the connection nodes could be regarded as logic operations [19], and a specialized learning algorithm based on the plasticity in dendrites was required to train the model [20].

In our previous research, a PNN model, in which the particular locations and types of synapses on the dendrite branches are formulated via learning, is proposed, and useless and superfluous synaptic and dendritic connections are eliminated. Thus, the efficiency of the model is enhanced [21, 22]. Similar to most other ANNs, PNN adopts the backpropagation (BP) algorithm as its learning method. However, learning algorithms are widely considered to have significant influences on the performances of ANNs [23, 24].

The BP algorithm and its variations [23, 25] are considered rather inefficient because of their obvious drawbacks such as their slow convergence [26], sensitivity to initialization [27], and a tendency to become trapped in local minima [28, 29]. Specifically, first, during the learning process, the error often remains large because the learning algorithm leads the ANNs to local minima instead of the global minimum. This problem is quite common in gradient-based learning approaches. Second, the convergence of the BP algorithm is strongly dependent on the initial values of the learning rate and momentum. The unsuitable values for these variables may even lead to divergence. Third, the learning time increases substantially when the dataset becomes larger [30]. Many researchers have focused on making improvements to resolve these shortcomings of BP, but each method has its disadvantages [31, 32]. These disadvantages make them unreliable for risk classification applications and inspire us to adopt other algorithms to train the neural model to avoid the computational inefficiency and local minimum problems.

In this study, we propose an EPNN model with a dendritic structure as a global optimization algorithm called the JADE algorithm [33]. With respect to the EPNN, the axons of the other neurons transmit the input signals to the synaptic layer; then, the interaction of the synaptic signals transfers to every branch of the dendrites. Next, the interactions are collected and sent to the membrane layer and then transformed to the soma body. In addition, the neuronal pruning function can remove extra synapses and dendrites and simultaneously achieve high accuracy. Specifically, during the training process, the superfluous inputs and dendrites are eliminated, while the useful and necessary ones are retained. Then, the neuronal pruning function can produce a simplified dendritic morphology without a loss of classification accuracy. Furthermore, the simplified topological morphology can operate similarly to a logic circuit composed merely of comparators and logic NOT, AND, and OR gates. Thus, applying EPNN to bankruptcy analysis can easily be implemented in hardware. To the best of our knowledge, we note that if achieved through hardware implementation, this technique will achieve the highest computation speed when compared with other methods. This demonstrates an excellent adoption possibility for financial institutions. JADE is a state-of-the-art variant of the differential evolution algorithm and uses a self-adaptive mechanism to select suitable parameters for each optimization problem. This imbues JADE with a better balance between exploration and exploitation compared to other heuristic algorithms [34]. In training the EPNN, JADE can avoid local minima and speed up the training process during the optimization process. Thus, JADE allows the EPNN to obtain satisfactory results and produce an effective logic circuit for each bankruptcy prediction problem.

In addition to avoiding misleading and contradictory conclusions, four key components are carefully defined to allow one to draw well-founded conclusions from the experimental results. First, the research has adopted both benchmark and application-oriented databases, namely, a Qualitative Bankruptcy dataset from the UCI Machine Learning Database Repository and a Distress dataset from the Kaggle dataset. Second, in the simulation, the two datasets are

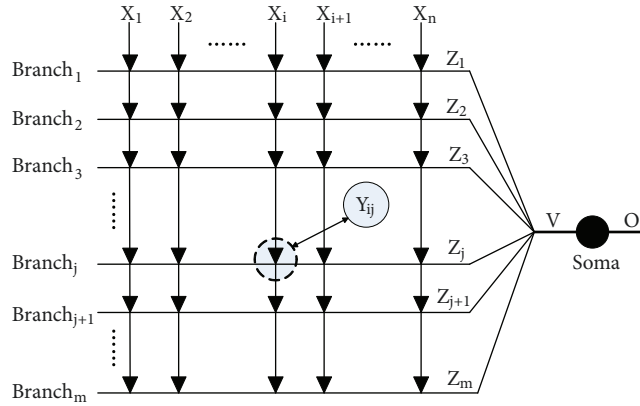


FIGURE 1: The morphological architecture of the PNN.

separated into a training set and a testing set at proportions of 50% each. Third, the average accuracy, sensitivity, specificity, convergence speed, and AUC are used as the evaluation metric framework; such metrics can be used to effectively and efficiently analyse the possibility of bankruptcy. Fourth, a nonparametric test called the Wilcoxon rank-sum test has been adopted to allow us to claim that the observed result differences in performance are statistically significant and not simply caused by random splitting effects.

To conclude, our main contributions are clarified as follows: first, a novel EPNN model is proposed in this paper which can adopt synaptic and dendritic pruning to simplify its neuron morphology during the training process. Second, the simplified model of EPNN can be completely replaced by logic circuits which be easily implemented on hardware. The logic circuits can maintain high classification accuracy and obtain extremely high computation speed, simultaneously. Last but not least, comprehensive comparison experiments have been implemented to demonstrate that the EPNN outperforms the MLP, PNN, and other commonly used classifiers on the bankruptcy prediction problems.

The remainder of this paper is constructed as follows. Section 2 presents an overview of the related theories in bankruptcy analysis. Section 3 introduces the proposed EPNN model in detail. Moreover, the EPNN's learning algorithm JADE is described. Section 4 presents the experimental results obtained using the EPNN and makes a comparison with other algorithms by adopting the Qualitative Bankruptcy dataset and Distress dataset. Section 5 concludes this paper.

2. Proposed Model

We build up the EPNN, which has a dendritic structure and which is trained by JADE, to achieve a high bankruptcy classification accuracy. The morphological architecture of the EPNN is shown in Figure 1. The network has four layers, namely, a synaptic layer, a dendritic layer, a membrane layer, and a soma layer. The inputs x_{1-n} from the axons of the prior neurons enter the synaptic layer; then, the interactions of the synaptic signals occur on each branch of dendrites. After that,

the interactions are collected and sent to the membrane layer; finally, they are sent to the soma body. During the training process, the necessary inputs and useful dendrites are held, whereas the unnecessary ones are filtered out. The cell would be motivated and would then send an output signal to other neurons through the axon terminal when the input of the soma exceeds its threshold. The morphological architecture of the EPNN model is presented below in detail.

2.1. Synaptic Layer. The synaptic layer of a neuron represents the specific area at which nerve impulses are transmitted among neurons, thereby passing through the axon terminal of a neuron where neurotransmitters are released in response to an impulse [35]. The impulse is implemented using a certain pattern of a specific ion. When an ion transmits to the receptor, the potential of the receptor is changed and determines the excitatory or inhibitory characteristic of a synapse [36]. The flow direction of the synaptic layer is feed-forward, which conventionally starts from a presynaptic neuron and transmits to a postsynaptic neuron. In the EPNN, these connections are formulated by a sigmoid function with a single input and a single output. The equation of the j^{th} ($j = 1, 2, 3, \dots, J$) synaptic layer receiving the i^{th} ($i = 1, 2, 3, \dots, I$) input is expressed as follows:

$$Y_{ij} = \frac{1}{1 + e^{-k(w_{ij}x_i - q_{ij})}}, \quad (1)$$

where k is a positive constant, w_{ij} and q_{ij} are synaptic parameters that need to be optimized by the learning algorithm, and x_i is the input of the synapse, with a range of $[0, 1]$. There are four types of connection states corresponding to different values of w_{ij} and q_{ij} : a direct connection, a reverse connection, a constant-1 connection, and a constant-0 connection, as shown in Figure 2. θ_{ij} represents the threshold of a synaptic layer; this threshold can be defined by the following equation,

$$\theta_{ij} = \frac{q_{ij}}{w_{ij}}. \quad (2)$$

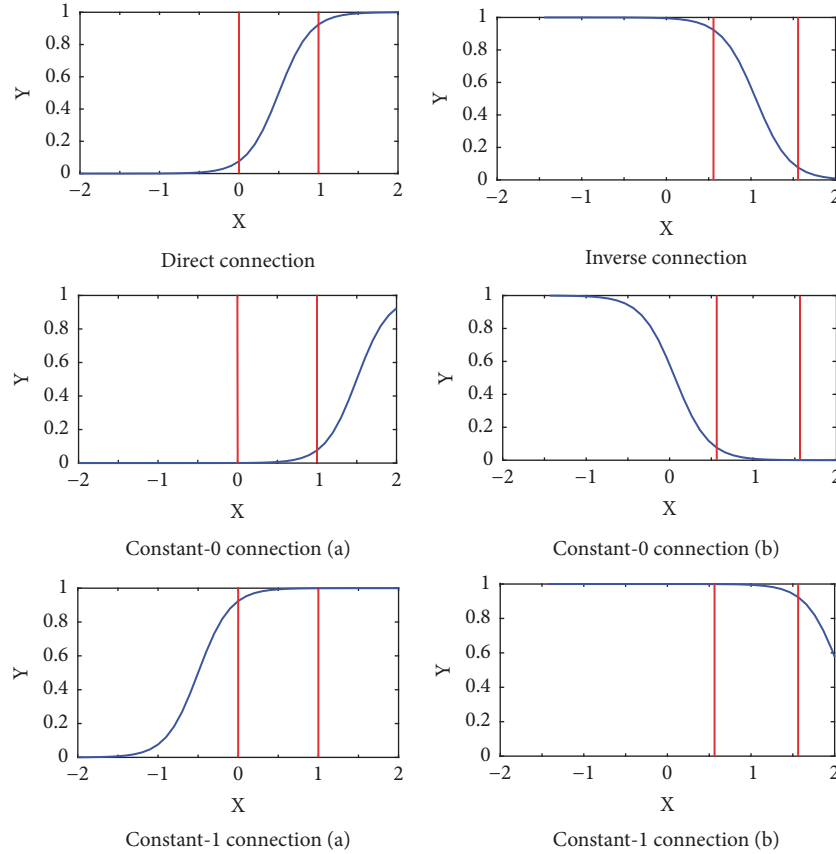


FIGURE 2: Six connection cases of the synaptic layer.

2.1.1. Direct Connection. $0 < q_{ij} < w_{ij}$, e.g., $q_{ij} = 0.5$ and $w_{ij} = 1.0$, corresponds to a direct connection. Once $x_i > \theta_{ij}$, the output Y_{ij} approximates to 1, the synapse becomes excitatory, and it depolarizes the soma layer. When $x_i \leq \theta_{ij}$, the corresponding output tends to be 0, the synapse becomes inhibitory, and it hyperpolarizes the soma layer in a transient manner. In general, regardless of the input values, the outputs always approximate the inputs.

2.1.2. Inverse Connection. $w_{ij} < q_{ij} < 0$, e.g., $q_{ij} = -0.5$ and $w_{ij} = -1.0$, leads to an inverse connection. Once $x_i > \theta_{ij}$, the output Y_{ij} is approximately 0, and the synapse becomes inhibitory. In addition, it will hyperpolarize the soma layer in a transient manner. In contrast, when $x_i \leq \theta_{ij}$, the output Y_{ij} is approximately 1, the synapse will become excitatory, and it depolarizes the soma layer. Briefly, regardless of the values of the inputs in $[0, 1]$, the output will receive an inverse signal triggered by the input. This can be regarded as a logic NOT operation.

2.1.3. Constant-0 Connection. There are two states in the constant-0 connection: $w_{ij} < 0 < q_{ij}$, e.g., $q_{ij} = 0.5$ and $w_{ij} = -1.0$, and $0 < w_{ij} < q_{ij}$, e.g., $q_{ij} = 1.5$ and $w_{ij} = 1.0$. Regardless of the value of the input, the outputs are always approximately 0.

2.1.4. Constant-1 Connection. There are two states in the constant-1 connection: $q_{ij} < 0 < w_{ij}$, e.g., $q_{ij} = -0.5$ and $w_{ij} = 1.0$, and $q_{ij} < w_{ij} < 0$, e.g., $q_{ij} = -1.5$ and $w_{ij} = -1.0$. The corresponding output tends to be 1 all the time regardless of whether the input signal x_i exceeds the threshold θ_{ij} . This means that the signals of the synaptic layer have minimal impact on the dendritic layer. Whenever the excitatory input signals transport in, depolarization occurs in the next membrane layer.

The values of w_{ij} and q_{ij} are initialized randomly between -1.5 and 1.5. This represents that the inputs connect to each dendritic branch in one of the four synaptic connection statuses randomly. When the values of w_{ij} and q_{ij} change, the connection states of the synaptic layer vary. In Figure 3, four marks are adopted to represent the four connection states: a direct connection (\bullet), an inverse connection (\blacksquare), a constant-1 connection (\odot), and a constant-0 connection (\ominus).

2.2. Dendrite Layer. A dendrite layer denotes a typical nonlinear interaction of the synaptic signals on each branch of dendrites. The multiplication operation plays a vital role in the process of transporting and disposing the neural information [37, 38]. Thus, the nonlinearity calculation of the synaptic layer can be implemented by a typical multiplication operation instead of summation. The interaction of a dendritic branch is equivalent to a logic AND operation.

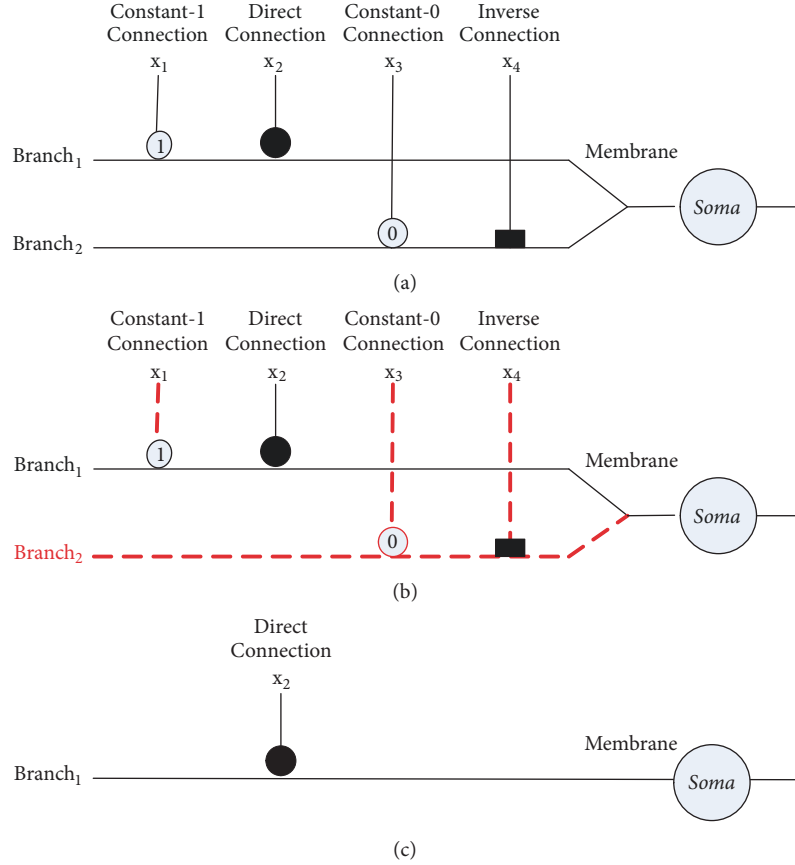


FIGURE 3: An example of a synaptic and dendritic pruning procedure.

The operation of the input variables will generate a 1 when and only when all input variables equal 1 simultaneously. The corresponding equation of the dendrite layer is defined as follows:

$$Z_j = \prod_{i=1}^I Y_{ij}. \quad (3)$$

2.3. Membrane Layer. A membrane layer accumulates the linear summation of the dendritic signals from the upper dendrite layer. It is similar to the logic OR operation in the binary cases. This logic OR operation generates a 1 when at least one of the variables is 1. Its equation is given below:

$$V = \sum_{j=1}^J Z_j. \quad (4)$$

2.4. Soma Layer. The output of the membrane layer is transmitted to the soma layer. Once the membrane potential exceeds the threshold, the neuron fires. A sigmoid operation is used to describe the function of the soma layer:

$$O = \frac{1}{1 + e^{-k_{soma}(V - \theta_{soma})}}. \quad (5)$$

2.5. Neural Pruning Function. The EPNN possesses the ability to perform a neural pruning function to simplify its topological morphology. The neural pruning technique represents omitting the extra nodes and weights by learning and training the neural network [39]. In the EPNN, the pruning function can eliminate unnecessary synapses and dendrites and then form a unique neural structure for a given problem. The function contains two parts: synaptic pruning and dendritic pruning.

Synaptic pruning: when the synaptic layer that receives the input from the axon is in the constant-1 connection case, the synaptic output is always 1. Because of the multiplication operation, the result of any arbitrary value multiplying 1 will equal itself in the dendrite layer. It is evident that the synaptic input in the constant-1 connection has minimal impact on the output of the dendrite layer. Therefore, this type of synaptic input can be neglected entirely.

Dendritic pruning: when the synaptic layer that receives the input signal is in the constant-0 connection case, the output is always 0. Consequently, the output of the adjacent dendrite layer becomes 0 because the result of any value multiplying 0 equals 0. This implies that this entire dendrite layer should be omitted because it has minimal influence on the result of the soma layer.

An example of a synaptic and dendritic pruning procedure is presented in Figure 3. The neural structure is

composed of four synaptic inputs, two dendrite branches, one membrane layer, and one soma layer as shown in Figure 3(a). The connection case of input x_1 is ① in Branch 1; this synaptic layer can be omitted according to the mechanism of synaptic pruning. In addition, the connection case of input x_3 is ② in Branch 2; Branch 2 can be completely deleted based on the dendritic pruning mechanism. The unnecessary synaptic inputs and dendritic branches, which are shown with dotted lines in Figure 3(b), should be removed. Finally, the simplified dendritic morphology can be obtained, as in Figure 3(c). Only the synaptic layer on Branch 1 remains in the structure because only the input x_2 can affect the final output of the soma.

3. Learning Algorithm

Actually, PNN suffers from the curse of dimensionality. When the dimension increases largely, any small change of the weights on one dendritic branch will produce a great disparity of its final results because of the multiplication operation. This is the main limitation of EPNN. Thus it needs us to propose more powerful optimization algorithms to figure it out. Conventional classifiers use BP to adjust the weights and threshold. However, BP suffers an inherent local minimum trapping problem and has difficulties in achieving the globally best values of its weights and thresholds. This disadvantage of BP has largely limited the computational capabilities of our neural mode. To improve the performance of the EPNN, we adopt JADE to train the model.

JADE has been regarded as one of a few “important variants of Differential Evolution (DE)” in a major DE review published in 2011 [34]. The vast popularity of DE algorithms has led to an increasing interest in developing their variants [46–48]. It is well known that the performances of many metaheuristic methods are influenced by the choice of their control parameters [49, 50]. JADE can use a self-adaptive mechanism to select suitable parameters F and CR for different optimization problems and implement a “DE/current-to- p best” mutation strategy with an optional external archive. The experimental results verified that JADE obtains a better balance between exploration and exploitation during the evolutionary process and is superior to other optimization algorithms in terms of solution quality and convergence rate [33]. JADE follows the general procedure of an evolutionary algorithm. After initialization, DE executes a loop of evolutionary operations: mutation, crossover, and selection. In addition, JADE dynamically updates control parameters as the evolutionary search proceeds.

Initialization: each agent in the initial population $x_{i,0} = x_{1,i,0}, x_{2,i,0}, \dots, x_{D,i,0}$, $i = 1, 2, \dots, NP$ is generated randomly according to a uniform distribution.

$$x_j^{low} \leq x_{j,i,0} \leq x_j^{up}, \quad j = 1, 2, \dots, D, \quad (6)$$

where D is the dimensionality of the problem and NP is the population size.

Mutation: At each generation g , the mutation vector $v_{i,g}$ is created based on the current parent population.

$$v_{i,g} = x_{i,g} + F_i * (x_{best,g}^p - x_{i,g}) + F_i * (x_{r_1,g} - x_{r_2,g}), \quad (7)$$

where the indices r_0, r_1 , and r_2 are distinct integers uniformly chosen from the set $\{1, 2, \dots, NP\}$; $x_{best,g}^p$ is chosen randomly as one of the top 100 $p\%$ individuals in the current population, with the probability $p \in (0, 1]$; and F_i is the mutation factor of the individual x_i that will be regenerated at each generation by the adaptation mechanism.

Crossover: a binomial crossover operation is adopted to generate the final offspring vector $u_{i,g} = u_{1,i,g}, u_{2,i,g}, \dots, u_{D,i,g}$.

$$u_{j,i,g} = \begin{cases} v_{j,i,g}, & \text{if } \text{rand}(0, 1) \leq CR_i \text{ or } j = j_{rand}, \\ x_{j,i,g}, & \text{otherwise,} \end{cases} \quad (8)$$

where $\text{rand}(0, 1)$ is a uniform random number on the interval $[0, 1]$. $j_{rand} = \text{rand int}(1, D)$ is an integer randomly extracted from the set $[1, 2, \dots, D - 1, D]$, where each individual i has its own crossover probability CR_i . The crossover probability $CR_i \in [0, 1]$ approximately corresponds to the fraction of vector components inherited from the mutation vector.

Selection: the selection operation compares the parent vector $x_{i,g}$ with the trial vector $u_{i,g}$ according to their fitness values $f(\cdot)$, and it chooses the better vector for the next generation. For example, if given a minimization problem, the selected vector is generated by the following equation:

$$x_{i,g+1} = \begin{cases} u_{i,g}, & \text{if } f(u_{i,g}) < f(x_{i,g}), \\ x_{i,g}, & \text{otherwise.} \end{cases} \quad (9)$$

In addition, if the trial vector $u_{i,g}$ is better than the parent vector $x_{i,g}$, the control parameters F_i and CR_i of the individual are called a successful mutation factor and a successful crossover probability, respectively.

Parameter adaptation: Better controlling the parameter values can result in individuals that have greater possibility to survive, and hence, these values should be retained in the next generation. At each generation g , the crossover rate CR_i is formed independently according to a normal distribution of mean μ_{CR} and standard deviation 0.1 and then normalized to the range $[0, 1]$, which can be described as follows:

$$CR_i = \text{rand } n_i(\mu_{CR}, 0.1), \quad (10)$$

where S_{CR} is the set that records all successful crossover rates CR_i at generation g . The initial value of μ_{CR} is set as 0.5; then, it is updated by the following equation at the end of each generation:

$$\mu_{CR} = (1 - c) * \mu_{CR} + c * \text{mean}_A(S_{CR}), \quad (11)$$

where c is a positive constant in the interval $[0, 1]$ and $\text{mean}_A(S_{CR})$ represents the arithmetic mean of the agents in S_{CR} .

Similarly, the mutation factor F_i is also independently generated according to a Cauchy distribution with location

parameter μ_F and scale parameter 0.1, subsequently being normalized to $[0, 1]$. This can be expressed as follows:

$$F_i = \text{rand}_c(\mu_F, 0.1). \quad (12)$$

Furthermore, the set S_F contains all the successful mutation factors in generation g . The initial value of μ_F of the Cauchy distribution is set to 0.5, and then, they are updated at the end of each generation by the following equation:

$$\mu_F = (1 - c) * \mu_F + c * \frac{\sum_{F \in S_F} F^2}{\sum_{F \in S_F} F}. \quad (13)$$

4. Application to Bankruptcy Classification

4.1. Bankruptcy Dataset Description. To evaluate the performance of the EPNN, both benchmark and application-oriented databases are adopted in our experiments. Each option has its advantages and disadvantages. The benchmark database allows future experiments to make extensive comparisons among different prediction models, but it cannot represent current socioeconomic statuses. Thus, the experiments may be out of date and lead to meaningless conclusions. In contrast, the application-oriented database is capable of addressing real-world problems, but it is difficult to employ for further comparison. Therefore, it is generally better to employ a mixture of both benchmark and application-oriented databases [51]. This study adopts a Qualitative Bankruptcy dataset from the UCI Machine Learning Database Repository and a Distress dataset from the Kaggle dataset to draw a significant and meaningful conclusion. In this paper, it is assumed that the state of a company's financial situation is expressed through a qualitative variable, such as the binary variable, where "1" represents a financially sound company and "0" denotes a company falling into bankruptcy.

The Qualitative Bankruptcy dataset is from the UCI repository, which has been applied successfully for bankruptcy classification in several previous works in the literature. The dataset is composed of 250 instances based on 6 attributes, with each corresponding to qualitative parameters concerning bankruptcy, namely, industrial risk, management risk, financial flexibility, credibility, competitiveness, and operating risk. The output has two classes of nominal types, which describe the instances as "Bankruptcy" (107 cases) or "Non-bankruptcy" (143 cases). The Distress dataset is from the Kaggle dataset and can be found in <https://www.kaggle.com/shebrahimi/financial-distress>. This dataset addresses financial distress prediction for a sampling of companies. The first column represents the sample companies, which include 422 companies. The second column shows different time periods to which the data belong. The time series length varies between 1 and 14 for each company. The third column, named the target variable, is the "Financial Distress". If this value is higher than -0.50, the company should be considered as healthy; otherwise, it is regarded as financially distressed. The fourth-to-last column denotes the features, which are denoted x_1 to x_{83} ; they represent some financial and nonfinancial characteristics of the sample companies. These features belong to the previous

period, which should be used to predict whether the company will be financially distressed (classification). Until now, there has been no relevant literature adopting these datasets to solve the problem of bankruptcy prediction.

4.2. Data Preprocessing. Generally, data preprocessing is an initial and basic step of data analysis. Because artificial neural networks require that every data sample be expressed as a real number vector, we need to change the nominal attributes of the data samples into numerical values before inputting them into the classifier [52].

There are no missing values in the Qualitative Bankruptcy dataset, but all the attributes are nominal. This dataset includes 250 samples, and each sample possesses 6 features. The 6 features are all represented by 3 labels: "P", "A", and "N". We convert the qualitative features into the values 1, 2, and 3, respectively.

The original Distress dataset is an extensive dataset; it includes 422 companies, and each company behaves differently in different time series. Moreover, this dataset is imbalanced and skewed; there are 136 financially distressed companies against 286 healthy ones, 136 firm-year observations are financially distressed, while 3546 firm-year observations are healthy. To make the structure of the distress dataset under observation be similar to the Qualitative Bankruptcy dataset, we perform some preprocessing. First, all the distressed companies are chosen from time series period 1 to period 14, and the total number of distressed companies is 126. In each time series period, 15 healthy companies are selected randomly, and the number of healthy companies is 210. Thus, there are 336 samples remaining in the newly generated dataset. Because each company presents 83 features, this dataset remains relatively large. We have adopted the minimal-redundancy-maximal-relevance (mRMR) criterion to generate the feature selection. The mRMR criterion offers an excellent way to maximize the dependence of the results on the input features by combining the max-relevance criterion with the min-redundancy criterion. Moreover, mRMR can not only enhance the appropriate feature selection but also achieve high classification accuracy and high computation speeds [53]. The max-relevance mechanism of mRMR is defined as follows:

$$\max D(S, c), D = \frac{1}{|S|} \sum_{x_i, x_j \in S} I(x_i, c), \quad (14)$$

where S is the set that both contains m individual features x_i and has the most considerable dependency on the target class c . If the features selected according to the max-relevance criterion are of high redundancy, there exists a large dependency among these features. To increase the respective class-discriminative power, the minimal redundancy (min-redundancy) condition is added to select noninteracting features [54],

$$\min D(S), R = \frac{1}{|S|^2} \sum_{x_i, x_j \in S} I(x_i, x_j). \quad (15)$$

TABLE 1: Parameter levels in EPNN of the Qualitative Bankruptcy data set.

D	K	θ_{soma}
8, 10, 12, 14	3, 5, 7, 10	0.1, 0.3, 0.5, 0.7

TABLE 2: $L_{16}(4^3)$ orthogonal array and factor assignment of the Qualitative Bankruptcy data set.

Expe. NO./Parameter	D	K	θ_{soma}	Testing accuracy
1	8	3	0.1	99.52 ± 1.00
2	8	5	0.3	99.01 ± 1.61
3	8	7	0.5	99.15 ± 1.26
4	8	10	0.7	99.01 ± 1.24
5	10	3	0.3	99.57 ± 0.55
6	10	5	0.1	99.23 ± 1.02
7	10	7	0.9	99.12 ± 1.32
8	10	10	0.7	99.12 ± 1.06
9	12	3	0.5	99.15 ± 1.28
10	12	5	0.7	99.23 ± 1.27
11	12	7	0.1	98.88 ± 1.83
12	12	10	0.3	99.12 ± 1.41
13	14	3	0.7	98.64 ± 2.10
14	14	5	0.5	99.15 ± 1.15
15	14	7	0.3	99.07 ± 1.56
16	14	10	0.1	98.93 ± 1.43

mRMR combines the two constraints through the operator $\Phi(D, R)$ by adopting a simple form to optimize D and R simultaneously.

$$\max \Phi(D, R), \Phi = D - R. \quad (16)$$

Using mRMR, we sort the features of this dataset. x_{43} , x_{75} , x_{52} , x_{45} , x_{80} , x_{64} , x_{60} , x_{48} , x_{34} , and x_{76} are the first ten max-dependent and min-redundant features and are used in our experiments. The updated Distress dataset includes 336 samples, where each sample has 10 features.

4.3. Optimal Parameter Setting. To realize a specific accuracy rate and achieve fast convergence in the training dataset, an optimal set of parameters must be selected. The Taguchi method is employed to decrease the number of experimental runs using orthogonal arrays [55]. Under this method, the time cost, human effort required, and material requirements can also be effectively controlled in our simulation. Selecting the orthogonal arrays that are proper for the simulation is a vital step. First, three parameters, D , K , θ_{soma} , are considered to be important in the EPNN. D denotes the branch number of the dendritic layer, K represents a parameter of the sigmoid function in the synaptic layer, and θ_{soma} denotes the threshold of the soma. Tables 1 and 3 show the ranges of parameter values of the two datasets. There are 3 parameter trials, and each parameter contains 4 values. The $L_{16}(4^3)$ orthogonal arrays of the two datasets are presented in Tables 2 and 4. To obtain a reliable average testing accuracy, each experiment is repeated 30 times. The population sizes are set to 50, and the maximum number of generations is set to 1000. The accuracy

TABLE 3: Parameter levels in EPNN of the Distress credit data set.

D	K	θ_{soma}
12, 15, 18, 21	3, 5, 7, 10	0.3, 0.5, 0.7, 0.9

rate results of the Qualitative Bankruptcy dataset and Distress dataset are shown in Tables 2 and 4.

From Table 2, it is obvious that the highest classification accuracy of the Qualitative Bankruptcy dataset is achieved by the combination of the parameters $D = 10$, $K = 3$, and $\theta_{soma} = 0.3$. In addition, from Table 4, the best performance of the Distress dataset is $D = 18$, $K = 10$, and $\theta_{soma} = 0.5$. These parameter sets are reasonable for obtaining acceptable performance, and they are optimal for further comparisons with other algorithms.

In addition, because both MLP and PNN are adopted as the competitors in our experiment, several other parameters for these algorithms are considered cautiously. Table 5 lists the relevant parameters.

Next, to make a fair comparison, the performances of the EPNN, MLP, and PNN should be compared with an approximately equal number of thresholds and weights. In addition, the learning rate of the PNN and MLP trained by using the back-error propagation algorithm is set to 0.01. The number of dendrites in the PNN should be defined in this simulation, as should the number of hidden layers and neurons of the MLP. The MLP's parameter number depends mainly on the number of neurons in the hidden layer. Thus, we denote MLP as a D -dimensional vector, which consists

TABLE 4: $L_{16}(4^3)$ orthogonal array and factor assignment of the Distress data set.

Expe. NO./Parameter	D	K	θ_{soma}	Testing accuracy
1	12	3	0.3	76.25 ± 4.04
2	12	5	0.5	75.73 ± 2.64
3	12	7	0.7	75.69 ± 3.14
4	12	10	0.9	74.78 ± 2.77
5	15	3	0.5	76.15 ± 3.24
6	15	5	0.3	76.31 ± 3.66
7	15	7	0.9	76.11 ± 3.81
8	15	10	0.7	75.83 ± 2.27
9	18	3	0.7	75.06 ± 2.93
10	18	5	0.9	75.10 ± 3.60
11	18	7	0.3	75.48 ± 3.02
12	18	10	0.5	76.41 ± 3.23
13	21	3	0.9	75.20 ± 2.42
14	21	5	0.7	75.08 ± 3.60
15	21	7	0.5	75.42 ± 3.53
16	21	10	0.3	76.27 ± 2.83

TABLE 5: Initial Parameters of the algorithms in comparison.

Methods	Relative Parameters	Parameters' values
EPNN	popSize	50
	Max-gen	1000
	D	10(Bankruptcy),18(Distress)
	K	3(Bankruptcy),10(Distress)
	θ_{soma}	0.3(Bankruptcy),0.5(Distress)
PNN	Max-gen	1000
	lr	0.01
	D	10(Bankruptcy),18(Distress)
	K	3(Bankruptcy),10(Distress)
	θ_{soma}	0.3(Bankruptcy),0.5(Distress)
MLP	epoch	1000
	lr	0.01

of weights and biases. The dimension number D can be calculated as follows:

$$D = (Input \times Hidden) + (Hidden \times Output) + Hidden_{bias} + Output_{bias}, \quad (17)$$

where $Input$, $Hidden$, and $Output$ denote the neuron numbers in the input, hidden, and output layers of the MLP, respectively. $Hidden_{bias}$ and $Output_{bias}$ represent the bias in the hidden and output layer [56].

Meanwhile, in the PNN and EPNN, the synaptic input number is $Input$, and the dendritic branch number is $Branch$. The dimension number D of the PNN and EPNN can be calculated in the following equation:

$$D = 2 * Input \times Branch. \quad (18)$$

The structural description of the MLP and EPNN is shown in Table 6. Both models have nearly equal parameter numbers for the two datasets.

To evaluate the performance of the classification methods, each dataset is separated randomly into two subsets: a training set and a testing set. The training subset is used to set up the classification model, and the testing dataset is adopted to test the model's accuracy. The splitting strategy is significantly relevant to achieve reliable model evaluation because the case data are usually very scarce. According to prior experiments, a larger training set results in a better classifier [57]. In this simulation, 50% of the samples are chosen randomly for training, while the remaining 50% is for testing to guarantee high test accuracy. The average value of the classification accuracy rate over 30 runs is regarded as the overall classification performance.

4.4. Performance Measures. In general, most performance evaluation metrics attempt to estimate how well the learned model predicts the correct class of new input samples; however, different metrics yield different orderings of model performances [58]. The classification accuracy has been by far

TABLE 6: Structures of EPNN and MLP for the Qualitative Bankruptcy and Distress data set.

Dataset	Model	No. of input	No. of Branch Hidden node	No. of output	No. of adjusted parameters
Qualitative Bankruptcy	EPNN	6	10	1	120
	MLP	6	15	1	121
Distress	EPNN	10	18	1	360
	MLP	10	30	1	361

TABLE 7: Contingency matrix of prediction results.

Hypothesis class	Real class	
	Positive	Negative
Positive	True Positive (TP)	False Positive (FP)
Negative	False Negative (FN)	True Negative (TN)

the most frequently adopted indicator of performance in the literature [51]. Sensitivity and specificity can be highlighted as straightforward indices. The AUC does not implicitly assume equal misclassification costs, and it corresponds to the most preferred score calculated as the empirical probability that a randomly chosen positive observation ranks above a randomly chosen negative sample [59]. Hence, the overall accuracy rate, sensitivity, specificity, and AUC are used to construct the performance evaluation system.

Table 7 demonstrates that the result of a classifier can be measured by a 2-dimensional contingency matrix. The accuracy rate is the critical indicator in evaluating the classification algorithms; another indicator of accuracy analysis is related to the possibility of misclassifying bankruptcies. The classification accuracy rate is measured by the following equation:

$$Accuracy\ rate = \frac{TP + TN}{TP + TN + FP + FN} (\%), \quad (19)$$

where TP, TN, FP, and FN represent true positive, true negative, false positive, and false negative, respectively. True positive (TP) means that the company is detected as healthy, and the teacher target label is healthy as well. True negative (TN) represents that the input and the teacher target label are detected as unhealthy simultaneously. False positive (FP) shows that the input is detected as healthy, whereas the teacher target label is unhealthy. False negative (FN) denotes that input is detected unhealthy, and the teacher target label shows the opposite,

$$Sensitivity = \frac{TP}{TP + FN} (\%), \quad (20)$$

$$Specificity = \frac{TN}{TN + FP} (\%), \quad (21)$$

$$AUC (\%) = \frac{1}{2} \left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP} \right) \times 100 (\%), \quad (22)$$

where the sensitivity and specificity are called the true positive rate and true negative rate, respectively. Sensitivity

measures the percentage of real positives that are identified correctly. This metric shows how successfully a classifier can identify regular records, which means that the companies are healthy in terms of bankruptcy analysis. Therefore, financial institutions can achieve correct and efficient analysis by adopting a classifier with a higher sensitivity. Specificity represents how successfully a classifier can distinguish abnormal records; i.e., it is the proportion of true negatives. Hence, a higher specificity can help financial institutions reduce the possibility of misclassifying healthy companies. AUC represents the ratio of companies that are not in danger of bankruptcy. In other words, a score of 100% indicates that two classes can be correctly discriminated by the classifier, whereas a score of 50% indicates that the classifier has an insignificant ability to classify companies correctly.

In addition to comparing different classification algorithms, the convergence performances of the two models, EPNN and MLP, are compared. When the mean squared error (MSE) achieves a predetermined minimum value, the learning tends to be completed. The training error is calculated as shown in (21),

$$MSE = \frac{1}{R} \sum_{a=1}^R \left[\frac{1}{S} \sum_{b=1}^S (E_{ab} - O_{ab})^2 \right], \quad (23)$$

where E_{ab} and O_{ab} are the desired output and the actual output, respectively; S represents the number of instances applied for training; and R denotes the number of simulation runs.

4.5. Performance Comparison. For a fair comparison, EPNN and PNN are equipped with the same parameters, and the learning rates of the PNN and MLP are the same. All three algorithms are run 30 times independently. Tables 8 and 13 show the classification performances obtained by these algorithms. In addition, to detect the significant differences among the results, a nonparametric test called the Wilcoxon rank-sum test [60] is adopted in this study. A review in the literature has summarized that it is preferable to use a nonparametric test instead of a parametric test to achieve high

TABLE 8: Experimental results for Qualitative Bankruptcy data set.

Method	Average accuracy (AVE±STD)	Wilcoxon rank-sum test (<i>p</i> -value)	Sensitivity	Specificity	AUC (AVE±STD)	Wilcoxon rank-sum test (<i>p</i> -value)
EPNN	99.57±0.5452	N/A	0.9976	0.9775	0.9984±0.0055	N/A
PNN	98.11±1.2690	$2.82E^{-05}$	0.9852	0.9185	0.9871±0.0195	$1.20E^{-03}$
MLP	94.59±3.3330	$8.76E^{-07}$	0.9393	0.9552	0.9868±0.0147	$1.50E^{-05}$

TABLE 9: Initialization parameters of other classification methods.

Method	Parameter	Value
KNN	k	5
RBF	Spread of radial basis function	1.0
RF	Number of trees	50
DT	Children	1
SVM	Type of kernel function	Radial basis function
	Degree	3
DA	Prior probabilities	uniform

statistical accuracy, especially when the sample size is small [61]. Thus, the calculated *p*-values of the Wilcoxon rank-sum test are presented in the tables as well. In the following comparison tables, N/A represents “Not Applicable”, which indicates that the relevant algorithm cannot be compared with itself in the test. In our experiments, the significance level is set to 5%. As a matter of routine, there is substantial evidence to reject the null hypothesis when *p*-values are less than 0.05. In order to further verify the superiority of EPNN, we compare it with other popularly applied classification methods, such as K-nearest neighbor algorithm (KNN) [62], radical basis function (RBF) [63], random forest (RF) [64], decision tree (DT) [65], support vector machine (SVM) [66], and discriminant analysis (DA) [67]. Each method runs 30 times independently. The initial parameters of each method are summarized in Table 9.

4.5.1. Qualitative Bankruptcy Dataset. For the Qualitative Bankruptcy dataset, as shown in Table 8, the proposed EPNN obtains an average testing accuracy of 99.57%, which is higher than the 98.11% obtained by the PNN and the 94.59% obtained by MLP. In addition, the statistical results also show that the EPNN achieves significantly better performances than the PNN and MLP. Moreover, the EPNN also performs better than the PNN and MLP in terms of sensitivity and specificity. A comparatively higher sensitivity value indicates the powerful capability of the EPNN in identifying the companies that are healthy. Higher specificity values represent the EPNN’s ability to not misclassify an unhealthy company. Furthermore, the convergence rate of the three models, EPNN, PNN, and MLP, are also compared in our experiments in Figure 4. As observed, the EPNN achieves the highest convergence rate compared to the PNN and MLP. Moreover, Figure 5 shows the ROC of the EPNN, the PNN, and MLP. The corresponding AUC value of the EPNN is larger than that of the PNN and of MLP. It is emphasized that the EPNN is superior to the PNN and MLP in solving the Qualitative Bankruptcy dataset problem.

TABLE 10: Classification performance of EPNN in comparison with other classification methods on Qualitative Bankruptcy dataset.

Method	Average accuracy MEAN±STD
KNN	99.12±0.6759
RBF	96.75±2.2425
RF	98.99±1.2238
DT	97.92±1.1810
SVM	99.39±0.5009
DA	99.52±0.4506
EPNN	99.57±0.5452

In addition, the performances’ comparisons between the EPNN and other commonly used classifiers are presented in Table 10. It is clear that the EPNN also shows its superiority in the average accuracy rate on Qualitative Bankruptcy dataset.

Since there are many proposed methods which are adopted to classify Qualitative Bankruptcy dataset in the relative literatures, we summarized the classification performances and compared them with that of the EPNN. Specifically, Table 11 presents some single classification methods and Table 12 demonstrates some hybrid classification methods, respectively. From Table 11, it can be observed that the accuracy rate of the EPNN is only slightly less than RBF-based SVM, Ant-miner, and Random Forest. As Table 12 shows, compared with other hybrid classification methods, the average accuracy rate of the EPNN is only slightly lower than the hybrid logistic regression-naive bayes. Thus, it can be concluded that although the EPNN adopted 50%-50% train-to-test ratio, it has still presented a competitive performance on the Qualitative Bankruptcy dataset. And it is worth mentioning that hybrid classification methods are not always superior to single classification methods based on the above experimental results.

TABLE 11: Classification accuracies for Qualitative Bankruptcy dataset obtained by other single classification methods in relative literatures.

Author (year)	Method (train-to-test ratios)	Average accuracy (%)
Kalyan Nagaraj, Amulyashree Sridhar (2015) [40]	Logistic Regression (2/3-1/3)	97.2
Kalyan Nagaraj, Amulyashree Sridhar (2015) [40]	Rotation Forest (2/3-1/3)	97.4
Kalyan Nagaraj, Amulyashree Sridhar (2015) [40]	Naive Bayes (2/3-1/3)	98.3
Kalyan Nagaraj, Amulyashree Sridhar (2015) [40]	RBF-based SVM (2/3-1/3)	99.6
E. K. Kornoushenko (2017) [41]	Nearest Neighborhood (50%-50%)	97.6
J.Uthayakumar et. al. (2017) [42]	Ant-Miner (10 fold cross validation)	100
J.Uthayakumar et. al. (2017) [42]	Logistic Regression (10 fold cross validation)	99.2
J.Uthayakumar et. al. (2017) [42]	MLP (10 fold cross validation)	99.2
J.Uthayakumar et. al. (2017) [42]	Random Forest (10 fold cross validation)	100
J.Uthayakumar et. al. (2017) [42]	Radical Basis Function (10 fold cross validation)	99.2
J.Uthayakumar et. al. (2018) [43]	Genetic Algorithm (Not Mentioned)	71.48
J.Uthayakumar et. al. (2018) [43]	Ant Colony Algorithm (Not Mentioned)	83.05
Our Method (2019)	EPNN (50%-50%)	99.57
Our Method (2019)	EPNN(10 fold cross validation)	99.68

TABLE 12: Classification accuracies for Qualitative Bankruptcy dataset obtained by other hybrid classification methods in relative literatures.

Author (year)	Method (train-to-test ratios)	Average accuracy (%)
Yi Tan et. al. (2016) [44]	Hybrid logistic regression-naive bayes (90%-10%)	99.64
Nanxi Wang (2017) [45]	Neural network model with robust logistic regression (50%-50%)	69.44
Nanxi Wang (2017) [45]	Neural network model with inductive learning algorithm (50%-50%)	89.7
Nanxi Wang (2017) [45]	Neural network model with genetic algorithm (50%-50%)	94
Nanxi Wang (2017) [45]	Neural network model with neural networks without dropout (50%-50%)	90.3
Nanxi Wang (2017) [45]	Neural network model with SVM (50%-50%)	98.67
Nanxi Wang (2017) [45]	Neural network model with decision tree (50%-50%)	99.33
J.Uthayakumar et. al. (2018) [43]	Genetic ant colony algorithm (Not mentioned)	91.32
J.Uthayakumar et. al. (2018) [43]	Fitness-scaling chaotic Genetic ant colony algorithm (Not mentioned)	92.14
J.Uthayakumar et. al. (2018) [43]	Improved K-means clustering and fitness-scaling chaotic genetic ant colony algorithm (Not mentioned)	97.93
Our Method (2019)	EPNN (50%-50%)	99.57
Our Method (2019)	EPNN (10 fold cross validation)	99.68

TABLE 13: Experimental results for Distress data set.

Method	Average accuracy (AVE±STD)	Wilcoxon rank-sum test (<i>p</i> -value)	Sensitivity	Specificity	AUC (AVE±STD)	Wilcoxon rank-sum test (<i>p</i> -value)
EPNN	76.41±3.2320	N/A	0.8108	0.6546	0.7762±0.0478	N/A
PNN	54.03±19.636	$2.43E^{-05}$	0.3166	0.8353	0.6338±0.1848	$6.67E^{-04}$
MLP	66.15±5.4324	$9.09E^{-07}$	0.8475	0.3575	0.6814±0.0906	$2.13E^{-05}$

4.5.2. *Distress Dataset.* Concerning the Distress dataset, as shown in Table 13, the EPNN acquires an average testing accuracy of 76.41%, which is higher than the 54.03% obtained by the PNN and the 66.15% accuracy obtained by MLP. In addition, the *p*-values of the Wilcoxon test show there are significance differences between EPNN and the other two methods. Although not all the sensitivity and specificity values of the EPNN are larger than those of the PNN and MLP, the PNN performs worse on sensitivity, and MLP is the worst on specificity. The EPNN achieves better performances on both sensitivity and specificity. The convergence curves of

the EPNN, the PNN, and MLP are compared in Figure 6. This figure shows that the EPNN achieves the highest convergence rate compared to the PNN and MLP. In addition, Figure 7 presents the ROC of the EPNN, PNN, and MLP. In addition, the corresponding AUC value of the EPNN is larger than that of the PNN and MLP. This implies that, compared with the PNN and MLP, the EPNN is a more effective classifier on the Distress dataset. Besides, the classification performance of the EPNN is compared with KNN, RBF, RF, DT, SVM, and DA, and the corresponding results are presented in Table 14. As Table 14 illustrated, EPNN performs better than all the other

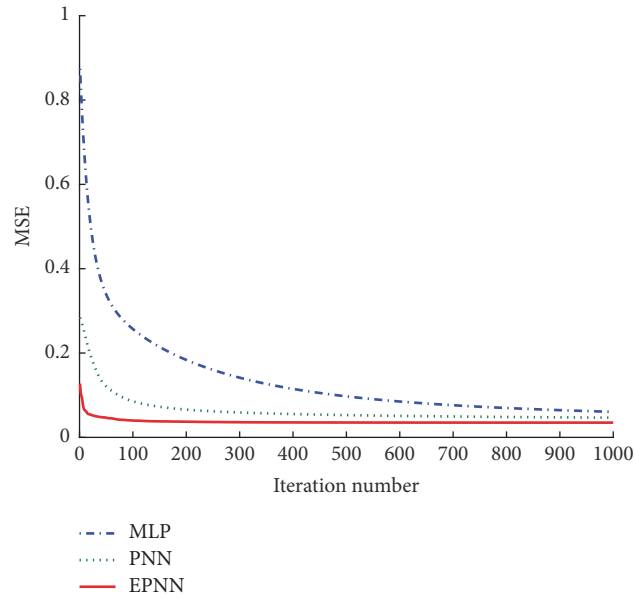


FIGURE 4: The convergency curve of the Qualitative Bankruptcy data set.

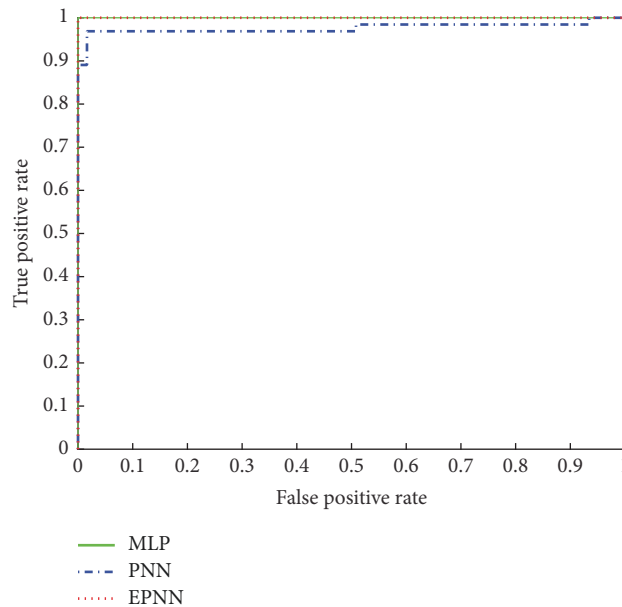


FIGURE 5: The ROC of the Qualitative Bankruptcy data set.

classification methods except RF. Since there are no other classification methods applied to classify Distress dataset in the literature, horizontal comparison can not be fulfilled for this dataset.

4.6. Dendrite Morphology Reconstruction

4.6.1. The Ultimate Synaptic and Dendritic Morphology. As mentioned above, the EPNN can implement synaptic pruning and dendritic pruning during the training process. Thus,

superfluous synapses and dendrites can be removed, and then, a simplified and distinct topological morphology is formed for each problem. Figure 8 shows the particular dendritic structure of the EPNN on the Qualitative Bankruptcy dataset after learning. The unnecessary dendrites (Branch 2, Branch 4 and Branch 9) of the PNN are presented in Figure 9, and superfluous synaptic layers are provided in Figure 10. Finally, the simplified structural morphology is described in Figure 11. It can be observed that 7 dendritic branches and 4 features are reserved in the structure. This means that the features x_1 and x_2 are not crucial for the EPNN and

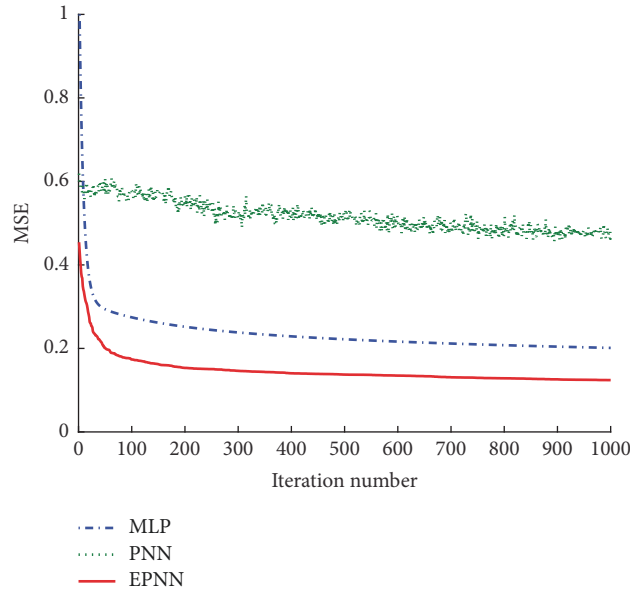


FIGURE 6: The convergency curve of the Distress data set.

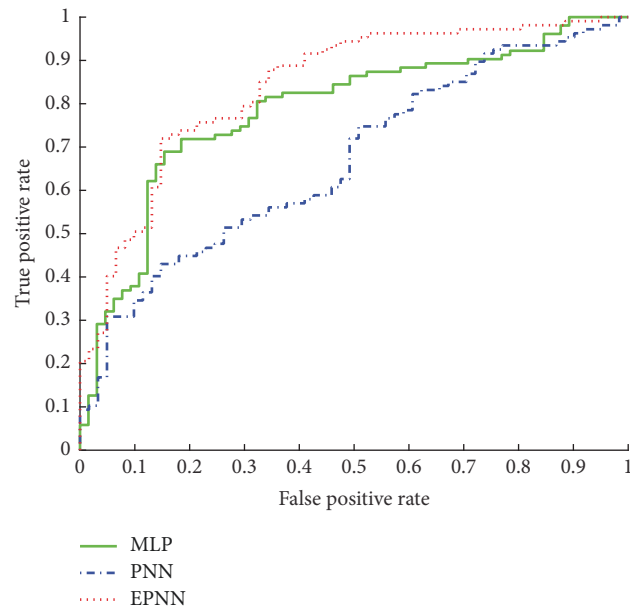


FIGURE 7: The ROC of the Distress data set.

have no contribution to solving the Qualitative Bankruptcy dataset problem. In addition, Figure 12 illustrates the unique dendritic morphology of the EPNN on the Distress dataset after learning. Figure 13 shows that all ineffective dendrites are removed, and Figure 14 rules out all ineffective synaptic layers. Thus, the final structural morphology is presented in Figure 15. Only four branches of the dendrites are remaining, and the feature x_6 is removed. As summarized in Table 15, it can be observed that synaptic pruning and dendritic pruning mechanism can largely simplify the structure of the EPNN. Thus, it is able to speed up the bankruptcy prediction analysis by the simplified EPNN obviously.

4.6.2. *The Simplified Logic Circuit of the After-Learning Morphology.* In addition to the neural pruning function, the other function worth emphasizing is that the simplified structure of the EPNN can form an approximate logic circuit applicable to hardware implementations. Figures 16 and 17 present further simplified logic circuits of the structural morphologies. We use an analog-to-digital converter, which can be regarded as a “comparator”, to compare the input with the threshold θ . Once the input x_i is less than the threshold θ , the “comparator” will output 0; otherwise, it will output 1. Using the logic circuits, we can classify the companies into bankrupt and not-bankrupt on both the Qualitative

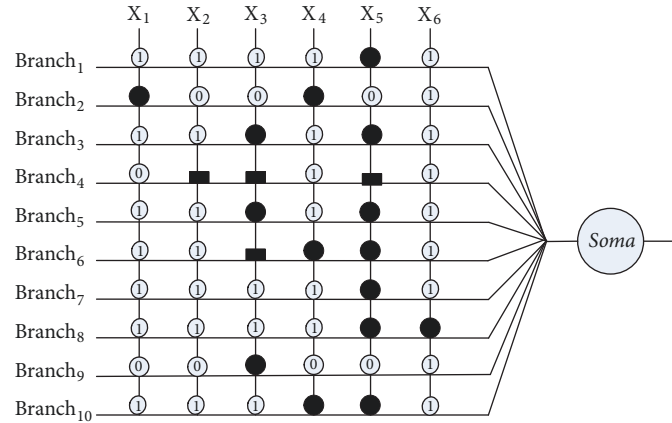


FIGURE 8: The dendritic morphology of the Qualitative Bankruptcy data set after learning.

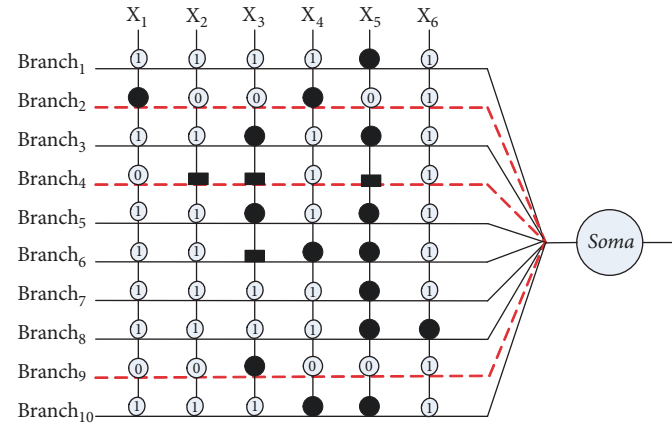


FIGURE 9: The dendritic morphology of the Qualitative Bankruptcy data set after dendritic pruning.

TABLE 14: Classification performance of EPNN in comparison with other classification methods on Distress dataset.

Method	Average accuracy
	MEAN±STD
KNN	76.35±3.1065
RBF	49.15±14.4445
RF	81.25±2.2907
DT	76.11±2.8599
SVM	62.42±2.5047
DA	74.74±3.4414
EPNN	76.41±3.2320

Bankruptcy dataset and Distress dataset. The accuracies of the logic circuits are shown in Table 16. Clearly, the test accuracies of the logic circuits do not decrease and are higher than those of the EPNN. Note that the logic circuits in Figures 16 and 17 are selected randomly from an arbitrarily chosen experiment, and they are not unique to each problem. Forming a logic circuit can further increase the classification speed of the

EPNN, thereby creating a more powerful method for the prediction of financial bankruptcy.

5. Conclusion

Artificial intelligence algorithms, such as neural network methods, are being widely applied in bankruptcy analysis. In this paper, we introduce a more realistic neural model called the EPNN to facilitate bankruptcy analysis. This technique adopts the JADE algorithm to train the model to obtain satisfactory classification performances. In contrast with the PNN and MLP, the proposed EPNN performs the best in terms of the average accuracy and AUC on both benchmark and application-oriented datasets, namely, the Qualitative Bankruptcy dataset and the Distress dataset. In addition, compared with other classification methods such as KNN, RBF, RF, DT, SVM, and DA, the EPNN also provides competitive and satisfactory classification performances. Note that the neuronal pruning mechanism is an important aspect of the EPNN. After synaptic pruning and dendritic pruning, the number of input features in both datasets is reduced, and

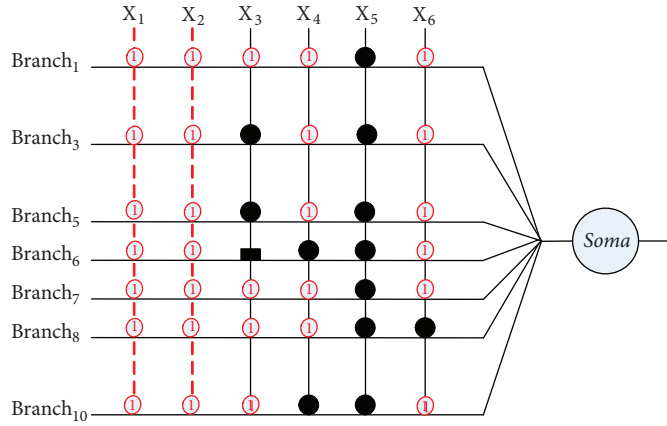


FIGURE 10: The dendritic morphology of the Qualitative Bankruptcy data set after synaptic pruning.

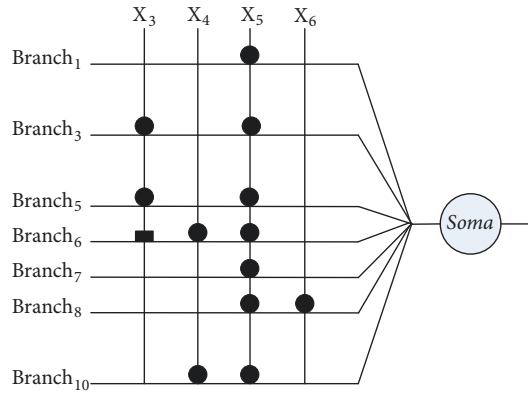


FIGURE 11: The final structure of the Qualitative Bankruptcy data set.

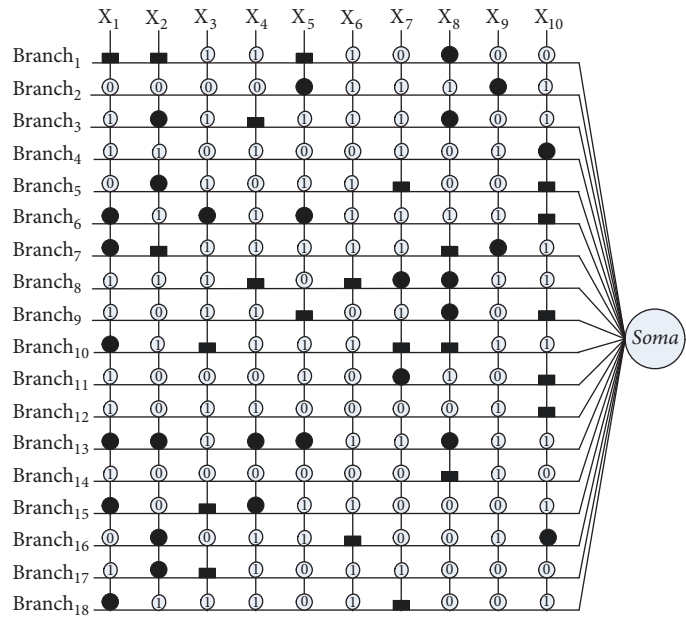


FIGURE 12: The dendritic morphology of the Distress data set after learning.

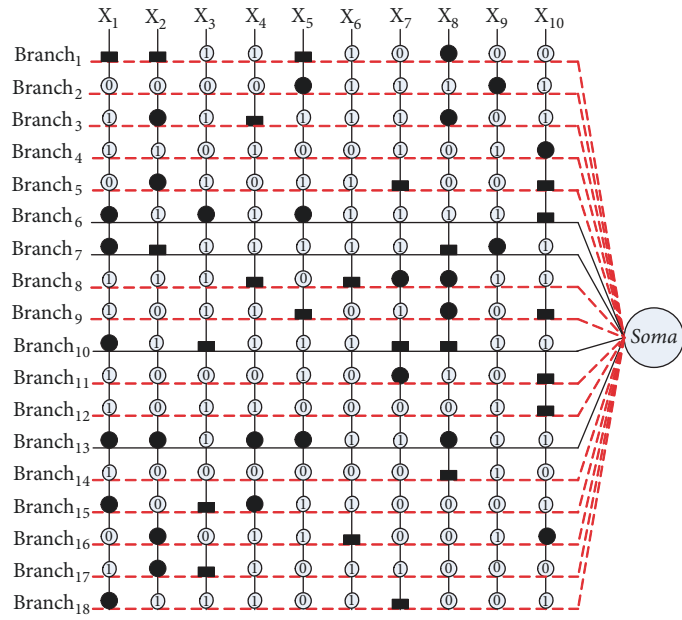


FIGURE 13: The dendritic morphology of the Distress data set after dendritic pruning.

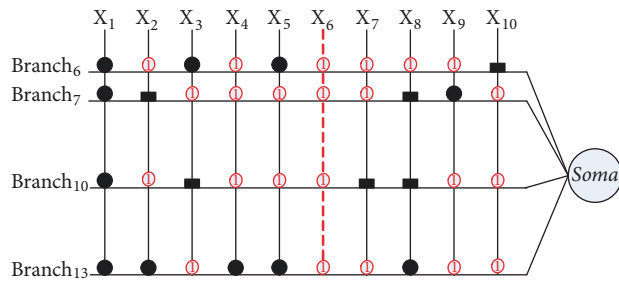


FIGURE 14: The dendritic morphology of the Distress data set after synaptic pruning.

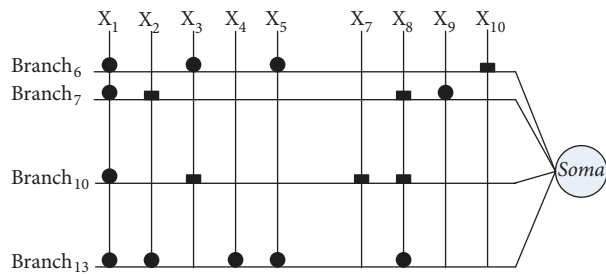


FIGURE 15: The final structure of the Distress data set.

TABLE 15: Model structure comparison between Qualitative Bankruptcy and Distress data sets.

Dataset	feature input		Dendritic layer		Adjusted weight	
	initial value	selected value	initial value	selected value	initial value	selected value
Qualitative Bankruptcy	6	4	10	7	120	56
Distress	10	9	18	4	360	72

TABLE 16: Verification of the logic circuits.

Dataset	Accuracy of the EPNN	Accuracy of the logic circuit
Qualitative Bankruptcy	99.57	100
Distress	76.41	79.17

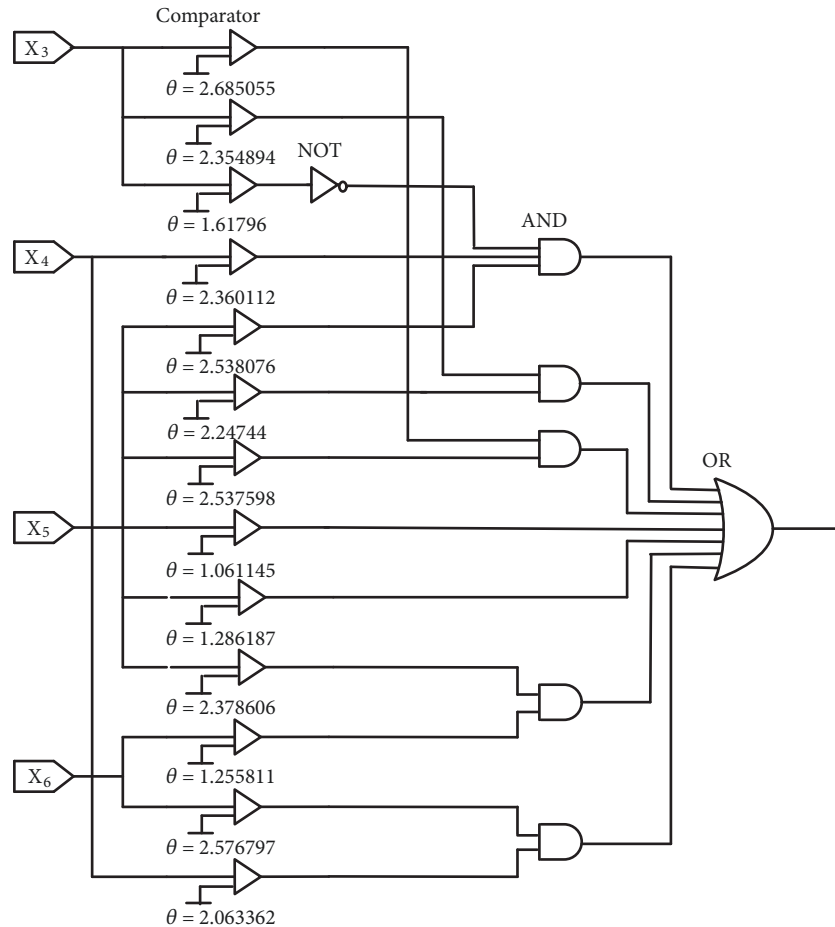


FIGURE 16: The logic circuit of the Qualitative Bankruptcy data set.

the structure of the neural network is simplified. Moreover, the simplified structural morphology can form a logic circuit which can also be employed as a powerful tool to solve bankruptcy prediction problems. Thus, the contribution of this paper can be summarized from three aspects: First, we provide a comprehensive study by comparing different classification models in terms of bankruptcy prediction problems. Although many novel algorithms are continually emerging, a large proportion of approaches still only focus on the bankruptcy prediction model's ability to improve the classification accuracy. Compared with some other models, the EPNN possesses a certain advantage with respect to average accuracy and AUC. Second, the EPNN can implement synaptic and dendritic pruning to realize pattern extraction and reconstruct a more compact neuronal morphology. The EPNN has a large initial neuronal topology, which makes it not very sensitive to its initial conditions, but it can utilize neuronal pruning after learning, which increases the efficiency of the neural network, speeds up the convergence, avoids becoming trapped in local minima, and reduces the operation time and computational cost. Third, the simplified models can be replaced by logic circuits, which can increase the classification accuracy and be easily implemented on the

hardware. Therefore, these findings provide details and offer insight into technical development for understanding and tracing the operating mechanisms and construction of single neurons. In addition, the results also imply that the proposed EPNN classifier possesses an excellent potential to be applied in other binary classification problems. The EPNN makes it possible to draw standard profiles of the failing companies and provide a theoretical contribution to the phenomenon of bankruptcy. It is believed that the EPNN will be suitable for not only bankruptcy prediction but also other fields of application within the scope of financial analysis such as performance analysis.

Data Availability

No data were used to support this study.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

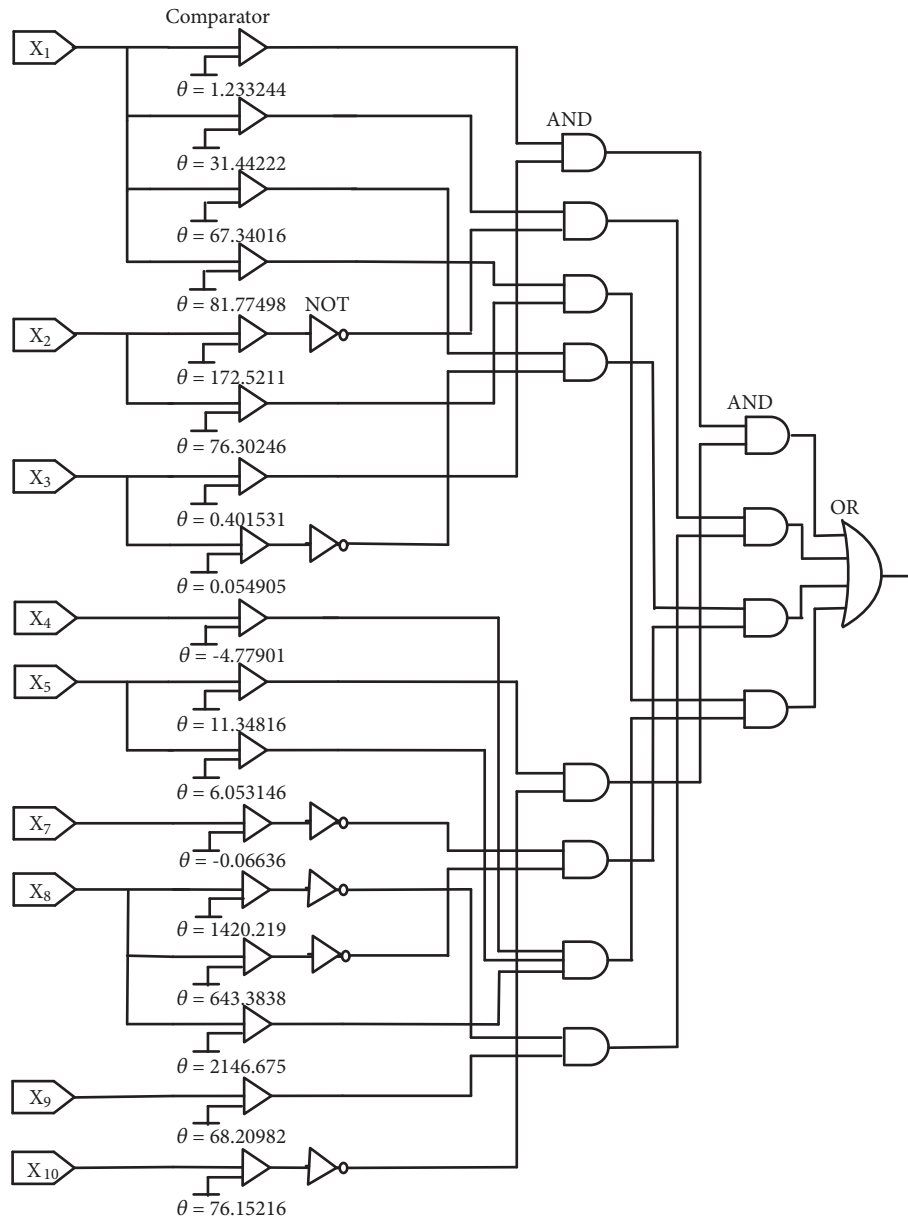


FIGURE 17: The logic circuit of the Distress data set.

References

- [1] E. I. Altman, "Financial ratios, discriminant analysis and the prediction of corporate bankruptcy," *The Journal of Finance*, vol. 23, no. 4, pp. 589–609, 1968.
- [2] J. A. Ohlson, "Financial ratios and the probabilistic prediction of bankruptcy," *Journal of Accounting Research*, vol. 18, no. 1, pp. 109–131, 1980.
- [3] J. Begley, J. Ming, and S. Watts, "Bankruptcy classification errors in the 1980s: an empirical analysis of altman's and ohlson's models," *Review of Accounting Studies*, vol. 1, no. 4, pp. 267–284, 1996.
- [4] J. Kruppa, A. Schwarz, G. Arminger, and A. Ziegler, "Consumer credit risk: Individual probability estimates using machine learning," *Expert Systems with Applications*, vol. 40, no. 13, pp. 5125–5131, 2013.
- [5] R. Pal, K. Kupka, A. P. Aneja, and J. Militky, "Business health characterization: A hybrid regression and support vector machine analysis," *Expert Systems with Applications*, vol. 49, pp. 48–59, 2016.
- [6] A. E. Celik and Y. Karatepe, "Evaluating and forecasting banking crises through neural network models: An application for Turkish banking sector," *Expert Systems with Applications*, vol. 33, no. 4, pp. 809–815, 2007.
- [7] M. Perez, "Artificial neural networks and bankruptcy forecasting: A state of the art," *Neural Computing and Applications*, vol. 15, no. 2, pp. 154–163, 2006.
- [8] H. A. Abdou and J. Pointon, "Credit scoring, statistical techniques and evaluation criteria: a review of the literature," *Intelligent Systems in Accounting, Finance and Management*, vol. 18, no. 2-3, pp. 59–88, 2011.

- [9] M. D. Odom and R. Sharda, "A neural network model for bankruptcy prediction," in *Proceedings of the International Joint Conference on Neural Prediction Networks (IJCNN '90)*, pp. 163–168, IEEE, 1990.
- [10] T. E. McKee and M. Greenstein, "Predicting bankruptcy using recursive partitioning and a realistically proportioned data set," *Journal of Forecasting*, vol. 19, no. 3, pp. 219–230, 2000.
- [11] G. Zhang, M. Y. Hu, B. E. Patuwo, and D. C. Indro, "Artificial neural networks in bankruptcy prediction: general framework and cross-validation analysis," *European Journal of Operational Research*, vol. 116, no. 1, pp. 16–32, 1999.
- [12] T. Bellotti and J. Crook, "Support vector machines for credit scoring and discovery of significant features," *Expert Systems with Applications*, vol. 36, no. 2, pp. 3302–3308, 2009.
- [13] K. Hornik, "Approximation capabilities of multilayer feedforward networks," *Neural Networks*, vol. 4, no. 2, pp. 251–257, 1991.
- [14] K. Hornik, "Some new results on neural network approximation," *Neural Networks*, vol. 6, no. 8, pp. 1069–1072, 1993.
- [15] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Networks*, vol. 2, no. 5, pp. 359–366, 1989.
- [16] W. Chen and Y. Du, "Using neural networks and data mining techniques for the financial distress prediction model," *Expert Systems with Applications*, vol. 36, no. 2, pp. 4075–4086, 2009.
- [17] H. Wang, Q. Xu, and L. Zhou, "Large unbalanced credit scoring using lasso-logistic regression ensemble," *PLoS ONE*, vol. 10, no. 2, Article ID e0117844, 2015.
- [18] C. Koch, T. Poggio, and V. Torre, "Nonlinear interactions in a dendritic tree: Localization, timing, and role in information processing," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 80, no. 9 I, pp. 2799–2802, 1983.
- [19] C. Koch, T. Poggio, and V. Torre, "Retinal ganglion cells: a functional interpretation of dendritic morphology," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 298, no. 1090, pp. 227–263, 1982.
- [20] C. Koch, "Computation and the single neuron," *Nature*, vol. 385, no. 6613, pp. 207–210, 1997.
- [21] J. Ji, S. Gao, J. Cheng, Z. Tang, and Y. Todo, "An approximate logic neuron model with a dendritic structure," *Neurocomputing*, vol. 173, pp. 1775–1783, 2016.
- [22] Y. Tang, J. Ji, S. Gao et al., "A pruning neural network model in credit classification analysis," *Computational Intelligence and Neuroscience*, vol. 2018, Article ID 9390410, 22 pages, 2018.
- [23] S. Piramuthu, M. J. Shaw, and J. A. Gentry, "A classification approach using multi-layered neural networks," *Decision Support Systems*, vol. 11, no. 5, pp. 509–525, 1994.
- [24] S. Gao, M. Zhou, Y. Wang, J. Cheng, H. Yachi, and J. Wang, "Dendritic neuron model with effective learning algorithms for classification, approximation, and prediction," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 2, pp. 601–614, 2019.
- [25] R. R. Trippi and E. Turban, *Neural Networks in Finance and Investing: Using Artificial Intelligence to Improve Real World Performance*, McGraw-Hill, Inc, 1992.
- [26] T. P. Vogl, J. K. Mangis, A. K. Rigler, W. T. Zink, and D. L. Alkon, "Accelerating the convergence of the back-propagation method," *Biological Cybernetics*, vol. 59, no. 4-5, pp. 257–263, 1988.
- [27] S. Mirjalili, S. M. Mirjalili, and A. Lewis, "Let a biogeography-based optimizer train your multi-layer perceptron," *Information Sciences*, vol. 269, pp. 188–209, 2014.
- [28] M. Gori and A. Tesi, "On the problem of local minima in backpropagation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, no. 1, pp. 76–86, 1992.
- [29] Y. Lee, S. Oh, and M. W. Kim, "An analysis of premature saturation in back propagation learning," *Neural Networks*, vol. 6, no. 5, pp. 719–728, 1993.
- [30] M. S. Hung and J. W. Denton, "Training neural networks with the GRG2 nonlinear optimizer," *European Journal of Operational Research*, vol. 69, no. 1, pp. 83–91, 1993.
- [31] M. K. Weir, "A method for self-determination of adaptive learning rates in back propagation," *Neural Networks*, vol. 4, no. 3, pp. 371–379, 1991.
- [32] A. Van Ooyen and B. Nienhuis, "Improving the convergence of the back-propagation algorithm," *Neural Networks*, vol. 5, no. 3, pp. 465–471, 1992.
- [33] J. Q. Zhang and A. C. Sanderson, "JADE: adaptive differential evolution with optional external archive," *IEEE Transactions on Evolutionary Computation*, vol. 13, no. 5, pp. 945–958, 2009.
- [34] S. Das and P. N. Suganthan, "Differential evolution: a survey of the state-of-the-art," *IEEE Transactions on Evolutionary Computation*, vol. 15, no. 1, pp. 4–31, 2011.
- [35] M. Morey, S. K. Yee, T. Herman, A. Nern, E. Blanco, and S. L. Zipursky, "Coordinate control of synaptic-layer specificity and rhodopsins in photoreceptor neurons," *Nature*, vol. 456, no. 7223, pp. 795–799, 2008.
- [36] C. Koch, *Biophysics of Computation: Information Processing in Single Neurons*, Oxford University Press, 2004.
- [37] E. Salinas and L. F. Abbott, "A model of multiplicative neural responses in parietal cortex," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 93, no. 21, pp. 11956–11961, 1996.
- [38] F. Gabbiani, H. G. Krapp, C. Koch, and G. Laurent, "Multiplicative computation in a visual neuron sensitive to looming," *Nature*, vol. 420, no. 6913, pp. 320–324, 2002.
- [39] J. Sietsma and R. J. Dow, "Neural net pruning-why and how," in *Proceedings of 1993 IEEE International Conference on Neural Networks (ICNN '93)*, vol. 1, pp. 325–333, IEEE, San Diego, CA, USA, 1988.
- [40] K. Nagaraj and A. Sridhar, "A predictive system for detection of bankruptcy using machine learning techniques," <https://arxiv.org/abs/1502.03601>.
- [41] E. K. Kornoushenko, "Classification algorithm based on pairwise comparison of features," *Automation and Remote Control*, vol. 78, no. 11, pp. 2062–2074, 2017.
- [42] J. Uthayakumar, T. Vengattaraman, and P. Dhavachelvan, "Swarm intelligence based classification rule induction (cri) framework for qualitative and quantitative approach: An application of bankruptcy prediction and credit risk analysis," *Journal of King Saud University-Computer and Information Sciences*, 2017.
- [43] J. Uthayakumar, N. Metawa, K. Shankar, and S. K. Lakshmanaprabu, "Intelligent hybrid model for financial crisis prediction using machine learning techniques," *Journal of Information Systems and e-Business Management*, pp. 1–29, 2018.
- [44] Y. Tan, P. P. Shenoy, M. W. Chan, and P. M. Romberg, "On construction of hybrid logistic regression-naive bayes model for classification," in *Proceedings of the Conference on Probabilistic Graphical Models*, pp. 523–534, 2016.
- [45] N. Wang, "Bankruptcy prediction using machine learning," *Journal of Mathematical Finance*, vol. 7, no. 04, p. 908, 2017.

- [46] R. Storn and K. Price, "Differential evolution—a simple and efficient heuristic for global optimization over continuous spaces," *Journal of Global Optimization*, vol. 11, no. 4, pp. 341–359, 1997.
- [47] Q. Lin, S. Liu, Q. Zhu et al., "Particle swarm optimization with a balanceable fitness estimation for many-objective optimization problems," *IEEE Transactions on Evolutionary Computation*, vol. 22, no. 1, pp. 32–46, 2018.
- [48] Q. Lin, S. Liu, K. Wong et al., "A clustering-based evolutionary algorithm for many-objective optimization problems," *IEEE Transactions on Evolutionary Computation*, vol. 23, no. 3, pp. 391–405, 2019.
- [49] A. E. Eiben and S. K. Smit, "Parameter tuning for configuring and analyzing evolutionary algorithms," *Swarm and Evolutionary Computation*, vol. 1, no. 1, pp. 19–31, 2011.
- [50] A. Zamuda and J. Brest, "Self-adaptive control parameters? randomization frequency and propagations in differential evolution," *Swarm and Evolutionary Computation*, vol. 25, pp. 72–99, 2015.
- [51] V. García, A. I. Marqués, and J. S. Sánchez, "An insight into the experimental design for credit risk and corporate bankruptcy prediction systems," *Journal of Intelligent Information Systems*, vol. 44, no. 1, pp. 159–189, 2015.
- [52] C.-M. Wang and Y.-F. Huang, "Evolutionary-based feature selection approaches with new criteria for data mining: A case study of credit approval data," *Expert Systems with Applications*, vol. 36, no. 3, pp. 5900–5908, 2009.
- [53] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.
- [54] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," *Journal of Bioinformatics and Computational Biology*, vol. 3, no. 2, pp. 185–205, 2005.
- [55] R. Jugulum, S. Taguchi et al., *Computer-Based Robust Engineering: Essentials for DFSS*, ASQ Quality Press, 2004.
- [56] Z. Beheshti, S. M. H. Shamsuddin, E. Beheshti, and S. S. Yuhaniz, "Enhancement of artificial neural network learning using centripetal accelerated particle swarm optimization for medical diseases diagnosis," *Soft Computing*, vol. 18, no. 11, pp. 2253–2270, 2013.
- [57] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, 2016.
- [58] D. J. Hand, "Assessing the performance of classification methods," *International Statistical Review*, vol. 80, no. 3, pp. 400–414, 2012.
- [59] N. M. Kiefer, "Default estimation for low-default portfolios," *Journal of Empirical Finance*, vol. 16, no. 1, pp. 164–173, 2009.
- [60] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945.
- [61] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.
- [62] T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967.
- [63] Y. Lu, N. Sundararajan, and P. Saratchandran, "Performance evaluation of a sequential minimal Radial Basis Function (RBF) neural network learning algorithm," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 9, no. 2, pp. 308–318, 1998.
- [64] V. Svetnik, A. Liaw, C. Tong, J. Christopher Culberson, R. P. Sheridan, and B. P. Feuston, "Random forest: a classification and regression tool for compound classification and QSAR modeling," *Journal of Chemical Information and Computer Sciences*, vol. 43, no. 6, pp. 1947–1958, 2003.
- [65] S. R. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 21, no. 3, pp. 660–674, 1991.
- [66] M. M. Adankon and M. Cheriet, "Support vector machine," *Encyclopedia of Biometrics*, pp. 1303–1308, 2009.
- [67] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K.-R. Muller, "Fisher discriminant analysis with kernels," in *Proceedings of the 9th IEEE Signal Processing Society Workshop on Neural Networks for Signal Processing (NNSP '99)*, pp. 41–48, Madison, Wis, USA, August 1999.

Research Article

Analysis of Financing Efficiency of Chinese Agricultural Listed Companies Based on Machine Learning

Lixia Liu ¹ and Xueli Zhan ²

¹School of Economics, Tianjin University of Commerce, Tianjin 300134, China

²School of Economics, Beijing Wuzi University, Beijing 101149, China

Correspondence should be addressed to Lixia Liu; liulixia77@163.com and Xueli Zhan; xuelz20163205@126.com

Received 4 April 2019; Revised 14 June 2019; Accepted 24 June 2019; Published 10 July 2019

Guest Editor: Benjamin M. Tabak

Copyright © 2019 Lixia Liu and Xueli Zhan. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Agricultural enterprises play a significant role in China's economic development. However, compared with other enterprises, agricultural enterprises are facing serious financial problems. Financing difficulty is essentially a question of financing efficiency. Based on the DEA method, this paper evaluates the financing efficiency of 39 agricultural listed companies in China from 2013 to 2017. The results suggest that the financing efficiency is generally low, and the Total Factor Productivity of agricultural enterprises' financing has a tendency to decrease first and then increase. The influencing factors of financing efficiency are analyzed using the Tobit regression model and the random forest regression model. And we find the following: (1) The random forest regression model significantly outperformed the Tobit regression model, with determination coefficients (R^2) greater than 0.9 in full sample sets. (2) Total liability, financial expenses, return on total assets, and inventory turnover rate are important factors affecting financing efficiency of agricultural listed companies. (3) Return on total assets and inventory turnover rate promote the financing efficiency, while total liability and financial expenses reduce financing efficiency. Finally, the paper makes some suggestions for the financing of agricultural enterprises.

1. Introduction

Agriculture not only provides us with the food and clothing, but also provides us with energy and chemical raw materials needed for industrial development. It is a basic industry related to economic development and social stability. Agricultural enterprises are the most important organizations in modern agricultural industrial system and the important bridge connecting farmers and the market. Agricultural enterprises are more difficult to operate than other enterprises, particularly in developing countries such as China [1]. They are not only affected by social factors, but also affected by natural factors, especially the weather. Under the influence of severe weather, the agricultural enterprises may be subjected to uncontrollable factors, which can increase the risk of corporate failure and default [2]. So, agricultural enterprises often face more severe financing problems [3]. Public sector funding is widely believed to be a more effective measure for agricultural progress [4]. However, government

funds are often limited. It is essential to enhance the external financing capacity and financing efficiency of agricultural enterprises. The research on agricultural financing mainly focuses on financing structure, financing mode, and agricultural financial policy [5–12]. Abate et al. [6] analyzed the impact of institutional finance on agricultural technology adoption in Ethiopia, and the results showed that the access to institutional finance had a significant positive impact on farmers' adoption of agricultural technology. However, few scholars pay attention to the financing efficiency of agricultural enterprises.

Financing efficiency is a key index to estimate an enterprises' efficiency of using their funds. From the literature, we note that the research on the enterprises' financing efficiency can be divided into three perspectives including regions [13, 14], industries [15–17], and capital market [18, 19]. Geng et al. [13] evaluated the financing efficiency of listed companies of machinery manufacturing industry in Jiangsu based on the Malmquist index model. Ma et al. [17] analyzed the

financing efficiency of 21 listed companies in LED lighting industry in photovoltaic industry, and the results indicated that the financing efficiency showed an upward trend, but the overall level was low. Dong et al. [18] analyzed the financing efficiency of 300 listed companies in Shanghai and Shenzhen Stock Exchange from 2008 to 2014, and the results showed that the financing efficiency of Chinese listed companies was generally low. Data envelopment analysis (DEA) first proposed by Charnes et al. is a common method used to evaluate financing efficiency [20]. Compared with other methods, the DEA method has many advantages: there is no need to estimate the production function, it is capable of handling multiple inputs and outputs, and it is capable of analyzing the reasons for the inefficiency of each evaluation unit. Prior studies also investigated the impact of internal and external factors on the enterprises' financing efficiency. The internal factors mainly include capital structure, financing cost, financing mode, property nature, firm age, and firm size [21–25]. The external factors mainly include macroeconomic situation, financial development, external financial support, legal environment, market competition, and interfirm trust [26–29]. The linear regression model has been the most commonly used method in the analysis of influencing factors of financing efficiency. With the development of computer technology, the application of machine learning and game theory in the economic field has gradually increased, but research in the field of corporate finance is still rare [30–33].

This paper selects 39 agricultural listed companies in Shanghai Stock Exchange and Shenzhen Stock Exchange from 2013 to 2017, evaluates financing efficiency of Chinese agricultural listed companies with DEA model, and explores the impact of internal and external factors on financing efficiency. We contribute to the existing literature on enterprises' financing efficiency in three respects. First, we focus purely on agricultural enterprises and hope this research can help to improve the overall level of agricultural enterprises' financing efficiency. Numerous researches have focused on the financing efficiency of regions, industries, and capital market. So far, there are relatively few studies on the financing efficiency of agricultural enterprises [29]. Second, we calculate the financing efficiency of agricultural listed companies in China's Stock Market. China's agricultural enterprises are a significant case study for our purposes. China is a big agricultural country with abundant agricultural resources, a long history of agriculture, and a huge rural population. Now, more than 20% of the Chinese population still lives on farms. In 2016, the number of agricultural industrialization organizations in China had reached 417,000, an increase of 8.01% over 2015. Agricultural industrialization is the development direction of China's agriculture, and the development of agricultural enterprises is related to the long-term development of China's agriculture. Our third contribution is methodological. In recent years, the methods such as game theory and machine learning have been applied more and more in the field of economics, but few people apply them to the analysis of financing efficiency [34–37]. Random forest is an ensemble machine learning methods of classification and regression proposed by Leo Breiman [38]. It has proven to be an effective analytical tool for studying

the relationship between predictors and response because of its excellence in interpretation, visualization, and abilities to handle complex nonlinearity [39–41]. The random forest regression model is used to explore the impact of internal and external factors on financing efficiency, and the results are compared with those of econometric regression analysis. The paper not only provides examples of application of machine learning methods on the research field of financing efficiency, but also has practical significance for empirical analysis on the financing efficiency of Chinese agricultural listed companies.

The rest of this paper is organized as follows. Section 2 introduces the models used throughout this paper. Section 3 describes the key variables and the data source. Section 4 provides the empirical results and discussion, which include the evaluation of financing efficiency of agricultural enterprises and study of its antecedents. The conclusion and policy suggestions are given in Section 5.

2. Methods

2.1. DEA Model. Data Envelopment Analysis (DEA) introduced by Charnes et al. [20] is a nonparametric method to measure relative efficiency of the analyzed objects with multiple inputs and multiple outputs. Different from other measuring efficiency methods, DEA model treats the DMU as a "black box." We don't need to determine the functional relationship between input and output metrics before using the DEA model. The method introduces linear programming to construct nonparametric piecewise surfaces of observed data and then computes efficiency relative to this frontier.

According to these assumptions, DEA model can be divided into two categories: constant return to scale (CRS) and the variable return to scale (VRS). VRS is an improvement to the CRS model, which is used to explain the variable scale income. When the enterprise is not satisfied with the optimal scale operation, VRS can avoid the confusion between the measurement result of the technical efficiency and scale efficiency. Obviously, we should use the VRS model to study agricultural enterprises' financing efficiency.

Suppose that there are I decision making units (DMUs), and each decision making unit has N inputs and M outputs. Let $X_i = (x_{1i}, x_{2i}, \dots, x_{Ni})^T$ and $Y_i = (y_{1i}, y_{2i}, \dots, y_{Mi})^T$ be the input vector and the output vector of DMUs i , respectively. The $N \times I$ input matrix and the $M \times I$ output matrix Y represent the data of all I decision making units. The DEA model can be shown as follows:

$$\begin{aligned} \min \quad & \theta \\ \text{s.t.} \quad & \sum_{i=1}^n \lambda_i x_i - \theta x_0 \leq 0 \\ & \sum_{i=1}^n \lambda_i y_i - x_0 \geq 0 \\ & \lambda_i \geq 0 \end{aligned} \tag{1}$$

where θ denotes the efficiency score of DMU i and λ denotes the weight of DMU i . When the decision unit θ is equal to 1,

the DMU is efficient; i.e., its inputs and outputs have reached optimal combination in the production system.

2.2. Malmquist Index Model. Malmquist [41] firstly proposed the Malmquist index and used this method to analyze the consumption behavior. Based on Malmquist's work, Caves et al. [42] put forward the Malmquist productivity index in 1982. The Malmquist productivity index is an effective method of measuring Total Factor Productivity (TFP). The Malmquist productivity index from t period to $t+1$ can be written as follows:

$$M(x^t, q^t, x^{t+1}, q^{t+1}) = \left(\frac{D^t(x^{t+1}, q^{t+1})}{D^t(x^t, q^t)} \times \frac{D^{t+1}(x^{t+1}, q^{t+1})}{D^{t+1}(x^t, q^t)} \right)^{1/2} \quad (2)$$

where x^t, q^t are the input and output vector of period t , respectively. $D^t(x^t, q^t)$ and $D^{t+1}(x^t, q^t)$ denote the distance function of the DMU of period t and $t+1$ when the period t is taken as reference.

Färe et al. [43] improved the model and decomposed Total Factor Productivity (TFP) into efficiency change (EC) and technical change (TC). The formulas are stated as follows:

$$M(x^t, q^t, x^{t+1}, q^{t+1}) = \frac{D^{t+1}(x^{t+1}, q^{t+1})}{D^t(x^t, q^t)} \times \left(\frac{D^t(x^{t+1}, q^{t+1})}{D^{t+1}(x^{t+1}, q^{t+1})} \times \frac{D^t(x^t, q^t)}{D^{t+1}(x^t, q^t)} \right)^{1/2} \quad (3)$$

$$EC = \frac{D^{t+1}(x^{t+1}, q^{t+1})}{D^t(x^t, q^t)} \quad (4)$$

$$TC = \left(\frac{D^t(x^{t+1}, q^{t+1})}{D^{t+1}(x^{t+1}, q^{t+1})} \times \frac{D^t(x^t, q^t)}{D^{t+1}(x^t, q^t)} \right)^{1/2} \quad (5)$$

The rate of technical change (TC) can be divided into pure technical efficiency change (PTEC) and scale efficiency change (SEC). The formulas are shown as follows:

$$PTEC = \left(\frac{D^t(x^{t+1}, q^{t+1})}{D^{t+1}(x^{t+1}, q^{t+1})} \right)^{1/2} \quad (6)$$

$$SEC = \left(\frac{D^t(x^t, q^t)}{D^{t+1}(x^t, q^t)} \right)^{1/2} \quad (7)$$

2.3. Tobit Regression Model. The value of financing efficiency is between 0 and 1, which is the truncated data. When we use financing efficiency as a dependent variable to analyze the effect of various factors on financing efficiency, there may be biased and inconsistent estimating results by ordinary linear regression. Tobit regression model, also called a censored regression model, is designed to estimate linear relationships between variables when there is either left or right censoring

in the dependent variable [44]. So we can use this method to resolve the above problems. The model is shown as follows:

$$y_i = \begin{cases} 0 & \text{if } y_i^* \leq 0 \\ \beta x_i' + \varepsilon_i & \text{if } 0 < y_i^* < c \\ c & \text{if } y_i^* \geq c \end{cases} \quad (8)$$

where y_i is the dependent variable, y_i^* is the latent variable, x_i' is the independent variables, β is the parameter vector, and $\varepsilon_i \sim N(0, \sigma^2)$ is a random perturbation.

2.4. Random Forest Regression Model. Random forest is a wonderful machine learning approach which is used for classification and regression as an ensemble learning [38]. It contains several decision trees trained by bootstrap resampling method. When a sample to be regressed is entered, the final regression result is determined by the vote of the output of these decision trees. Random forest overcomes the problem of overfitting and has good tolerance to noise and anomaly values. It is a fully nonparametric statistical method that optimizes predictive accuracy by fitting an ensemble of trees to stabilize model estimates.

The steps to generate a random forest can be represented as follows:

- (1) The bootstrap resampling method is applied to randomly extract k samples from the original training sets, and then n regression trees are generated.
- (2) For each of the bootstrap samples, an unpruned regression tree is grown. At each node, m of the predictors are chosen randomly and the best split is chosen among those predictors.
- (3) Predict new data by aggregating the predictions of the n trees (i.e., average for regression).

The mean square error (MSE) and the decision coefficient (R^2) are used as criteria for evaluating the model error. The calculation formulas are as follows:

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - y_i')^2 \quad (9)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - y_i')^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (10)$$

where y_i is the actual value of dependent variable, y_i' is the predictive value of dependent variable, and \bar{y} is the mean value of the dependent variable.

3. Indicator Selection and Data Sources

3.1. Financing Efficiency Evaluation Variables. The selection of optimum indicators is the hinge of analyzing the financing efficiency of agricultural listed companies using the DEA model. In this paper, we select total assets, operating cost, and equity ratio as the input indicators and select asset turnover ratio, earnings per share, and rate of return on common stockholders' equity (ROE) as output indicators.

- (1) Total assets (X_1). The total assets are an indicator reflecting the financing ability of the enterprise. It's generally believed that the larger the total assets of the enterprise are, the larger the scale of the enterprise is, and the stronger the financing ability of the enterprise is.
- (2) Operating cost (X_2). Operating cost refers to the costs associated with a company's main operating activities. The indicator can be used to indicate the use of corporate funds. As a rule, the higher the operating cost of the enterprise is, the higher the capital use cost of enterprise is.
- (3) Equity ratio (X_3). Equity ratio is the ratio of the total liabilities to owner's equity. It is an important indicator for evaluating the rationality of capital structure. The lower the enterprise's equity ratio is, the stronger its long-term solvency is.
- (4) Asset turnover ratio (Y_1). Asset turnover measures the efficiency ratio with which a company uses its assets to generate sales. It can be used as an indicator for evaluating the management quality and utilization efficiency of enterprise assets.
- (5) Earnings per share (Y_2). Earnings per share are a financial ratio, which measures net earnings earned per share of stock outstanding. The larger the earnings per share are, the stronger the enterprise's equity financing ability is.
- (6) Rate of return on common stockholders' equity (Y_3). Rate of return on common stockholders' equity is computed by dividing net income after interest, taxes, and preferred dividends to average common stockholders' equity. The higher the rate of return on common stockholders' equity ratio is, the higher the return of investors is, and the stronger the enterprise's profitability is. This indicator can be used to reflect the efficiency of the enterprise in using its own capital.

The values of the selected indicators are supposed to be positive in the DEA model, but some values selected in this paper are negative, so the data needs to be dimensionless. The approach is listed as follows:

$$Z_i' = 0.9 * \frac{(Z_i - \min(Z_i))}{(\max(Z_i) - \min(Z_i))} + 0.1 \quad (11)$$

where $i = 1, 2 \dots, 6$, $\min(Z_i)$ and $\max(Z_i)$ are the minimum and maximum values of each variable, respectively.

3.2. Regression Variables. The multiple linear regression models are established in (12). In this paper, the dependent variable is financing efficiency calculated by the DEA model. Total liability, financial expense, return on total assets, inventory turnover rate, price index of agricultural means of production, and gross domestic product (GDP) are selected to investigate the effect of these factors on the financing efficiency of agricultural enterprises.

- (1) Total liability (TL). Total liability refers to the aggregate debt for which agricultural enterprises are liable. Debt management is the most important means of agricultural enterprises' operation. It can alleviate agricultural enterprises' financing difficulties, expand their production scale, improve their market competitiveness, and promote their rapid development, while, when the total liabilities are too high, agricultural enterprise will face more financial risk.
- (2) Financial expense (FE). Financial expenses refer to the expenses incurred by the enterprise in order to raise the funds needed for its operation. It is an indicator used to reflect the cost of enterprises to raise funds. Generally, the higher the financial expenses are, the higher the enterprises' financing cost is.
- (3) Return on total assets (RT). Profitability is a measure of the enterprise's ability to pay off debts. Strong profitability means that the enterprises can get better returns and be able to repay their debts on time [45]. Thus, profitability is an indirect factor affecting the financing efficiency of enterprise. The indicators reflecting enterprises' profitability include gross profit margin, net profit margin, return on net assets, return on total assets, and earnings per share. Return on total assets is an important indicator of listed companies, which can reflect the efficiency of enterprises' asset operation and evaluate the ability of enterprise to manage assets.
- (4) Inventory turnover rate (ITR). The indicator of turnover rate is usually used to indicate the operating efficiency of enterprise, including accounts receivable turnover rate, inventory turnover rate, current assets turnover rate, fixed asset turnover rate, and total asset turnover rate. The inventory turnover rate is an important indicator for measuring how efficiently a firm turns its inventory into sales. Generally, the higher the inventory turnover rate is, the lower the inventory occupancy level is, and the stronger the liquidity is, which will enhance the short-term solvency and profitability of the enterprise.
- (5) Price index of agricultural means of production (PI). The price index of agricultural means of production measures changes in the price level of agricultural production materials. The agricultural production means mainly include agricultural hand tools, feed, animal products, semimechanized farm tools, mechanized farm tools, and so on. The higher the price of agricultural means of production is, the higher the market demands for agricultural enterprises' products are, and the higher the profitability of agricultural enterprises is.
- (6) Gross domestic product (GDP). GDP is the total value of all the final goods and services produced within a country's borders in a specific time period. It is often used as an indicator for measuring the economic situation of a country.

To satisfy the requirement of stationarity, the explanatory variables of total liability, financial expenses, and GDP are logarithmically transformed. The factors are standardized by taking natural logarithms. The model of the effect of influence factors on the financing efficiency of agricultural enterprises is as follows:

$$TE = \beta_0 + \beta_1 \ln(TL) + \beta_2 \ln(FE) + \beta_3 RT + \beta_4 ITR + \beta_5 PI + \beta_6 \ln(GDP) + \varepsilon \quad (12)$$

where TE is the values of comprehensive technical efficiency (TE) of agricultural enterprises' financing calculated by DEA model, β_0 denotes the intercept term, $\beta_1, \beta_2, \beta_3, \dots, \beta_6$ represent the regression coefficients of variables, and ε is the residual term of the regression model. Since financial expenses involve negative numbers, in order to facilitate the logarithm, financial expenses are translated as follows:

$$FE_i = Fe_i + \min(|Fe|) \quad (13)$$

where Fe_i denotes the original value of financial expenses, and FE_i represents the translated values of financial expenses.

3.3. Data Sources. We select agricultural listed companies in China from 2013 to 2017. In the selection process, the enterprises that have been given special treatment by the Shenzhen Stock Exchange (SZSE) and the Shanghai Stock Exchange (SSE) or lack the selected variables values are excluded. Finally, we choose 39 enterprises as our sample. The information of 39 agricultural listed companies is shown in Table 1. The data are mostly derived from Wind Financial Terminal (<http://www.eastmoney.com>) and China Statistical Yearbook.

4. Empirical Analysis

4.1. Descriptive Statistics. Before analyzing the financing efficiency of agricultural listed companies, descriptive statistics of the relevant variables will be discussed. Table 2 presents descriptive statistics regarding of all agricultural listed companies and macroeconomic indicators.

4.2. Measurement of Financing Efficiency Based on DEA Model. The financing efficiency of agricultural listed companies in China from 2013 to 2017 is measured by using DEA model. The results are shown in Table 3. We can see that the financing efficiency of agricultural enterprises in China is low in general. Comprehensive technical efficiency (TE), pure technical efficiency (PE), and scale efficiency (SE) show a significant downward trend during 2013-2016. Financing efficiency decreased from a relatively high base of 0.754 in 2013 to 0.661 in 2016. And due to the increase in pure technical efficiency, the financing efficiency raised from 0.661 in 2016 to 0.730 in 2017. In the period between 2013 and 2017, the number of financing efficient enterprises is 11, 8, 7, 5, and 6, respectively. The proportion of financing efficient enterprises is 28.21%, 20.51%, 17.95%, 12.82%, and 15.38%, respectively, suggesting that more than 70% of agricultural enterprises are at a very low level of financing efficiency. From the

TABLE 1: 39 agricultural listed companies.

No	Stock code	Stock market
1	000592	Shenzhen Stock Exchange
2	000735	Shenzhen Stock Exchange
3	000798	Shenzhen Stock Exchange
4	000998	Shenzhen Stock Exchange
5	002041	Shenzhen Stock Exchange
6	002069	Shenzhen Stock Exchange
7	002086	Shenzhen Stock Exchange
8	002200	Shenzhen Stock Exchange
9	002234	Shenzhen Stock Exchange
10	002299	Shenzhen Stock Exchange
11	002321	Shenzhen Stock Exchange
12	002458	Shenzhen Stock Exchange
13	002477	Shenzhen Stock Exchange
14	002679	Shenzhen Stock Exchange
15	002696	Shenzhen Stock Exchange
16	002714	Shenzhen Stock Exchange
17	002746	Shenzhen Stock Exchange
18	002772	Shenzhen Stock Exchange
19	200992	Shenzhen Stock Exchange
20	300087	Shenzhen Stock Exchange
21	300094	Shenzhen Stock Exchange
22	300106	Shenzhen Stock Exchange
23	300189	Shenzhen Stock Exchange
24	300511	Shenzhen Stock Exchange
25	300761	Shenzhen Stock Exchange
26	600097	Shanghai Stock Exchange
27	600108	Shanghai Stock Exchange
28	600257	Shanghai Stock Exchange
29	600313	Shanghai Stock Exchange
30	600354	Shanghai Stock Exchange
31	600359	Shanghai Stock Exchange
32	600371	Shanghai Stock Exchange
33	600467	Shanghai Stock Exchange
34	600506	Shanghai Stock Exchange
35	600540	Shanghai Stock Exchange
36	600598	Shanghai Stock Exchange
37	600965	Shanghai Stock Exchange
38	600975	Shanghai Stock Exchange
39	601118	Shanghai Stock Exchange

distribution of the financing efficiency, both scale efficiency and pure technical efficiency are less than 0.9, which are the main reasons for the low financing efficiency.

4.3. Measurement of Total Factor Productivity Based on Malmquist Index. We analyze the financing efficiency of agricultural enterprises with Malmquist index model. The results shown in Table 4 indicate that the Malmquist indices of the first three periods were 0.984, 0.998, and 0.824, respectively, and exhibit a downward trend. And due to the increase in pure technical efficiency, the Malmquist

TABLE 2: Descriptive statistics.

Variables		Cases	Mean	S.D.	Max	Min
Financing efficiency evaluation variables	X_1	195	3.84	3.73	24	0.288
	X_2	195	1.58	1.97	11.1	0.044
	X_3	195	1.052	1.049	8.782	0.052
	Y_1	195	0.562	0.348	1.924	0.083
	Y_2	195	0.150	0.564	2.25	-2.197
	Y_3	195	2.273	18.293	62.715	-101.659
Influencing factors indicators	TL	195	1.774	2.07	16.416	0.015
	FE	195	0.051	0.084	0.788	-0.034
	RT	195	3.358	8.095	31.408	-43.175
	ITR	195	2.979	2.580	12.887	0.042
	PI	195	673.99	2.660	677.791	670.1
	GDP	195	6.962	0.792	8.207	5.930

TABLE 3: Financing efficiency values of agricultural enterprises.

	2013	2014	2015	2016	2017
Comprehensive technical efficiency (TE)	0.754	0.755	0.731	0.661	0.730
Pure technical efficiency (PE)	0.826	0.863	0.869	0.744	0.880
Sale efficiency (SE)	0.896	0.862	0.830	0.881	0.823

TABLE 4: Malmquist index of agricultural listed companies' financing.

Time	TC	EC	PTEC	SEC	TFP
2013-2014	1.013	0.971	1.054	0.961	0.984
2014-2015	0.969	1.03	1.008	0.962	0.998
2015-2016	0.897	0.919	0.846	1.06	0.824
2016-2017	1.12	1.279	1.199	0.934	1.433
mean	0.996	1.041	1.019	0.978	1.038

TABLE 5: Results of Tobit regression analysis.

Variables	Coefficient	Standard error	T-statistic	Prob
TL	-0.152	0.025	-6.14	0.001
FE	-0.057	0.025	-2.25	0.025
RT	0.012	0.002	5.97	0.001
ITR	0.011	0.005	1.94	0.053
PI	0.005	0.003	1.76	0.080
GDP	-0.042	0.097	-0.43	0.669
β_0	2.279	1.746	1.31	0.193

index raised to 1.433 during 2016-2017, and the Total Factor Productivity (TFP) growth rate was 43.3%. In the period between 2013 and 2017, the average TFP of agricultural enterprises' financing was 1.038, which indicates that the TFP increased by an average annual rate of 3.8%. In terms of composition, the average annual growth rate of technical change, efficiency change, pure technical efficiency change, and scale efficiency change was -0.4%, 4.1%, 1.9%, and -2.2%, respectively. The results indicate that the financing efficiency of agricultural enterprises is promoted by efficiency change and pure technical efficiency change and hindered by scale efficiency change.

4.4. Influencing Factors Analysis Based on Tobit Regression. We examine the effect of influencing factors on the financing efficiency of agricultural enterprises using Stata 14 software. The results shown in Table 5 indicate that, in addition to GDP, the other five influencing factors pass the significance test, and the performance of Tobit regression was remarkably good.

Total liability has a significant negative impact on agricultural enterprises' financing efficiency. The indicator of total liability is used to reflect the liability scale of enterprises. A 1% increase in total asset leads to a 0.152% decrease in agricultural enterprises' financing efficiency. This result

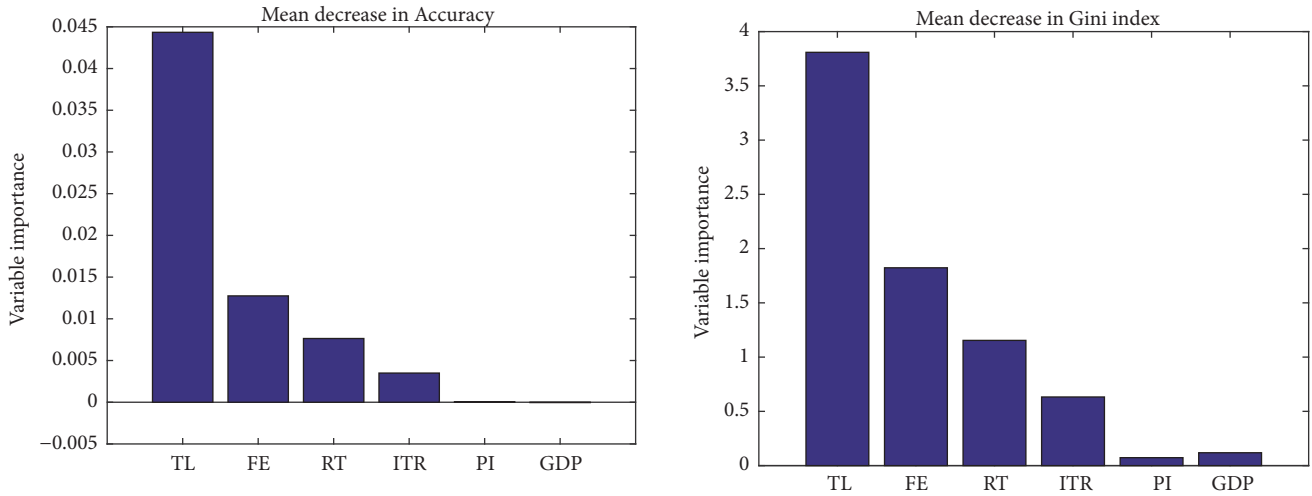


FIGURE 1: The importance ranking of influencing factors.

is consistent with the finding of Pan et al. for China's environmental protection industry [22]. The funds obtained by enterprises through debt need to be repaid, which will reduce the free capital of enterprises. Therefore, excessive debt financing will have a negative impact on the financing efficiency of agricultural enterprises.

Financial expense has a significant negative impact on agricultural enterprises' financing efficiency. A 1% increase in financial expense leads to a 0.057% decrease in agricultural enterprises' financing efficiency. This result is consistent with the finding of Gatti and Love (2010) and Wang and Zhang (2018), who found that higher financial costs lead to the lower financing efficiency of agricultural enterprises [23, 24].

Return on total assets is positively correlated with the financing efficiency of agricultural enterprises. A 1% increase in return on total assets leads to a 0.012% increase in financing efficiency. The result suggests that the stronger the enterprises' capital management ability is, the stronger the enterprises' profitability is, and the higher the financing efficiency of agricultural enterprises is, which confirms the findings of Wu and Zeng for SMEs (2019) [25].

Inventory turnover rate has a significant positive influence on the financing efficiency of agricultural enterprises. A 1% increase in asset-liability ratio leads to a 0.011% increase in financing efficiency. The result suggests that the higher the operating efficiency is, the higher the efficiency of enterprises financing is.

Price index of agricultural means of production has a significant negative impact on agricultural enterprises' financing efficiency. However, compared with other indicators, price index has little effect on the financing efficiency of agricultural enterprises. A 1% increase in price index of agricultural means of production only leads to a 0.005% increase in financing efficiency. This confirms the result of Pan et al. [22]. PI is a measure widely used to track agricultural production materials price inflation. Once inflation occurs, the rise of agricultural products prices will be a fatal blow to agricultural enterprises, which will inevitably affect the financing efficiency of enterprises.

GDP is negatively correlated with the financing efficiency, but it does not pass the significance test. The result suggests that GDP has no relation with the financing efficiency of agricultural enterprises.

On the whole, the negative impact of debt size and financing cost on financial efficiency is far greater than the positive impact of profitability and operating efficiency, while the impact of macroeconomic environment on the financing efficiency of agricultural enterprises is very limited.

4.5. Influencing Factors Analysis Based on Random Forest Regression. The impact of influencing factors on the financing efficiency of agricultural enterprises is also analyzed using random forest regression. Matlab package random forest developed by Abhishek Jaientilal is used in this paper [46]. Firstly, we examine the importance ranking of influencing factors under the random forest approach. The results presented in Figure 1 indicate that the factors order of importance from strong to weak is total liability, financial expense, return on total assets, inventory turnover rate, price index of agricultural means of production, and GDP. Among them, external environmental factors, including price index of agricultural means of production and GDP, have little impact on financing efficiency of agricultural enterprises. This result is in agreement with that of Tobit regression.

Figure 2 presents the results of the impact of each factor on financing efficiency. It can be observed that the impact of total liability (TL), financial expense (FE), and GDP on financing efficiency is declining. When $TL > 4$, $FE > 0.3$, and $GDP > 7.15$, the impact of these three factors in financing efficiency tends to be stable. The impact of return on total assets (RT) on financing efficiency is on the rise. When $RT > 18$, the impact of RT on financing efficiency tends to be stable. The impact of inventory turnover rate (ITR) on financing efficiency shows a downward trend from 0 to 4 and then an upward trend from 4 to 6. When $ITR > 6$, the impact of ITR on financing efficiency tends to be stable. The influence of price

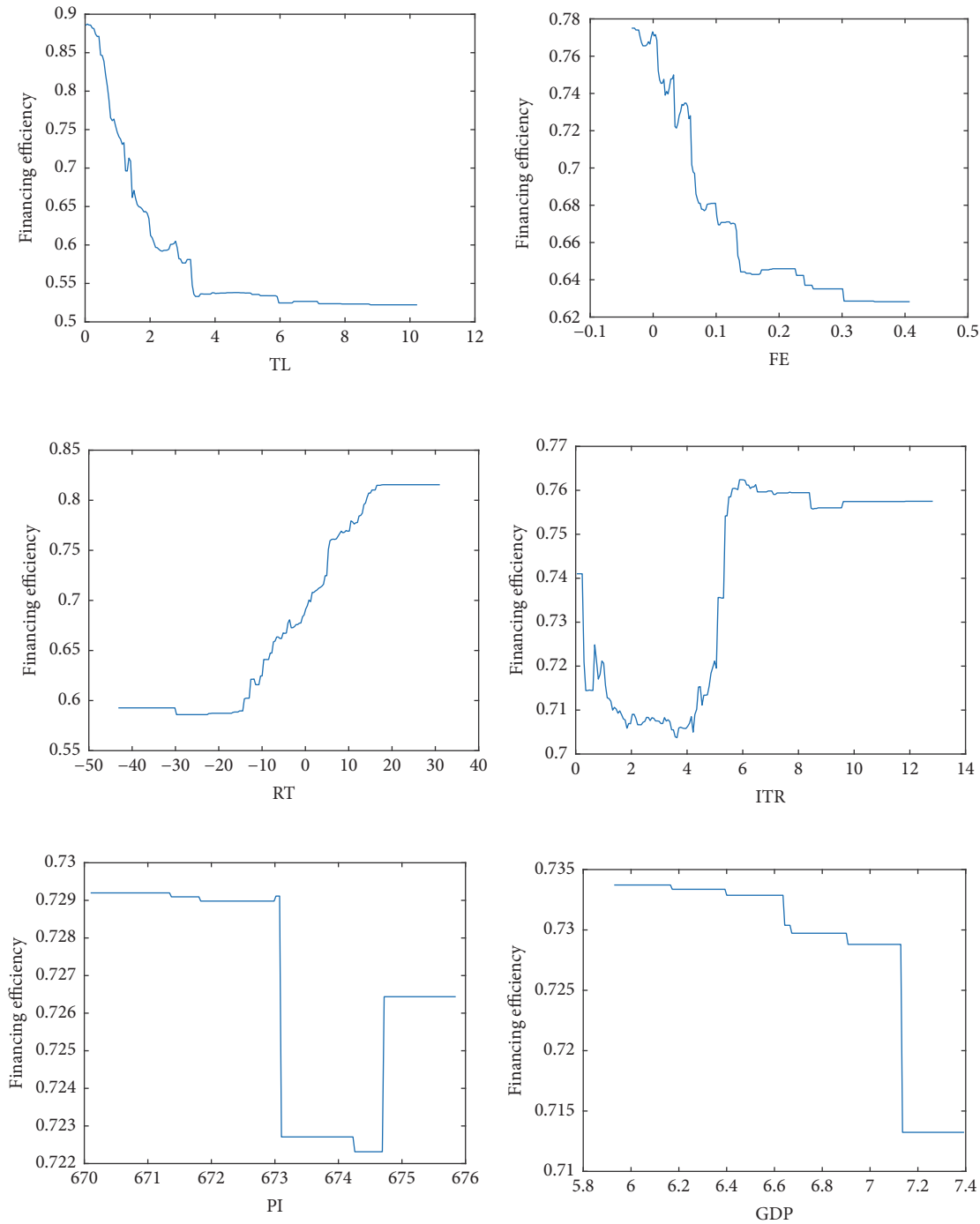


FIGURE 2: The impact of influencing factors on financing efficiency.

index of agricultural means of production (PI) on financing efficiency is U-shaped.

Taking the data of the first four years as training variables and the data of the last one year as testing variables, the empirical analysis is conducted with random forest regression model. Figure 3 shows that the R^2 value in training data and testing data is 0.946 and 0.748, respectively. We also compare the two regression methods. As shown in Figure 4, the R^2

value of the full data set based on random forest regression and Tobit regression is 0.913 and 0.577, respectively. The results suggest that, compared with Tobit regression, the analysis of financing efficiency based on random forest regression has higher R^2 values and better prediction results. Probably, the reason is the inability of the Tobit regression model in capturing the nonlinearity between financing efficiency of agricultural enterprises and its influencing factors.

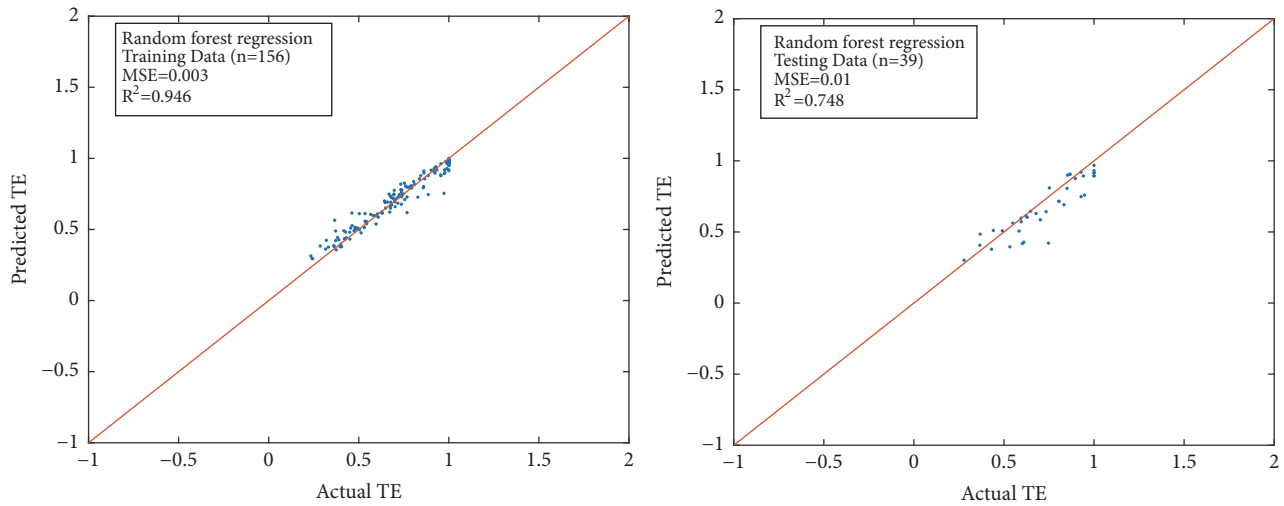


FIGURE 3: Predicted TE using random forest regression.

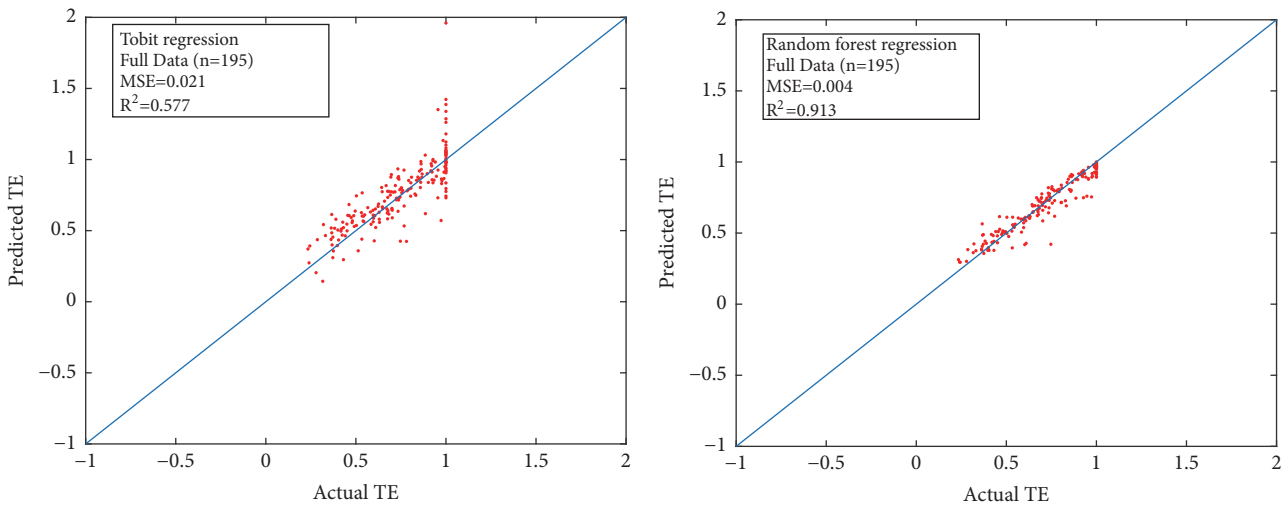


FIGURE 4: Comparison of predicted financing efficiency based on two regression models.

5. Conclusions and Recommendations

Using DEA model, this paper calculates financing efficiency of 39 agricultural listed companies in China from 2013 to 2017. The results reveal that the overall efficiency of financing of agricultural listed companies is low, and less than 30% of agricultural enterprises have achieved DEA effectiveness. Results of Malmquist index analysis indicate that, in the period of 2013-2017, the Total Factor Productivity (TFP) of agricultural enterprises has shown an upward trend due to the increase of efficiency change and pure technical efficiency change. Tobit regression and random forest regression have been applied to the analysis of influencing factors of financing efficiency of agricultural listed companies. The results indicate that random forest regression outperformed Tobit regression in terms of MSE and R^2 . The improvement of return on total assets, inventory turnover rate, and the price index of agricultural means of production promote the increase of agricultural enterprises' financing efficiency.

However, the significant increase in total liability and the expenditure of financial expenses is the main reasons for low financing efficiency of agricultural listed companies. In order to improve the financing efficiency of agricultural enterprises in China, the following suggestions are put forward.

- (1) Improving the capital management ability of agricultural listed companies. Profitability is considered as a major factor for enhancing the enterprise's financing efficiency. Therefore, the enterprises should increase their project identification capabilities, invest their funds to agricultural projects with high returns, and strengthen their capital utilization efficiency. At the same time, the enterprises should establish an effective internal management system, effectively manage financial risks, and reduce unnecessary financial expenses.
- (2) Expanding agricultural listed companies' financing channels and optimizing the financing structure.

Compared with other enterprises, agricultural enterprises are usually small in scale, weak in economic strength, vulnerable to the natural environment and social economy, and slow for the accumulation of funds. These characteristics result in relatively slow accumulation of internal funds and limited scale of external financing of agricultural enterprises. The enterprises should not be confined to bank loans, but take full advantage of various short-term and long-term financing resources such as microfinance, financial leasing, factoring, and bill discounting, and appropriately control the scale of debt financing of agricultural enterprises.

- (3) Improving the government's ability to provide financing services for agricultural enterprises. The government should increase its support for the financing of agricultural listed companies and improve the policies for agricultural enterprise development from the aspects of market access, fair competition, and policy incentives. It ought to encourage innovation in financial products and businesses and build bridges between agricultural enterprises and financial institutions, so that the agricultural enterprises can find low cost funds. And the government also should strengthen regulation of listed companies, guard against illegal and irregular acts in the process of enterprises financing, protect the legitimate rights and interests of investors, and create a favorable financing environment for agricultural enterprises.

Data Availability

The data in this paper mainly come from the data of Listed Companies in China, which has been explained in Section 3.3 of the paper.

Conflicts of Interest

The authors declare no conflict of interest regarding the publication of this paper.

Acknowledgments

This paper is supported by the Tianjin Planning Leading Group Office of Philosophy and Social Sciences under Grant Number TJYY17-017.

References

- [1] R. Holst, X. Yu, and C. Grün, "Climate change, risk and grain yields in China," *Journal of Integrative Agriculture*, vol. 12, no. 7, pp. 1279–1291, 2013.
- [2] H. Fu, J. Li, Y. Li et al., "Risk transfer mechanism for agricultural products supply chain based on weather index insurance," *Complexity*, vol. 2018, Article ID 2369423, 17 pages, 2018.
- [3] J. Huang and Y. Wang, "Financing sustainable agriculture under climate change," *Journal of Integrative Agriculture*, vol. 13, no. 4, pp. 698–712, 2014.
- [4] J. Beynon, "The state's role in financing agricultural research," *Food Policy*, vol. 20, no. 6, pp. 545–550, 1995.
- [5] E. Gelb and Y. Kislev, "Farmers' financing of agricultural research in Israel," *Research Policy*, vol. 11, no. 5, pp. 300–327, 1982.
- [6] G. T. Abate, S. Rashid, C. Borzaga, and K. Getnet, "Rural finance and agricultural technology adoption in Ethiopia: does the institutional design of lending organizations matter?" *World Development*, vol. 84, pp. 235–253, 2016.
- [7] A. Duncan, "Financing agricultural services in sub-Saharan Africa," *Food Policy*, vol. 18, no. 6, pp. 463–465, 1993.
- [8] A. Magnan, "The financialization of agri-food in Canada and Australia: Corporate farmland and farm ownership in the grains and oilseed sector," *Journal of Rural Studies*, vol. 41, pp. 1–12, 2015.
- [9] P. Maitra, S. Mitra, D. Mookherjee, A. Motta, and S. Visaria, "Financing smallholder agriculture: An experiment with agent-intermediated microloans in India," *Journal of Development Economics*, vol. 127, pp. 306–337, 2017.
- [10] M. van Bergen, M. Steeman, M. Reindorp, and L. Gelsomino, "Supply chain finance schemes in the procurement of agricultural products," *Journal of Purchasing and Supply Management*, vol. 25, no. 2, pp. 172–184, 2019.
- [11] J. F. Swinnen and H. R. Gow, "Agricultural credit problems and policies during the transition to a market economy in Central and Eastern Europe," *Food Policy*, vol. 24, no. 1, pp. 21–47, 1999.
- [12] P. Newton, A. E. Gomez, S. Jung et al., "Overcoming barriers to low carbon agriculture and forest restoration in Brazil: the Rural Sustentável project," *World Development Perspectives*, vol. 4, pp. 5–7, 2016.
- [13] C. Geng, H. E. and L. Wang, "Research on the financing efficiency of listed machinery manufacturing companies in Jiangsu province based on the malmquist index model," *International Journal of Arts & Sciences*, vol. 9, no. 4, pp. 243–252, 2017.
- [14] Q. Li, "The evaluation of equity financing efficiency for listed companies based on super-efficiency DEA model," *Evista de la Facultad de Ingeniería*, vol. 32, no. 4, pp. 652–658, 2017.
- [15] P. Sunega and M. Lux, "Market-based housing finance efficiency in the Czech Republic," *European Journal of Housing Policy*, vol. 7, no. 3, pp. 241–273, 2007.
- [16] D. Fletschner, C. Guirkinger, and S. Boucher, "Risk, credit constraints and financial efficiency in Peruvian agriculture," *The Journal of Development Studies*, vol. 46, no. 6, pp. 981–1002, 2010.
- [17] L. Ma, F. Xu, and Y. Yu, "Evaluation of financing efficiency of listed companies in photovoltaic industry based on deamodel," *Light & Engineering*, vol. 25, no. 3, pp. 71–78, 2017.
- [18] J. Dong, L. Zhu, B. Wang et al., "The evaluation of financing efficiency of China's stock market," *Mathematical Problems in Engineering*, vol. 2016, no. 6, Article ID 3236897, 13 pages, 2016.
- [19] C. Sheng and L. Zhang, "A research on financing efficiency of SMEs in neeq market: Private placement based on three stage DEA model," *Journal of Audit & Economics*, vol. 32, no. 3, pp. 78–86, 2017.
- [20] A. Charnes, W. W. Cooper, and E. Rhodes, "Measuring the efficiency of decision making units," *European Journal of Operational Research*, vol. 2, no. 6, pp. 429–444, 1978.
- [21] R. He, "The study on the problem of the relationship between the 'heterogeneity' of the neighbor enterprise and the financing efficiency of SMEs in China—empirical data from China industrial enterprises database," *Open Journal of Applied Sciences*, vol. 7, no. 4, pp. 171–183, 2017.

- [22] Y. Pan, Q. Yu, and M. Zhu, "A study on financing efficiency evaluation and its influencing factors of environmental protection industry in China," *East China Economic Management*, vol. 30, no. 2, pp. 77–83, 2016.
- [23] Y. Wang and Y. Zhang, "The impact of collaboration between internal governance and external financing on enterprises' productivity," *Industrial Economic Review*, vol. 9, no. 5, pp. 103–113, 2018.
- [24] R. Gatti and I. Love, "Does access to credit improve productivity? evidence from Bulgarian firms," *The Economics of Transition*, vol. 16, no. 3, pp. 445–465, 2008.
- [25] Y. Wu and F. Zeng, "Research on the measurement of SMEs financing efficiency in Chinas New Third Board," *Hubei Social Sciences*, vol. 385, no. 1, pp. 71–79, 2019.
- [26] R. G. Rajan and L. Zingales, "Financial dependence and growth," *The American Economic Review*, vol. 88, no. 3, pp. 559–586, 1998.
- [27] Y. Zhang and L. Zhao, "Government support and financial development, social capital and the financing efficiency of S&T innovation enterprises," *Science Research Management*, vol. 36, no. 11, pp. 55–63, 2015.
- [28] L. Wang, "An empirical study on the impact of guanxi and trust on external financing efficiency in clusters," *iBusiness*, vol. 07, no. 01, pp. 18–24, 2015.
- [29] V. Maksimovic and A. Demirgüçkunt, "Institutions, financial markets, and firm debt maturity," *Journal of Financial Economics*, vol. 54, no. 3, pp. 295–336, 1999.
- [30] J. Ma and Z. Guo, "The parameter basin and complex of dynamic game with estimation and two-stage consideration," *Applied Mathematics and Computation*, vol. 248, pp. 131–142, 2014.
- [31] J. Ma and L. Xie, "The comparison and complex analysis on dual-channel supply chain under different channel power structures and uncertain demand," *Nonlinear Dynamics*, vol. 83, no. 3, pp. 1379–1393, 2016.
- [32] J. Ma and X. Ma, "Measure of the bullwhip effect considering the market competition between two retailers," *International Journal of Production Research*, vol. 55, no. 2, pp. 313–326, 2016.
- [33] Y. Li, L. Yang, B. Yang, N. Wang, and T. Wu, "Application of interpretable machine learning models for the intelligent decision," *Neurocomputing*, vol. 333, pp. 273–283, 2019.
- [34] J. Ma and H. Wang, "Complexity analysis of dynamic non-cooperative game models for closed-loop supply chain with product recovery," *Applied Mathematical Modelling: Simulation and Computation for Engineering and Environmental Systems*, vol. 38, no. 23, pp. 5562–5572, 2014.
- [35] J. Ma and B. Bao, "Research on bullwhip effect in energy-efficient air conditioning supply chain," *Journal of Cleaner Production*, vol. 143, pp. 854–865, 2017.
- [36] H. Cao, T. Lin, Y. Li, and H. Zhang, "Stock price pattern prediction based on complex network and machine learning," *Complexity*, vol. 2019, Article ID 4132485, 12 pages, 2019.
- [37] J. Ma, W. Yang, and W. Lou, "Research on bifurcation and chaos in a dynamic mixed game system with oligopolies under carbon emission constraint," *International Journal of Bifurcation and Chaos*, vol. 27, no. 10, Article ID 1750158, 2017.
- [38] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [39] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer Series in Statistics, New York, NY, USA, 2001.
- [40] H. Gong, Y. Sun, X. Shu, and B. Huang, "Use of random forests regression for predicting IRI of asphalt pavements," *Construction and Building Materials*, vol. 189, pp. 890–897, 2018.
- [41] S. Malmquist, "Index numbers and indifference surfaces," *Trabajos de Estadística*, vol. 4, pp. 209–242, 1953.
- [42] D. W. Caves, L. R. Christensen, and W. E. Diewert, "The economic theory of index numbers and the measurement of input, output, and productivity," *Econometrica*, vol. 50, no. 6, pp. 1393–1414, 1982.
- [43] R. Fare, S. Grosskopf, and M. Norris, "Productivity growth, technical progress, and efficiency change in industrialized countries," *American Economic Review*, vol. 84, no. 55, pp. 1040–1044, 1997.
- [44] J. Tobin, "Estimation of relationships for limited dependent variables," *Econometrica*, vol. 26, no. 1, pp. 24–36, 1958.
- [45] D. Cumming and S. Johan, "Is it the law or the lawyers? investment covenants around the world," *European Financial Management*, vol. 12, no. 4, pp. 535–574, 2006.
- [46] A. Jaiantilal, "Randomforest-matlab," 2019, <http://code.google.com/p/randomforest-matlab/downloads/list>.

Research Article

Application of BP Neural Network Model in Risk Evaluation of Railway Construction

Yang Changwei ¹, **Li Zonghao** ¹, **Guo Xueyan** ¹, **Yu Wenyong** ²,
Jin Jing ³, and **Zhu Liang** ⁴

¹Department of Civil Engineering, Southwest Jiaotong University, 610031 Chengdu, China

²Oklahoma State University, Stillwater, OK, USA

³China Academy of Railway Sciences, 100081 Beijing, China

⁴China Railway Corporation, 100844 Beijing, China

Correspondence should be addressed to Yu Wenyong; yuwenyong@my.swjtu.edu.cn

Received 15 February 2019; Accepted 30 April 2019; Published 2 June 2019

Guest Editor: Thiago Christiano Silva

Copyright © 2019 Yang Changwei et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Chinese railway construction project is an important part of the implementation of the “Belt and Road” strategy, and the risk evaluation of overseas railway construction is the primary link of the project. Firstly, this paper mainly analyzes the Asian and European countries along the railway construction project, establishes a railway construction project risk evaluation system, and synthesizes various risk factors. Secondly, it establishes two independent BP neural network models by using different training algorithms because of the different political, economic, and cultural elements between the two continents.

1. Introduction

Since the “the Belt and Road” cooperation initiative was put forward, China is committed to the construction of overseas railways. Railways are the key means of forming crucial international channels and achieving interconnection [1–3]. At the same time, China’s railway technology is technologically advanced, safe and reliable, compatible, and cost-effective. These advantages make this line a hot project which attract the countries involved in the Belt and Road. However, since “the Belt and Road” involves many countries, the national conditions and needs of different countries vary, which makes it necessary to design different policies according to different conditions. Besides, a railway project needs a large investment, a long construction period that involves many links. Therefore, the various risk factors are complex and need to be treated differently according to differently countries, aiming to assess project risks and provide decision support for railway construction in order to avoid and control risks [4, 5]. The traditional risk evaluation method cannot meet the current complicated situation, so it is of great practical significance to seek a new method

suitable for overseas railway investment and construction risk evaluation. Figure 1 shows the overall trend of “the Belt and Road” Eurasia region railway [6, 7].

This paper mainly analyzes countries along “the Belt and Road” strategic line in Asia and Europe, establish a risk evaluation index system, and use the BP neural network toolbox in MATLAB under this risk evaluation index system to process the data to build a risk prediction model.

2. Railway Construction Risk Evaluation Method

2.1. Evaluation Index System. The railway construction technology expert resource and investment should be integrated based on the national conditions of Asian and European countries, the factors needed to be considered for railway construction, and the principles of being scientifically, systematically, typically, and feasibly practical. The opinions of experts, market operation experts, and venture capital experts establish an evaluation index system, as shown in Figure 2 [8, 9].



FIGURE 1: “The Belt and Road” railway roadmap.

The risk is roughly divided into several aspects.

(1) *Economic Risk.* The risk refers to the risk that may be brought about by changes in the social economic situation of the country where the project is located, including the country's import and export volume, per capital GDP, the turnover of foreign contracted projects, inflation rate, and economic prospects. It affects willingness and ability to pay of a country to some extent.

(2) *Population Risk.* The risk is analyzed according to the average quality of people's life of the target country, including population density, population growth rate, average life expectancy, infectious disease mortality, and the number of refugees. It reflects the living standards of the people in that country to a certain extent.

(3) *Traffic Risk.* The risk is analyzed based on the current national overall traffic level of the target country. It includes

elements such as transportation services, power coverage, and traffic accident rate, which reflect the current development of the country's transportation industry to a certain extent.

(4) *Political Risk.* The risk is analyzed based on the current domestic policies and regulations of the target country, including the cooperative relationship with China, the law, civil political freedom, and political stability. The unstable political situation of Central Asia and the Middle East is quite dangerous. If there is any big change, there will be huge impact on the project, even Casualties.

2.2. Comprehensive Evaluation Criteria. Among the 17 factors that affect the evaluation index of railway investment and construction project risks, there is no direct relationship between each factor, and the dimensions are very different. Therefore, before the neural network model being

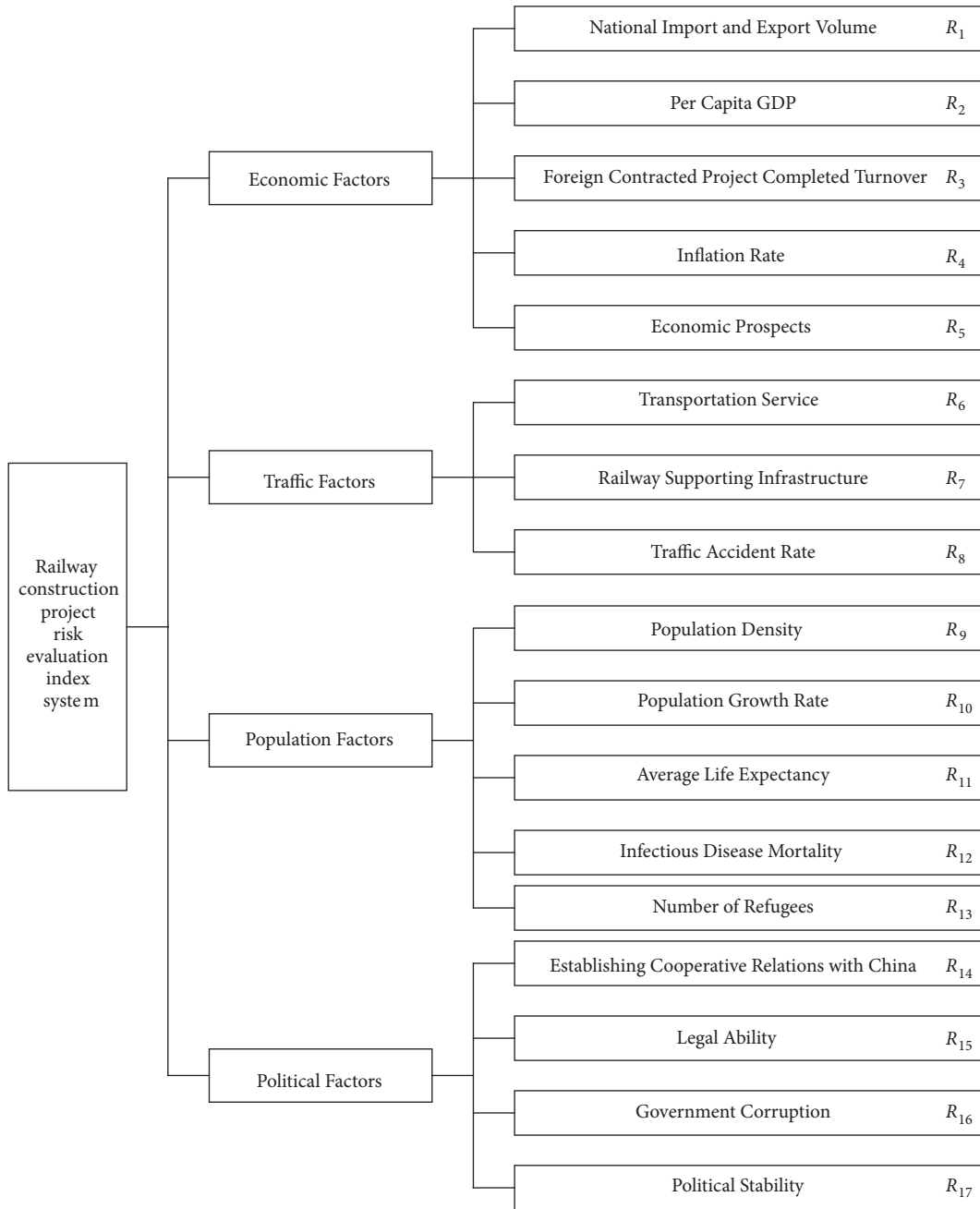


FIGURE 2: Railway construction project risk evaluation index system.

established, the data should be normalized. That is to say, the data is limited to a certain range and classified by the expert after the data is processed. When scoring, it should be fully reviewed by experts; the specific results are shown in Table 1.

Due to the large social and cultural differences between Asia and Europe, as well as the large gap in economic development, this paper will use the same system for evaluation in Asia and Europe but score and model independently, which means the criteria for judging in Asia and Europe are independent. Experts will score 17 factors according to different environments in Asia and Europe and work out the

final score based on the final comprehensive opinions and the scores of each factor, which also satisfies the scoring standard of Table 1; that is, the higher the value the lower the risk of building a railway in that country and vice versa.

3. Establish BP Neural Network Model

3.1. Introduction to BP Neural Network. Artificial neural network is a model based on human brain [10–13]. It has a neuron system composed of many neurons, which has the advantages of massive parallelism, distributed processing,

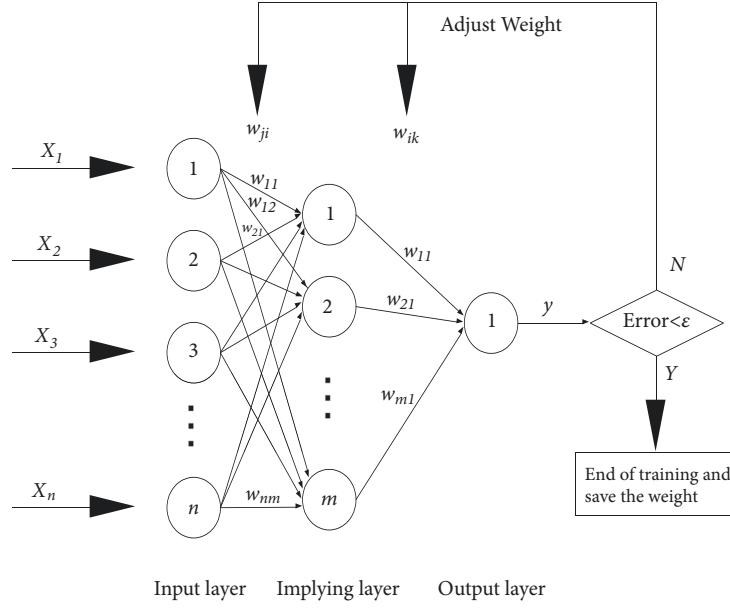


FIGURE 3: Typical 3-layer BP network flow chart.

TABLE 1: Experts' criteria for risk.

Risk Grade	High	Mid-High	Medium	Mid-Low	Low
Expert Grading	0~0.2	0.2~0.4	0.4~0.6	0.6~0.8	0.8~1.0

self-organization, and self-learning. Among these models, multilayer forward BP network is the most widely used neural network form. It has the universal advantages of all neural networks, self-learning and self-adaptive ability, nonlinear mapping ability, and high fault tolerance rate. Many problems that cannot be solved by traditional information processing methods have made some progress after using neural networks. For example, the risk evaluation of this paper and BP neural network can play a vital role.

Figure 3 is a typical BP neural network flow chart, which shows the whole process of BP neural network for data processing: firstly, input the variables (input layer), then deal with the data through the processing of the function, adjust the weight of each inputted variable, and finally compare the output value with the target value. If the satisfied accuracy is not met, then readjust the weight until the output value meets the error requirement. In this case, the scores of the 17 risk evaluation factor are the input values; the synthesized score of each country is the target value. By inputting 17 risk scores and synthesized scores of each country, the BP neural network functions like experts who can give an overall score; thus it forms a railway construction risk evaluation model.

3.2. Basic Principle of BP Neural Network. There are many training functions in BP neural network; the most common ones are *traind* (gradient descent method), *traindm* (momentum gradient descent method), *trainda* (adaptive learning rate gradient descent method), *traindx* (adaptive learning rate and momentum gradient descent method),

traincgf/*traincgp*/*traincgb* (three conjugate gradient methods), etc. Each one of them has different parameters and different training methods that are based on the basic principles of BP neural network [14–16]. The basic principles are as follows.

The essence of neural networks is the following function:

$$\vec{y} = f_{network}(\vec{x}) \quad (1)$$

In the input layer, the input node spreads the inputted information to the hidden layer node through activating function $f(u)$. There are two kinds of activation functions: Sigmoid and Purein. They are nonlinear for either nonlinear data or linear data. Thus, this model is nonlinear; the expression of sigmoid is

$$f(u) = [1 + \exp(-u)]^{-1} \quad (2)$$

Supposing that w is the input layer and the hidden layer, the random weight between the hidden layer and the output layer, the vector is \vec{w} , the output of the hidden layer node is represented by O , and then there is

$$O = \text{sigmoid}(\vec{w}^T \cdot \vec{x}) \quad (3)$$

Similarly, the output of the output layer node is represented by P ,

$$P = \text{sigmoid}(\vec{w}^T \cdot \vec{O}) \quad (4)$$

After the output is figured out, the node output error δ can be calculated as follows:

$$\delta = P(1 - P)(T - P) \quad (5)$$

T is the target value of the output node, and for hidden layer nodes,

$$\delta_i = a_i(1 - a_i) \sum_{k \in \text{outputs}} w_{ki} \delta_k \quad (6)$$

where δ_i is the error term of the hidden layer node i , a_i is the output of the hidden layer node i , and w_{ki} is the weight of the node i to the next layer node k . And δ_k is the error term of the node k also the next layer of the node i . Finally, the weight on each link is updated:

$$w_{ji} \leftarrow w_{ji} + \eta \delta_j x_{ji} \quad (7)$$

In this expression, w_{ji} is the weight of node i to node j , η is the learning rate constant, δ_j is the error term of node j , and x_{ji} is the input passed by node i to node j .

The content above is the calculation of the error term of each node of the BP neural network and the weight update method. To calculate the error term of a node, it is necessary to firstly calculate the error term of each node connected to it. This requires to begin with calculating the error terms of the output layer, then the error terms of each hidden layer are calculated reversely until the hidden layer connected to the input layer. This is the meaning of the name of the reverse propagation (BP) algorithm.

3.3. Model and Algorithmic Principles for Asia. As mentioned earlier in this article, the railway construction project spans the Eurasian continent and passes through dozens of countries; each country has different national conditions. The regional differences between the two continents are obvious. Therefore, the target countries are divided into Asian district and European district. Thus, different learning functions and separate training models are applied.

In general, the number of hidden layers increases, complicating the network, thereby increasing the training time of the network and the tendency of "overfitting". After many tests and extensive statistical analysis, the neural network is designed to be 3 layers containing one hidden layer based on the calculate accuracy requirement and training time, which is the same as other research results [17–21].

From a view of the overall situation in Asia, the economic development of some country is relatively backward, the political situation of some country is turbulent, and the overall situation is even more complicated. This paper adopts the traindx (adaptive learning rate momentum gradient descent method) algorithm that is attached to the most basic gradient descent method. The momentum and automatic adjustment of the learning rate: the standard BP algorithm is essentially a simple steepest descent static optimization method, which does not consider the direction of the error gradient descent, so that the learning process is often oscillated, the convergence is slow, and the additional momentum is the part of the last weight adjustment. It is added to the weight adjustment

amount calculated according to the current error. As the actual weight adjustment amount, the essence of this is the influence of the last weight change, which is transmitted by a momentum factor. The weight adjustment formula with an additional momentum factor is

$$\begin{aligned} W(k+1) - W(k) &= \Delta W(k+1) \\ &= (1 - mc) \eta \times \delta_i O_j \\ &\quad \times [W(k) - W(k-1)] \end{aligned} \quad (8)$$

In this expression, W is the weight, ΔW is the weight increment, k is the training number, mc is the momentum factor, η is the learning rate, δ_i is the error of the output node i , and O_j is the input of the input node j . The role of momentum is to remember the direction of the change of the last connection weight (positive and negative values), so that you can use a higher learning rate to improve learning speed.

Based on the additional momentum, the traindx algorithm also uses the adaptive learning rate method. An important reason for the slow convergence rate of the standard BP algorithm is that the learning rate is not properly selected, and the rate cannot be changed after the rate is determined. When the learning rate is low, the training time will be long and the convergence will be slow. If the learning rate is too high, the overfitting will be caused and the data oscillation finally will diverge, so the adaptive learning rate method is used to solve this problem. The principle is as follows: check whether the modification of each weight changes the error. It means that the selected learning rate value is small if the error becomes smaller, and an appropriate amount can be added to it; if the error increases, then it should be reduced. The value of the learning rate can be expressed by the following equation:

$$\eta(K+1) = \begin{cases} a\eta(k) & E[W(K+1)] < E[W(K)] \\ b\eta(k) & E[W(K+1)] < kE[W(K)] \\ \eta(k) & \text{Other} \end{cases} \quad (9)$$

where η is the learning rate, K is the number of trainings, and E is the error function, $E = \sum_{i=1}^n (y_i - \hat{y}_i)^2$, $i = 1, 2, \dots, n$, y_i is the actual output value after network training, \hat{y}_i is the expected output value of the learning sample, and n is the number of learning samples.

Additional momentum can help to find better solutions, and adaptive learning can shorten the network training time. The combination of the two is the traindx training algorithm. There are many countries in Asia which have complicated situations. The traindx algorithm is relatively suitable for these countries. The sample set of the training network should be the authoritative evaluation result with high credibility, which can be obtained by experts in some Asian countries. The trained model is stored, and the risk prediction is performed on the project that needs to be predicted. If the corresponding risk evaluation value is input, the neural network system can calculate the comprehensive risk evaluation value of the project through the previously calculated weights and thresholds. And outputting the output

layer as a network result, you can get the final evaluation result.

For the confirmation of the number of neurons in the hidden layer, the formula $m = \sqrt{n+l} + a$ is used in this manuscript (m is the number of neurons in the hidden layer, n is the number of neurons in the input layer, l is the number of neurons in the output layer, and a is an arbitrary integer from 0 to 5), which also is used in other researches [22–26]. In order to avoid “overfitting” during training and ensure high enough network performance and generalization ability, the basic principle for determining the number of hidden layer neurons is using the most compact structure while satisfying the accuracy requirements. That is, take as few hidden layer neurons as possible. When $a=0$, $m=4.24$, take 5, and then the number (5,6,7,8,9,10,11) of hidden layer neurons is used to verify the error rate until the best error rate is obtained. Table 1 shows the error rate of each selected hidden layer neuron, and the error meets accuracy requirement (error < 5%) when the number of hidden layer neurons is 6. Therefore, 6 neurons in the hidden layer are selected in this paper.

Based on the above BP neural network model and risk evaluation system, the macrorisks of 36 Asian countries along “the Belt and Road” were evaluated; Table 2 shows these countries. First, based on the expert experience method, the 17 risks of each country are evaluated, and then the 17 scores are combined to evaluate the risk of railway construction in each country. The results are shown in Table 3. In the model for Asian countries, the data for training set comes from United Arab Emirates/Oman/Azerbaijan/Pakistan/Bahrain/Bhutan/Philippines/Georgia/Hassackstein/Korea/Qatar/Kuwait/Laos/Lebanon/Maldives/Malaysia/Mongolia/Bangladesh/Saudi Arabia/Sri Lanka/Turkey/Iran. The training function uses *traindx* to adjust the training parameters until the error requirements of the training set are met, as shown in Table 4. The parameters are selected as follows: 6000 cycles of maximum epochs, 0.001 for training accuracy, and 0.05 for learning rate (*lr*), the learning rate growth ratio is 1.05 (*lr_inc*), the learning rate reduction ratio (*lr_dec*) is 0.7, the maximum performance increment (*max_perf_inc*) is 1.04, the momentum factor (*mc*) is 0.9, and the rest are default values. The training results are shown in Table 5.

It can be seen from Table 5 that the average relative error of the 22 training data in the Asian training set countries is 2.70%, and the maximum error in the sample set is 4.88% which satisfies the accuracy requirement of 5%; this means the model training learning result is good. Then the verification samples are replaced by the remaining 14 countries to verify whether the model is valid; eventually the results will be compared with the expert score results. The results are shown in Table 6.

It can be seen from Table 6 that the average relative error of the verification data set is 2.99% and the maximum relative error of the prediction set is 4.22%, which meets the accuracy requirement of 5%, indicating that the neural network model can achieve the accuracy required by the project and can be used for risk assessment and prediction of Asian countries in railway construction projects.

3.4. Modeling and Algorithmic Principles for Europe. In Europe, the gap between various factors in various countries is not very large, the situation is relatively stable, the economy is more developed, and *traindx* is suitable for networks that have more complex and more data and may be used for smaller models. The error is increased so that *trainda* (adaptive learning rate algorithm) is selected; that is, no momentum is added except that the adaptive learning rate is added. The principle is shown in the previous section, Table 7 shows these countries; the expert risk evaluation in Europe is shown in Table 8. In the model for European countries, the data for training set comes from Albania/Estonia/Belarus/Bulgaria/Bosnia/Poland/Russia/Montenegro/Croatia/Latvia/Lithuania/Romania/Macedonia/Czech/Moldova. *Trainda* is chosen as the training function; then adjusting the training parameters should be done until the error requirements of the training set are met. The parameter selection is as follows: the European model hidden layer neurons are determined in the same way as the Asian model. Table 9 shows the error of the number of neurons, and the error meets accuracy requirement (error < 5%) when the number of hidden layer neurons is 6. Therefore, 6 neurons in the hidden layer are selected in this manuscript. The maximum number of cycles (epochs) is 5000 times, the training accuracy (goal) is 0.001, the learning rate (*lr*) takes 0.01, learning rate growth ratio (*lr_inc*) is 1.05, the learning rate reduction ratio (*lr_dec*) is 0.8, and the maximum performance increment (*max_perf_inc*) is 1.04. The rest are taken as default values. The training results are shown in Table 10.

It can be seen from Table 10 that the average relative error of the 15 training data in Europe is 3.69%, and the maximum error in the sample set is 4.77% which meets the accuracy requirement of 5%; that is to say, the model training learning result is better. Then the verification sample is replaced by the remaining 5 countries to verify whether the model is valid or not; lastly compare the results with the expert score results. The results are shown in Table 11.

It can be seen from Table 11 that the average relative error of the verification data set is 4.37% and the maximum relative error of the prediction set is 4.97%, indicating that the neural network model has been able to achieve the accuracy required by the project and can be used for railway outbuilding projects in European national risk evaluation forecast.

3.5. Model Robustness Test. It can be seen from the above training and testing that the accuracy of the model is good, and for machine learning, robustness is also an important feature, and the artificial neural network itself is robust [27–29]. The following is a robust test using the European model as an example: it is assumed that some data in the first 10 countries are randomly interfered and the data is shown in Table 12. The interference data is substituted into the model for simulation calculation, and the error between the verification result and the score given by the expert is shown in Table 13. The error between the verification result and the model score is shown in Table 14.

TABLE 2: Number 36 countries in Asia.

Number	1	2	3	4	5	6	7	8	9
Country	United Arab Emirates	Oman	Azerbaijan	Pakistan	Bahrain	Bhutan	Philippines	Georgia	Kazakhstan
Number	10	11	12	13	14	15	16	17	18
Country	Korea	Qatar	Kuwait	Laos	Lebanon	Maldives	Malaysia	Mongolia	Bangladesh
Number	19	20	21	22	23	24	25	26	27
Country	Saudi Arabia	Sri Lanka	Turkey	Iran	Thailand	Turkmenistan	Brunei	Uzbekistan	Singapore
Number	28	29	30	31	32	33	34	35	36
Country	Syria	Armenia	Yemen	Iraq	Israel	India	Indonesia	Jordan	Vietnam

TABLE 3: Risk scores for railway construction projects in 36 countries in Asia.

Country	Risk																	score
	R ₁	R ₂	R ₃	R ₄	R ₅	R ₆	R ₇	R ₈	R ₉	R ₁₀	R ₁₁	R ₁₂	R ₁₃	R ₁₄	R ₁₅	R ₁₆	R ₁₇	
1	0.78	0.90	0.62	0.80	0.90	0.52	0.78	0.67	0.50	0.71	0.70	0.37	0.61	0.62	0.01	0.81	0.86	0.819
2	0.65	0.71	0.44	0.86	0.79	0.81	0.56	0.71	0.14	0.98	0.75	0.37	0.78	0.27	0.82	0.71	0.78	0.792
3	0.19	0.61	0.34	0.72	0.45	0.41	0.32	0.45	0.50	0.48	0.45	0.79	0.53	0.37	0.74	0.39	0.42	0.545
4	0.61	0.22	0.91	0.29	0.40	0.38	0.56	0.39	0.64	0.77	0.39	0.11	0.04	0.77	0.97	0.41	0.12	0.491
5	0.21	0.77	0.04	0.82	0.87	0.88	0.23	0.70	0.86	0.67	0.78	0.78	0.70	0.37	0.28	0.59	0.29	0.602
6	0.03	0.39	0.00	0.52	0.62	0.67	0.46	0.67	0.16	0.56	0.67	0.26	0.28	0.13	0.83	0.82	0.86	0.505
7	0.72	0.41	0.62	0.79	0.61	0.67	0.76	0.63	0.66	0.59	0.58	0.32	0.77	0.63	0.32	0.66	0.32	0.663
8	0.15	0.52	0.38	0.70	0.84	0.87	0.41	0.71	0.31	0.23	0.78	0.95	0.57	0.27	0.49	0.76	0.45	0.674
9	0.70	0.69	0.79	0.33	0.51	0.59	0.53	0.56	0.06	0.53	0.55	0.75	0.58	0.88	0.67	0.44	0.59	0.612
10	0.99	0.80	0.37	0.83	0.95	0.95	0.86	0.72	0.75	0.27	0.87	0.62	0.38	0.77	0.76	0.75	0.71	0.691
11	0.53	0.99	0.62	0.82	0.91	0.90	0.75	0.72	0.55	0.90	0.87	0.31	0.79	0.63	0.92	0.73	0.95	0.733
12	0.61	0.92	0.57	0.87	0.71	0.75	0.65	0.69	0.59	0.89	0.60	0.65	0.61	0.37	0.85	0.68	0.67	0.773
13	0.33	0.30	0.74	0.73	0.35	0.29	0.35	0.39	0.21	0.61	0.23	0.09	0.99	0.89	0.82	0.36	0.69	0.535
14	0.33	0.62	0.07	0.77	0.48	0.43	0.41	0.47	0.76	0.91	0.49	0.86	0.06	0.27	0.53	0.28	0.18	0.542
15	0.07	0.61	0.22	0.51	0.51	0.45	0.65	0.51	0.87	0.80	0.50	0.55	0.98	0.49	0.67	0.56	0.70	0.503
16	0.95	0.69	0.81	0.80	0.88	0.89	0.78	0.71	0.42	0.54	0.80	0.35	0.29	0.89	0.39	0.70	0.66	0.791
17	0.52	0.53	0.53	0.38	0.52	0.58	0.63	0.58	0.03	0.67	0.56	0.59	0.90	0.89	0.61	0.67	0.78	0.651
18	0.57	0.13	0.68	0.52	0.34	0.29	0.32	0.38	0.86	0.43	0.31	0.23	0.19	0.77	0.52	0.49	0.25	0.563
19	0.84	0.78	0.99	0.85	0.73	0.76	0.69	0.69	0.14	0.82	0.67	0.37	0.62	0.88	0.55	0.71	0.50	0.743
20	0.43	0.45	0.71	0.61	0.64	0.69	0.58	0.65	0.67	0.34	0.61	0.66	0.67	0.63	0.72	0.66	0.46	0.635
21	0.65	0.68	0.58	0.43	0.79	0.81	0.63	0.70	0.43	0.54	0.76	0.85	0.05	0.63	0.62	0.61	0.27	0.714
22	0.73	0.61	0.72	0.16	0.41	0.51	0.52	0.47	0.26	0.46	0.53	0.65	0.12	0.89	0.04	0.52	0.27	0.536
23	0.87	0.54	0.63	0.82	0.73	0.77	0.41	0.69	0.54	0.27	0.68	0.35	0.23	0.88	0.18	0.55	0.28	0.635
24	0.52	0.61	0.57	0.70	0.18	0.01	0.32	0.21	0.09	0.47	0.03	0.44	0.86	0.63	0.66	0.03	0.67	0.442
25	0.22	0.89	0.17	0.85	0.92	0.90	0.68	0.72	0.35	0.54	0.85	0.77	0.99	0.63	0.21	0.72	0.94	0.406
26	0.42	0.31	0.46	0.18	0.30	0.50	0.36	0.40	0.35	0.74	0.44	0.58	0.77	0.62	0.30	0.25	0.45	0.457
27	0.91	0.94	0.81	0.83	0.99	0.99	0.81	0.71	0.99	0.65	0.99	0.68	0.92	0.48	0.10	0.93	0.92	0.865
28	0.19	0.30	0.12	0.17	0.26	0.50	0.12	0.39	0.51	0.23	0.45	0.37	0.13	0.27	0.51	0.15	0.01	0.135
29	0.13	0.46	0.14	0.72	0.62	0.69	0.42	0.67	0.51	0.26	0.60	0.93	0.32	0.38	0.17	0.61	0.61	0.442
30	0.45	0.22	0.33	0.23	0.30	0.21	0.36	0.34	0.26	0.83	0.18	0.17	0.17	0.13	0.53	0.13	0.06	0.263
31	0.70	0.56	0.87	0.82	0.21	0.14	0.06	0.25	0.35	0.86	0.04	0.20	0.22	0.63	0.08	0.17	0.15	0.314
32	0.55	0.86	0.22	0.79	0.96	0.95	0.58	0.71	0.74	0.76	0.92	0.87	0.30	0.49	0.31	0.76	0.28	0.741
33	0.88	0.29	0.98	0.59	0.60	0.66	0.75	0.66	0.74	0.45	0.59	0.19	0.21	0.63	0.91	0.68	0.28	0.736
34	0.83	0.44	0.95	0.62	0.57	0.63	0.36	0.61	0.55	0.48	0.56	0.36	0.41	0.88	0.62	0.64	0.42	0.743
35	0.38	0.52	0.24	0.67	0.77	0.79	0.62	0.67	0.42	0.86	0.69	0.54	0.01	0.62	0.43	0.68	0.43	0.577
36	0.88	0.32	0.89	0.35	0.51	0.50	0.45	0.57	0.65	0.39	0.53	0.49	0.68	0.89	0.75	0.63	0.68	0.683

TABLE 4: Asian model error corresponding to the number of neurons in different hidden layers.

Country	Neurons						
	5	6	7	8	9	10	11
1	5.76%	0.23%	3.96%	7.30%	8.17%	7.09%	6.94%
2	1.68%	4.52%	2.09%	4.40%	3.85%	5.65%	13.55%
3	1.75%	0.39%	4.36%	2.29%	5.11%	4.45%	6.79%
4	4.58%	4.37%	3.24%	1.47%	8.47%	5.44%	2.16%
5	3.38%	0.77%	3.50%	1.58%	3.91%	3.55%	6.48%
6	3.97%	4.78%	4.90%	4.36%	8.70%	4.52%	11.33%
7	5.00%	4.88%	4.19%	2.12%	7.08%	16.59%	14.55%
8	3.55%	2.24%	3.47%	0.33%	5.61%	6.87%	2.17%
9	3.12%	0.75%	2.96%	5.29%	2.26%	1.45%	12.32%
10	2.06%	3.99%	4.43%	9.33%	4.76%	4.29%	7.69%
11	2.98%	1.90%	3.20%	6.13%	5.90%	7.19%	7.29%
12	4.17%	3.44%	2.41%	4.54%	6.86%	5.35%	14.23%
13	1.27%	1.28%	3.82%	2.82%	7.02%	6.57%	0.39%
14	1.89%	2.40%	3.24%	2.20%	0.29%	6.25%	12.37%
15	5.39%	4.24%	3.32%	5.37%	1.85%	3.79%	15.91%
16	1.97%	3.15%	4.46%	5.29%	0.84%	3.38%	14.56%
17	2.69%	1.11%	3.69%	1.61%	5.40%	0.89%	11.06%
18	2.75%	1.22%	4.76%	4.08%	5.57%	7.74%	7.56%
19	2.58%	4.22%	4.27%	4.80%	11.97%	11.60%	0.81%
20	4.70%	4.65%	4.20%	6.97%	1.85%	10.02%	3.82%
21	3.91%	2.80%	0.32%	5.02%	1.39%	0.05%	7.49%
22	5.18%	2.10%	2.17%	4.40%	2.08%	6.89%	16.78%

TABLE 5: BP neural network results of training on railway construction risks in 22 Asian countries.

NO.	1	2	3	4	5	6	7	8	9	10	11
Expert mark	0.819	0.792	0.545	0.491	0.602	0.505	0.663	0.674	0.612	0.691	0.733
Training results	0.817	0.756	0.542	0.512	0.597	0.480	0.695	0.689	0.616	0.718	0.747
Error(%)	0.23	4.52	0.39	4.37	0.77	4.78	4.88	2.24	0.75	3.99	1.90
NO.	12	13	14	15	16	17	18	19	20	21	22
Expert mark	0.773	0.535	0.542	0.503	0.791	0.651	0.563	0.743	0.635	0.714	0.536
Training results	0.746	0.541	0.529	0.524	0.766	0.658	0.556	0.774	0.664	0.694	0.524
Error(%)	3.44	1.28	2.40	4.24	3.15	1.11	1.22	4.22	4.65	2.80	2.10

TABLE 6: Validation results of BP neural network on construction risk of railways in 14 Asian countries.

NO.	23	24	25	26	27	28	29	30	31	32	33	34	35	36
Expert mark	0.635	0.442	0.406	0.457	0.865	0.135	0.442	0.263	0.314	0.741	0.736	0.743	0.577	0.683
Training results	0.612	0.452	0.390	0.451	0.833	0.129	0.451	0.274	0.322	0.711	0.709	0.729	0.575	0.656
Error (%)	3.62	2.47	3.92	1.31	3.64	4.15	2.16	4.22	2.55	4.05	3.65	1.88	0.35	3.95

TABLE 7: Number 20 countries in Europe.

Number	1	2	3	4	5	6	7	8	9	10
Country	Albania	Estonia	Belarus	Bulgaria	Bosnia	Poland	Russia	Montenegro	Czech	Croatia
Number	11	12	13	14	15	16	17	18	19	20
Country	Latvia	Lithuania	Romania	Macedonia	Moldova	Serbia	Slovakia	Slovenia	Ukraine	Hungary

TABLE 8: Risk scores for railway construction projects in 20 countries in Europe.

NO.	R ₁	R ₂	R ₃	R ₄	R ₅	R ₆	R ₇	R ₈	R ₉	R ₁₀	R ₁₁	R ₁₂	R ₁₃	R ₁₄	R ₁₅	R ₁₆	R ₁₇	score
1	0.33	0.27	0.35	0.66	0.32	0.40	0.36	0.68	0.73	0.47	0.34	0.50	0.91	0.38	0.45	0.41	0.47	0.525
2	0.51	0.87	0.23	0.52	0.96	0.98	0.56	0.35	0.24	0.43	0.89	0.66	0.95	0.18	0.45	0.85	0.71	0.563
3	0.46	0.48	0.79	0.36	0.16	0.36	0.63	0.78	0.33	0.66	0.43	0.33	0.63	0.90	0.72	0.55	0.44	0.532
4	0.53	0.56	0.15	0.26	0.75	0.72	0.62	0.32	0.44	0.25	0.97	0.90	0.32	0.58	0.28	0.57	0.54	0.623
5	0.11	0.40	0.76	0.63	0.34	0.33	0.36	0.74	0.67	0.51	0.31	0.79	0.37	0.33	0.57	0.42	0.25	0.543
6	0.88	0.72	0.44	0.62	0.79	0.80	0.76	0.43	0.87	0.61	0.73	0.47	0.23	0.90	0.52	0.72	0.88	0.811
7	0.85	0.73	0.77	0.39	0.42	0.27	0.82	0.76	0.19	0.92	0.26	0.42	0.04	0.90	0.15	0.38	0.33	0.802
8	0.16	0.49	0.78	0.66	0.56	0.54	0.42	0.56	0.33	0.58	0.53	0.90	0.45	0.38	0.40	0.55	0.60	0.621
9	0.84	0.90	0.42	0.63	0.94	0.93	0.75	0.29	0.93	0.82	0.85	0.62	0.49	0.72	0.38	0.69	0.93	0.767
10	0.42	0.68	0.73	0.72	0.66	0.64	0.62	0.50	0.52	0.34	0.63	0.73	0.70	0.43	0.68	0.62	0.69	0.621
11	0.42	0.75	0.13	0.58	0.85	0.83	0.65	0.39	0.26	0.16	0.80	0.55	0.83	0.18	0.67	0.66	0.63	0.602
12	0.48	0.77	0.11	0.62	0.89	0.86	0.56	0.34	0.35	0.10	0.81	0.41	0.62	0.18	0.36	0.74	0.79	0.654
13	0.66	0.60	0.58	0.51	0.60	0.57	0.81	0.54	0.63	0.31	0.59	0.70	0.51	0.58	0.87	0.61	0.50	0.625
14	0.23	0.37	0.57	0.58	0.55	0.52	0.31	0.58	0.63	0.86	0.51	0.96	0.58	0.38	0.76	0.46	0.31	0.571
15	0.09	0.12	0.09	0.44	0.37	0.35	0.42	0.71	0.85	0.56	0.39	0.43	0.76	0.38	0.08	0.34	0.36	0.455
16	0.35	0.40	0.62	0.46	0.49	0.47	0.45	0.64	0.61	0.27	0.44	0.97	0.16	0.90	0.35	0.46	0.41	0.620
17	0.73	0.83	0.20	0.69	0.73	0.71	0.68	0.45	0.79	0.77	0.68	0.52	0.66	0.38	0.58	0.65	0.93	0.725
18	0.59	0.99	0.04	0.66	0.85	0.82	0.62	0.38	0.71	0.84	0.77	0.30	0.77	0.18	0.80	0.81	0.84	0.651
19	0.83	0.20	0.68	0.35	0.21	0.21	0.82	0.79	0.54	0.39	0.20	0.36	0.45	0.72	0.83	0.39	0.17	0.589
20	0.79	0.70	0.33	0.55	0.71	0.70	0.52	0.44	0.77	0.36	0.70	0.83	0.42	0.38	0.92	0.59	0.75	0.713

TABLE 9: Europe model error corresponding to the number of neurons in different hidden layers.

Country	Neurons						
	5	6	7	8	9	10	11
1	1.44%	4.22%	4.31%	3.58%	6.85%	2.67%	4.94%
2	4.73%	4.14%	3.00%	8.54%	4.49%	14.19%	8.70%
3	6.48%	2.29%	3.78%	7.36%	1.99%	10.95%	10.09%
4	7.05%	4.08%	6.23%	1.94%	6.62%	3.82%	4.45%
5	2.18%	2.94%	2.40%	8.07%	4.55%	5.03%	1.36%
6	2.68%	3.73%	5.04%	4.71%	5.29%	16.39%	9.16%
7	7.91%	4.77%	8.36%	0.22%	4.77%	5.95%	9.87%
8	4.08%	4.11%	2.31%	5.98%	3.27%	7.28%	2.45%
9	1.80%	3.56%	0.55%	6.61%	1.88%	14.33%	2.11%
10	1.74%	0.12%	4.44%	3.07%	8.49%	5.57%	5.64%
11	3.56%	4.15%	4.57%	2.27%	8.56%	4.15%	10.01%
12	4.57%	4.46%	1.52%	7.78%	5.83%	3.82%	5.41%
13	2.22%	4.36%	1.78%	5.51%	7.02%	6.27%	0.27%
14	2.67%	3.92%	2.72%	10.75%	7.89%	0.13%	0.69%
15	3.99%	4.56%	8.37%	6.01%	6.85%	7.86%	15.18%

It can be seen from the data in the table that even if the input sample data is partially interfered with the data, the model can guarantee the result that the accuracy requirement (error < 5%) is met after the network is running; that means the robustness test result is good.

4. Conclusion

According to different situations in Asia and Europe, BP neural network model is established by using different functions for risk evaluation. Through the created neural

network model, only experts can give the scores of the various risks in the macrorisk evaluation of the target country, and the overall construction risk score of the target country can be obtained without cumbersome manual scoring.

Due to the complex relationship between the various influencing factors in this project, it is not intuitive to use a linear expression to carry out risk prediction. BP neural network is selected to nonlinearly process the data to obtain the quantitative value of risk assessment. BP neural network does not need corresponding function equations for a set

TABLE 10: BP neural network results of training on railway construction risks in 15 European countries.

NO	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Expert mark	0.525	0.563	0.532	0.623	0.543	0.811	0.802	0.621	0.767	0.621	0.602	0.654	0.625	0.571	0.455
Training results	0.547	0.586	0.544	0.648	0.559	0.781	0.764	0.595	0.794	0.620	0.577	0.625	0.652	0.549	0.476
Error (%)	4.22	4.14	2.29	4.08	2.94	3.73	4.77	4.11	3.56	0.12	4.15	4.46	4.36	3.92	4.56

TABLE 11: Validation results of BP neural network on construction risk of railways in 5 European countries.

NO.	16	17	18	19	20
Expert mark	0.620	0.725	0.651	0.589	0.713
Training results	0.590	0.761	0.682	0.569	0.686
Error (%)	4.91	4.97	4.76	3.44	3.76

TABLE 12: Interfered data in the top 10 countries in Europe.

NO	1	2	3	4	5	6	7	8	9	10
R ₁	0.35	0.51	0.46	0.53	0.19	0.88	0.85	0.16	0.84	0.42
R ₂	0.27	0.79	0.48	0.56	0.40	0.72	0.73	0.49	0.86	0.68
R ₃	0.35	0.23	0.79	0.24	0.76	0.46	0.86	0.78	0.42	0.73
R ₄	0.66	0.52	0.36	0.26	0.63	0.62	0.39	0.72	0.63	0.72
R ₅	0.32	0.96	0.25	0.75	0.34	0.79	0.42	0.56	0.94	0.62
R ₆	0.40	0.98	0.36	0.72	0.33	0.80	0.27	0.54	0.93	0.64
R ₇	0.36	0.56	0.63	0.62	0.36	0.76	0.82	0.42	0.75	0.62
R ₈	0.68	0.43	0.78	0.32	0.74	0.48	0.76	0.56	0.29	0.50
R ₉	0.64	0.24	0.33	0.44	0.67	0.87	0.19	0.46	0.93	0.52
R ₁₀	0.47	0.43	0.61	0.25	0.59	0.61	0.82	0.58	0.75	0.34
R ₁₁	0.34	0.89	0.43	0.97	0.31	0.73	0.26	0.53	0.85	0.58
R ₁₂	0.50	0.66	0.33	0.81	0.79	0.47	0.42	0.90	0.62	0.73
R ₁₃	0.91	0.87	0.63	0.32	0.37	0.23	0.13	0.45	0.49	0.70
R ₁₄	0.38	0.18	0.90	0.58	0.28	0.85	0.90	0.27	0.68	0.43
R ₁₅	0.58	0.45	0.85	0.28	0.57	0.52	0.15	0.40	0.38	0.59
R ₁₆	0.41	0.85	0.55	0.64	0.42	0.72	0.38	0.55	0.69	0.62
R ₁₇	0.47	0.71	0.44	0.54	0.25	0.88	0.33	0.60	0.93	0.69

TABLE 13: Error test of interference data score and expert score.

NO	1	2	3	4	5	6	7	8	9	10
Expert score	0.525	0.563	0.532	0.623	0.543	0.811	0.802	0.621	0.767	0.621
Interference data score	0.541	0.589	0.549	0.609	0.545	0.773	0.782	0.593	0.762	0.652
Error (%)	3.05	4.64	3.15	2.32	0.43	4.68	2.55	4.46	0.66	4.96

TABLE 14: Error test of interference data score and model score.

NO	1	2	3	4	5	6	7	8	9	10
Model score	0.547	0.586	0.544	0.648	0.559	0.781	0.764	0.595	0.794	0.620
Interference data score	0.541	0.589	0.549	0.609	0.545	0.773	0.782	0.593	0.762	0.652
Error (%)	1.14	0.46	0.88	6.47	2.57	1.01	2.33	0.42	4.24	4.87

of nonlinear data; instead it iterates out the corresponding results and obtains an equation model that meets the requirements through its own training, which can meet the requirements of the project. It is more effective and

convenient than traditional methods; neural networks have broad application prospects in such nonlinear fields.

This model is mainly for the risk evaluation of railway macroconstruction. For specific railway project risks such as

construction risks and environmental risks, it is necessary to consider the actual target needs and actual risk indicators.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors of this manuscript do not have any conflicts of interest regarding the publication of this article.

Acknowledgments

This study is supported by Science and Technology Research and Development Project of China Railways Corporation (no. K2018T003); Intelligent High-Speed Rail Strategy Research (2035) of Chinese Academy of Engineering Consultative Project (no. 2018-ZD-05); Sichuan Provincial Science and Technology Support Project (nos. 2016GZ0338, 18MZGC0247, 18MZGC0186, and 2019JDRC0133); 2017-2019 Young Elite Scientist Sponsorship Program by CAST (YESS); Nanchang Railway Bureau Scientific Research Project (nos. 20171106 and 201710); Technology Research and Development Project of China Railway Eryuan Engineering Group Co. Ltd. (nos. KYY2017069(17-17) and KYY2019070(19-20)).

References

- [1] National Development and Reform Commission, Ministry of Foreign Affairs, and Ministry of Commerce, *Vision and action to promote the Silk Road Economic Belt and the 21st Century Maritime Silk Road*, vol. 3-29, No. 4, The People Daily, 2015.
- [2] A. Vangeli, "China's engagement with the sixteen countries of central, east and southeast Europe under the belt and road initiative," *China and World Economy*, vol. 25, no. 5, pp. 101-124, 2017.
- [3] E. Fardella and G. Prodi, "The belt and road initiative impact on Europe: an Italian perspective," *China and World Economy*, vol. 25, no. 5, pp. 125-138, 2017.
- [4] L. Zhang, H. Du, Y. Zhao, R. Wu, and X. Zhang, "Urban networks among Chinese cities along "the belt and road": a case of web search activity in cyberspace," *PLoS ONE*, vol. 12, no. 12, pp. 1-20, 2017.
- [5] Y. Zhao, J. Guo, X. Zhao, L. Tang, and Y. Wang, "The characteristics of city network along Belt and Road in China based on railway and aviation passenger transport," *Journal of Arid Land Resources and Environment*, vol. 31, no. 01, pp. 51-57, 2017.
- [6] W. Cui, H. Xuexi, Z. Shihong, and Z. Hongliang, "Risk evaluation model of agricultural high-tech investment project based on BP neural network," *Journal of Northwest Agriculture and Forestry University (Natural Science Edition)*, vol. 34, no. 7, pp. 160-164, 2006.
- [7] X. Wang, S. Feng, Y. Lei, and L. Yang, *43 case analysis of MATLAB neural network*, Beijing University of Aeronautics and Astronautics Press, 2013.
- [8] W. McCluskey, P. Davis, M. Haran, M. Mccord, and D. Mcilhatton, "The potential of artificial neural networks in mass appraisal: the case revisited," *Journal of Financial Management of Property and Construction*, vol. 17, no. 3, pp. 274-292, 2012.
- [9] T.-S. Quah and B. Srinivasan, "Improving returns on stock investment through neural network selection," *Expert Systems with Applications*, vol. 17, no. 4, pp. 295-301, 1999.
- [10] A. Saidi and M. Mirzaei, "Application of gold-labeled antibody biosensor in simultaneous determination of total aflatoxins using artificial neural network," *Journal of the Iranian Chemical Society*, vol. 11, no. 2, pp. 391-398, 2014.
- [11] G. Satyanarayana, G. Swami Naidu, and N. H. Babu, "Artificial neural network and regression modelling to study the effect of reinforcement and deformation on volumetric wear of red mud nano particle reinforced aluminium matrix composites synthesized by stir casting," *Boletín de la Sociedad Española de Cerámica y Vidrio*, vol. 57, no. 3, pp. 91-100, 2018.
- [12] L. L. Kien, C. Li-Lee, V. C. Chong et al., "Automated identification of copepods using digital image processing and artificial neural network," *BMC Bioinformatics*, vol. 16, no. S18, 2015.
- [13] G. B. José, H. J. Alfredo, A. M. Joel et al., "Estimation of umbilical cord blood leptin and insulin based on anthropometric data by means of artificial neural network approach: identifying key maternal and neonatal factors," *Bmc Pregnancy & Childbirth*, vol. 16, no. 1, p. 17, 2016.
- [14] A. N. Ran, *Research on Risk Evaluation of Logistics Park Construction Based on BP Neural Network*, ChangAn University, 2013.
- [15] A. B. Badiru and D. B. Sieger, "Neural network as a simulation metamodel in economic analysis of risky projects," *European Journal of Operational Research*, vol. 105, no. 1, pp. 130-142, 1998.
- [16] M. Chen, *Principles and Examples of MATLAB Neural Network*, Tsinghua University Press, 2013.
- [17] L. Zhang, K. Wu, Y. Zhong, and P. Li, "A new sub-pixel mapping algorithm based on a BP neural network with an observation model," *Neurocomputing*, vol. 71, no. 10-12, pp. 2046-2054, 2008.
- [18] F. Yu and X. Z. Xu, "A short-term load forecasting model of natural gas based on optimized genetic algorithm and improved BP neural network," *Applied Energy*, vol. 134, pp. 102-113, 2014.
- [19] B. H. M. Sadeghi, "A BP-neural network predictor model for plastic injection molding process," *Journal of Materials Processing Technology*, vol. 103, no. 3, pp. 411-416, 2000.
- [20] S. Wang, N. Zhang, L. Wu, and Y. Wang, "Wind speed forecasting based on the hybrid ensemble empirical mode decomposition and GA-BP neural network method," *Journal of Renewable Energy*, vol. 94, pp. 629-636, 2016.
- [21] C. Ren, N. An, J. Wang, L. Li, B. Hu, and D. Shang, "Optimal parameters selection for BP neural network based on particle swarm optimization: a case study of wind speed forecasting," *Knowledge-Based Systems*, vol. 56, pp. 226-239, 2014.
- [22] S. Ding and C. Su, "Application of optimizing BP neural networks algorithm based on genetic algorithm," in *Proceedings of the 29th Chinese Control Conference, CCC'10*, pp. 2425-2428, China, July 2010.
- [23] Z.-H. Guo, J. Wu, H.-Y. Lu, and J.-Z. Wang, "A case study on a hybrid wind speed forecasting method using BP neural network," *Knowledge-Based Systems*, vol. 24, no. 7, pp. 1048-1056, 2011.
- [24] W. K. Wong, C. W. M. Yuen, D. D. Fan, L. K. Chan, and E. H. K. Fung, "Stitching defect detection and classification using wavelet transform and BP neural network," *Expert Systems with Applications*, vol. 36, no. 2, pp. 3845-3856, 2009.

- [25] L. J. Guo, J. J. Gao, J. F. Yang, and J. X. Kang, "Criticality evaluation of petrochemical equipment based on fuzzy comprehensive evaluation and a BP neural network," *Journal of Loss Prevention in the Process Industries*, vol. 22, no. 4, pp. 469–476, 2009.
- [26] Z. Guoyi and H. Zheng, "Improved BP neural network model and its stability analysis," *Journal of Central South University*, vol. 42, no. 1, pp. 115–124, 2011.
- [27] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *Proceedings of the 2017 IEEE Symposium on Security and Privacy, SP 2017*, pp. 39–57, May 2017.
- [28] C. Melchiorre, M. Matteucci, and J. Remondo, "Artificial neural networks and robustness analysis in landslide susceptibility zonation," in *Proceedings of the International Joint Conference on Neural Networks 2006, IJCNN '06*, pp. 4375–4381, Canada, July 2006.
- [29] S. Ding and Q. Wu, "Research on robustness of BP neural network based inverse model for induction motor drives," in *Proceedings of the International Conference on Electronics and Optoelectronics (ICEOE '11)*, vol. 2, pp. V2127–V2131, July 2011.

Research Article

Stock Price Pattern Prediction Based on Complex Network and Machine Learning

Hongduo Cao , **Tiantian Lin**, **Ying Li** , and **Hanyu Zhang**

Business School, Sun Yat-sen University, Guangzhou 510275, China

Correspondence should be addressed to Hongduo Cao; caohd@mail.sysu.edu.cn and Ying Li; mnsliy@mail.sysu.edu.cn

Received 7 March 2019; Accepted 14 May 2019; Published 28 May 2019

Guest Editor: Benjamin M. Tabak

Copyright © 2019 Hongduo Cao et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Complex networks in stock market and stock price volatility pattern prediction are the important issues in stock price research. Previous studies have used historical information regarding a single stock to predict the future trend of the stock's price, seldom considering comovement among stocks in the same market. In this study, in order to extract the information about relation stocks for prediction, we try to combine the complex network method with machine learning to predict stock price patterns. Firstly, we propose a new pattern network construction method for multivariate stock time series. The price volatility combination patterns of the Standard & Poor's 500 Index (S&P 500), the NASDAQ Composite Index (NASDAQ), and the Dow Jones Industrial Average (DJIA) are transformed into directed weighted networks. It is found that network topology characteristics, such as average degree centrality, average strength, average shortest path length, and closeness centrality, can identify periods of sharp fluctuations in the stock market. Next, the topology characteristic variables for each combination symbolic pattern are used as the input variables for K-nearest neighbors (KNN) and support vector machine (SVM) algorithms to predict the next-day volatility patterns of a single stock. The results show that the optimal models corresponding to the two algorithms can be found through cross-validation and search methods, respectively. The prediction accuracy rates for the three indexes in relation to the testing data set are greater than 70%. In general, the prediction ability of SVM algorithms is better than that of KNN algorithms.

1. Introduction

Stock price volatility patterns classification and prediction is a very important problem in stock market research. The prediction of stock price trends is actually a classified prediction of stock price fluctuation patterns [1]. Literature showed that forecasting stock price patterns is sufficient to generate profitable trades and enable the execution of profitable trading strategies [2]. Therefore, many studies have focused on predicting stock price patterns rather than predicting the absolute prices of stocks [2–4].

To date, most studies have focused on the volatility patterns of a single stock based on its own historical attributes [5, 6] and have paid less attention to the comovement of related stocks and information pertaining to the overall market. A few studies have used historical information regarding related stocks as the input variables for prediction and shown that the price fluctuations in a single stock are not isolated and are often influenced by the trends of multiple related stocks [7, 8].

Thus, how to extract the comovement of multiple stocks and apply this information to the prediction of the fluctuation patterns of a single stock is a problem worth studying.

Complex network analysis provides a new explanation for stock market behavior from a systematic perspective. Using complex network theory to study stock prices not only allows us to analyze the relationship between different stocks, but also allows us to explore the macroaspects of the comovement characteristics of the market in different periods [9–11]. Previous studies have proposed a variety of methods to build complex networks using the time series of stock prices, including visibility graphs [12–14], recurrence networks [15–17], correlation networks [11, 18, 19], pattern networks [10, 20], and K-neighbors networks [21, 22]. Of all the network construction methods, the symbolic pattern network is favored by many scholars because it can more accurately reflect the degree of correlation and direction of the primitive elements in a complex system [10, 20, 23, 24]. In a stock price volatility pattern network, each volatility pattern is regarded

as a network node, and the relationship between patterns is regarded as a connection between nodes [10]. By analyzing the topological properties of the network, the characteristics of stock price fluctuations can be better understood. Huang et al. used coarse-grained symbolization methods to construct a network of market prices and transaction volume data in different periods based on the Shanghai Stock Exchange (SSE) composite index, and the results showed that the out-degree distribution of network nodes obeyed the power law and the basic fluctuations exhibited different patterns during different periods [24]. Wang et al. converted the yields of gasoline and crude oil stocks into five patterns and studied the characteristics of crude oil and gasoline node networks in different periods using sliding windows and then accurately predicted the crude oil and gasoline stock price pattern based on the conversion characteristics of the price network [10, 20].

However, most of the existing studies on stock price volatility pattern networks have focused on univariate time series. On this basis, we propose a new network construction method to build the volatility pattern networks of the three most important indexes in the US stock market, namely, the Standard & Poor's 500 Index (S&P 500), the NASDAQ Composite Index (NASDAQ), and the Dow Jones Industrial Average (DJIA). Firstly, the combination symbolic patterns for the three stock indexes are derived using a coarse-grained method. Then, the combination symbolic patterns are used as the nodes of the network, and the frequencies and directions of the conversion of the patterns are used as the weights and directions of the network connections. Finally, we construct directed and weighted networks for the US stock market. By analyzing the network topology properties, we can identify periods of sharp fluctuations in the market.

Meanwhile, many machine learning algorithms have been applied to stock price volatility classification and prediction, such as neural networks [25], random forests [26], decision trees [27], support vector machines (SVM) [3, 7], and K-nearest neighbors (KNN) [1, 28]. Among them, K-nearest neighbors (KNN) and support vector machine (SVM) algorithms have been widely used in pattern recognition and forecasting, machine learning, information retrieval, and data mining. KNN is a simple and effective classification method that is easy to calculate and its performance is comparable to the most advanced classification methods [29, 30]. SVM, which can map nonlinear separable data into high-dimensional space and use hyperplanes for classification, is highly suitable for small sample classification because of its excellent classification ability [26]. Both KNN and SVM algorithms have a mature theoretical basis in relation to classification prediction. Ballings et al. also compared the accuracy of SVM, KNN, and other algorithms in predicting stock price movements one year ahead for 5767 publicly listed European companies, and the results showed that SVM has the better prediction ability than KNN [2]. Teixeira proposed an automatic stock trading method that combined technical analysis with KNN classification. Using 15 stocks from Sao Paulo Stock Exchange (Bovespa), they found that the proposed method generated considerably higher profits than the buy-and-hold method for most of the companies, with few buy actions generated [1]. Huang et al. used SVM algorithms

to predict the weekly fluctuations in the Nikkei 225 index and found that SVM outperformed the other classification methods, such as quadratic discriminant analysis and Elman backpropagation neural networks [3].

Literature has demonstrated the ability of SVM and KNN to predict stock patterns. However, they predicted the stock price based on the information of the single stock itself, without considering the information of the network system composed of the relevant stocks. Therefore, another aim of this study is to predict the next-day pattern of a single stock for each combination mode of stocks using the network topology properties as input variables for SVM and KNN algorithms. To the best of our knowledge, this should be the first attempt in existing research. Then, we compare the prediction accuracy using the testing data set after identifying the best models using the training set. The stock price volatility pattern network includes price information for single stocks and related stocks and portrays the macronature of the market, which contains more information than is available using only historical information relating to single stocks. The results show that the pattern network can provide some information to enable us to forecast the price volatility patterns of single stocks. Of the two prediction methods, the optimal parameter search strategy combined with cross-validation and search methods enables us to find the models that perform well on the testing data set. Overall, the performance of SVM algorithms is better than that of KNN algorithms. Combining with complex network and machine learning can provide investors with information on profitability strategies.

The remainder of this paper is organized as follows. In the next section, we introduce the theoretical background for KNN and SVM algorithms. In Section 3, the methodology of constructing the network and of predicting the next-day patterns for each stock index is presented. In Section 4, we show the empirical results and compare the prediction accuracy for KNN and SVM. The last section is devoted to a summary.

2. Theoretical Background for KNN and SVM

2.1. KNN. K-nearest neighbor (KNN) algorithm is a non-parametric classification algorithm that assigns query data to be classified to the category to which most of its K neighbors belong [31]. We use the Euclidean distance metric to find K -nearest neighbors from a sample set of known classifications. Suppose that the known data set has four feature variables $\{f_1, f_2, f_3, f_4\}$ and four categories $\{y_1, y_2, y_3, y_4\}$. The steps to search the category of the new data i through the KNN algorithm are as follows.

Firstly, the Euclidean distance of the feature variables of the data i and the other data j ($j = 1, 2, \dots, n$) in the training data set is calculated:

$$D_j = \sqrt{\sum_{g=1}^4 (f_g(i) - f_g(j))^2}, \quad j = 1, 2, \dots, n. \quad (1)$$

Secondly, all the data in the training set are sorted in ascending order according to the distance from data i .

Thirdly, K data points with the smallest distance from data i are selected.

Finally, the category with the largest proportion of these K data points will be considered as the category of data i .

An important parameter to be determined in the KNN algorithm is K , which represents the number of the nearest neighbors to be considered when classifying unknown samples [1, 2].

2.2. SVM. SVM was introduced by Vapnik [32] and has been widely used in pattern prediction in recent years. The basic idea of SVM is to nonlinearly transform the input vector into a high-dimensional feature space and then search the optimal linear classification surface in this feature space to maximize the distance between the classification plane and the nearest point. The training samples closest to the classification plane are called support vectors. SVM algorithm can be briefly described as follows.

Consider the binary linear classification problem of training data set (\mathbf{x}_i, y_i) ($i = 1, 2, \dots, n$), $\mathbf{x}_i \in R^n$; $y_i \in \{\pm 1\}$, where \mathbf{x}_i is a feature vector and y_i is a class label. Suppose these two classes can be separated by a linear hyperplane $(\mathbf{w} \cdot \mathbf{x}) + b = 0$. In order to make the correct classification and get the largest classification interval, the optimization problem of constructing the optimal plane is described as

$$\begin{aligned} \min \quad & \varphi(\mathbf{w}) = \frac{1}{2} \|\mathbf{w}\|^2 = \frac{1}{2} (\mathbf{w}' \cdot \mathbf{w}), \\ \text{s.t.} \quad & y_i [(\mathbf{w} \cdot \mathbf{x}_i) + b] \geq 1. \end{aligned} \quad (2)$$

The optimal solution of \mathbf{w} and b can be solved by introducing the Lagrange multiplier. Then we can obtain the optimal classification problem like (3).

$$f(x) = \text{sgn} \{ \mathbf{w}^* \cdot \mathbf{x} + \mathbf{b}^* \}. \quad (3)$$

For a nonlinear classification problem, the feature vector is transformed into high-dimensional space vector firstly. Then the optimal classification hyperplane is constructed. Suppose the transformation function is Φ , then the optimal problem can be described as

$$\begin{aligned} \min \quad & \varphi(\mathbf{w}) = \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^N \xi_i = \frac{1}{2} (\mathbf{w}' \cdot \mathbf{w}) + C \sum_{i=1}^N \xi_i, \\ \text{s.t.} \quad & y_i [(\mathbf{w}^T \cdot \Phi(\mathbf{x}_i)) + b] + \xi_i \geq 1, \\ & \xi_i \geq 0, \quad i = 1, 2, \dots, n, \end{aligned} \quad (4)$$

where C is the penalty parameter, which specifies the trade-off between classification distance and misclassification [2]. Finally, the optimal classification hyperplane can be described in (5).

$$f(x) = \text{sgn} \{ \mathbf{w}^* \cdot \Phi(\mathbf{x}) + \mathbf{b}^* \}. \quad (5)$$

$$k(\mathbf{x}_i, \mathbf{x}_j) = \Phi^T(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j). \quad (6)$$

The function (6) is called a kernel function. Because the performance of the Gaussian radial basis function (RBF)

is excellent when the additional information of the data is limited, it is widely used in the financial time series analysis [3]. The Gaussian radial basis function (RBF) is used as the kernel function to implement the SVM algorithm in this study. The RBF kernel function can be expressed as

$$k(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\sigma \|\mathbf{x}_i - \mathbf{x}_j\|^2\right), \quad \sigma > 0, \quad (7)$$

where σ is the constant of the radial basis function. Before implementing the SVM algorithm, the parameter σ and parameter C need to be determined.

For multiclassification problem, it can be converted into multiple two-classification problems [33]. In this study, a four-classification problem is transferred into six two-classification problems by the ‘‘one-versus-one’’ approach of SVM.

3. Methodology

In this section, we introduce the methodology for predicting stock price patterns using network topology characteristic variables. Figure 1 shows a general framework of the proposed pattern prediction system. It consists of two parts: complex network analysis and pattern prediction using machine learning. We present a more detail procedure in the subsections.

3.1. Constructing a Pattern Network for the Stock Market. Using the daily closing price of each stock index, a sliding window is used to calculate the one-day return r , five-day return R , and five-day volatility V corresponding to day t :

$$r = \ln \frac{\text{Close}(t)}{\text{Close}(t-1)}, \quad (8)$$

$$R = \ln \frac{\text{Close}(t)}{\text{Close}(t-5)}, \quad (9)$$

$$V = \text{std}(r_1, \dots, r_5) * \sqrt{5}, \quad (10)$$

where $\text{Close}(t)$ is the closing price on day t , $\text{Close}(t-1)$ is the previous day's closing price, and $\text{std}(r_1, \dots, r_5)$ is the standard deviation of the yield from the first to the fifth day.

Then, we can calculate the average five-day volatility \bar{V} for each stock index:

$$\bar{V} = \frac{1}{N} \sum V, \quad (11)$$

where N is the number of trading days in a time series. Suppose we study M stocks in the stock market. Then, we can obtain the average volatility for the overall market as follows:

$$V' = \frac{1}{M} \sum_{i=1}^M \bar{V}_i. \quad (12)$$

Based on the sign of the five-day return R and the magnitude of the five-day volatility V each day, each stock can be classified into one of four patterns:

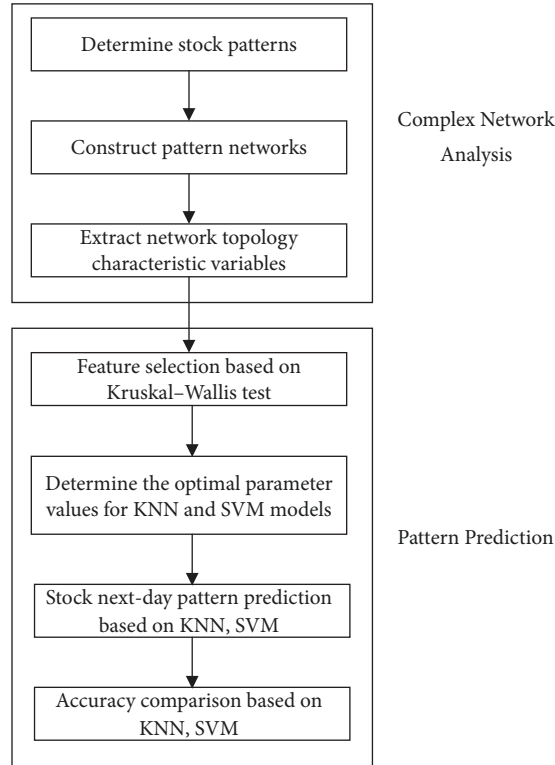


FIGURE 1: General framework of the stock index pattern prediction.

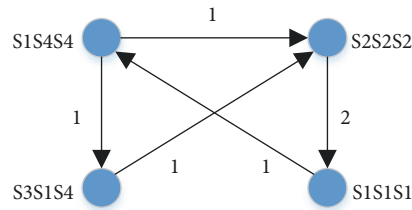


FIGURE 2: Sample directed weighted network.

$$f = \begin{cases} S1, & \text{if } R \geq 0 \text{ and } V \geq V' \text{ (sharp rise)} \\ S2, & \text{if } R \geq 0 \text{ and } V < V' \text{ (stable rise)} \\ S3, & \text{if } R < 0 \text{ and } V < V' \text{ (stable decline)} \\ S4, & \text{if } R < 0 \text{ and } V \geq V' \text{ (sharp decline)}. \end{cases} \quad (13)$$

By combining the patterns of each stock index, we can obtain the corresponding combination symbolic pattern for each day. Assuming that we study three stock indexes, we can obtain a maximum of $4^3 = 64$ combination modes. Taking the daily combination patterns as the nodes of the network, the edges and weights of the network can be determined in time order. If the pattern on day t is $S1S4S4$ and that on day $t+1$ is $S2S2S2$, there is a directed edge from $S1S4S4$ to $S2S2S2$ with a weight of 1. If the conversion frequency from $S1S4S4$ to $S2S2S2$ is w , the weight of the directed edges from $S1S4S4$ to $S2S2S2$ is w . For example, if the current patterns of the S&P 500, NASDAQ, and DJIA are $S1$, $S3$, and $S4$, respectively,

the current price combination pattern is $S1S3S4$. Suppose that the pattern transformation over a certain period of time is $\{S1S4S4, S2S2S2, S1S1S1, S1S4S4, S3S1S4, S2S2S2, S1S1S1\}$. Then, we can obtain the directed weighted network shown in Figure 2.

The key to the sliding window selection problem is how to effectively keep the quality and quantity of original time series information while reducing the computational complexity to the most extent [34]. In this study, we apply a sliding window with a length of 30 days (about one month in daily life and half a quarter in the stock market) and a step of one day to the stock indexes time series. So we can obtain a pattern network every 30 days. Table 1 shows the process of using the sliding window.

3.2. Computing Network Topology Characteristic Variables. Next, we calculate the network topology characteristic variables for every 30-day network.

TABLE 1: The process of using the sliding window.

Date	Stock 1	Stock 2	Stock 3	Combination Pattern
1	S1	S2	S3	S1S2S3
2	S2	S2	S2	S2S2S2
3	S3	S3	S3	S3S3S3
4	S1	S1	S1	S1S1S1
...
30	S1	S1	S2	S1S1S2
31	S2	S2	S2	S2S2S2
32	S3	S4	S4	S3S4S4
...

3.2.1. Network Average Degree Centrality. In undirected networks, the average degree centrality of the network reflects the level of connection between one node and other nodes in the network, that is, whether one node is connected with the other nodes or not [17]. The formula is as follows:

$$\rho = \frac{1}{N(N-1)} \sum_{i,j=1}^N a_{ij}, \quad (14)$$

where N is the number of nodes in the network and $a_{i,j}$ is the value of the adjacency matrix of an undirected network. $a_{i,j} = 1$ if node i and node j are connected, otherwise $a_{i,j} = 0$. The adjacency matrix of an undirected network is a symmetric matrix. However, $a_{i,j} = 1$ does not mean that $a_{j,i} = 1$ in a directed network. In the directed network, we must consider the out-degree and in-degree. We connect the nodes in time order so that the in-degree and the out-degree are the same except for the first node and the last node. Therefore, we only select the in-degree for analysis, and calculate the average in-degree centrality as follows:

$$\rho = \frac{1}{N^2} \sum_{j=1}^N \sum_{i=1}^N a_{ij}. \quad (15)$$

In terms of narration, in the next sections, we describe average in-degree centrality as average degree centrality. Average degree centrality measures the ratio of the actual number of connections to the maximum number of connections, that is, the edge density of the network. The greater the average degree centrality, the more connections between nodes in the price pattern network, the higher the accessibility between nodes, and the greater the density of the overall network [35].

3.2.2. Average Network Strength. In a network, the strength of the connection from node i to node j is the weight w_{ij} of the directed edge from node i to node j . Similar to the in-degree and out-degree of the directed network, the strength of the directed weighted network can also be divided into in-strength and out-strength [10, 35]. In this study, we describe average out-strength as average network strength:

$$S = \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^N w_{ij}. \quad (16)$$

The greater the average strength of the network, the fewer the number of network nodes, the simpler the composition of the price volatility patterns, the smaller the complexity of the network, and the higher the frequency of the same node. The simpler price patterns reflect the fact that the consistency of price changes of different stocks is stronger and lasts for longer.

3.2.3. Network Average Shortest Path Length. The average shortest path length of the network describes the degree of separation between nodes in the network, that is, the size of the network. The average shortest path length can be used to characterize a “small-world” network in a complex network [17]. The distance from node i to node j is defined as the minimum number of edges needed to pass from node i to node j . The average shortest path length of the network is the average length of all the shortest paths in the network:

$$L = \frac{1}{N(N-1)} \sum_{i=1}^N \sum_{j=1}^N d_{ij}, \quad (i \neq j). \quad (17)$$

The shorter the average shortest path, the less intermediate patterns are required for conversion between stock price modes. Modes can be connected by fewer edges, and price modes can interact with each other through fewer other modes. As a result, the conversion efficiency and speed of the overall network are both greater.

3.2.4. Network Closeness Centrality. The closeness centrality of node i is the reciprocal of the average shortest path length from other nodes to node i [15, 36]:

$$C_i = \frac{N-1}{\sum_{j=1}^N d_{ji}}. \quad (18)$$

The closer a point is to other points, the easier it is to transmit information. Now, we consider the weighted shortest path l_{ij} , which is defined as the shortest weighted distance from node i to node j . Then, we can obtain the closeness centrality of the network [37]:

$$C = \frac{1}{N} \sum_{i=1}^N \frac{n-1}{N-1} \frac{n-1}{\sum_{j=1}^{n-1} l_{ij}}, \quad (i \neq j). \quad (19)$$

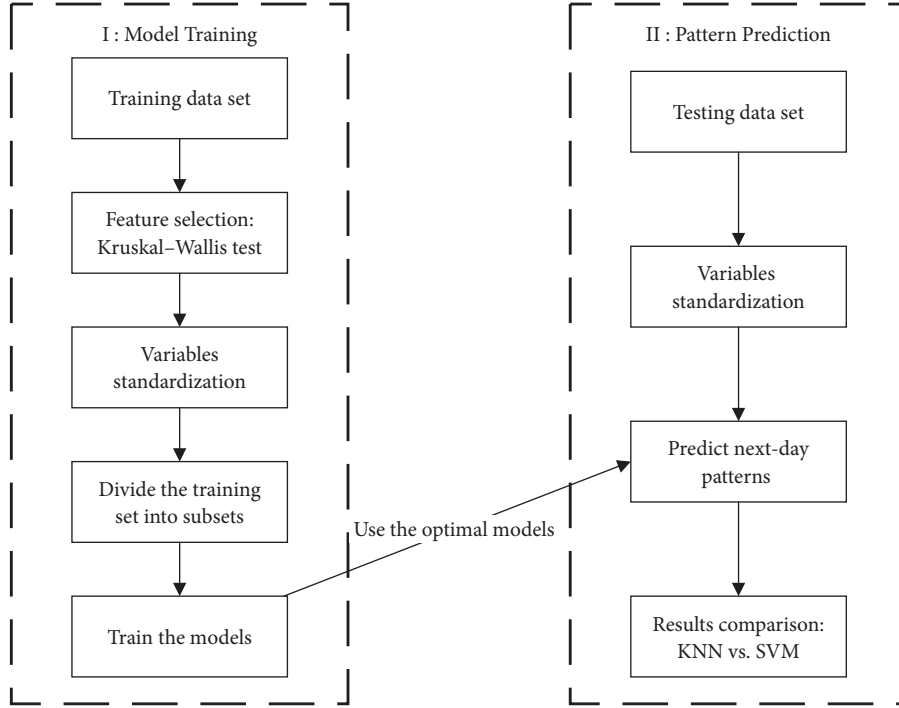


FIGURE 3: The road map of the pattern prediction experiment.

where n is the number of nodes reachable from node i . The greater the closeness centrality of the network, the smaller the shortest weighted distance between the network nodes, the fewer the conversion times between different price modes, and the smaller the conversion cycle. Pattern nodes tend to transform on their own, and thus the transformation area of the overall network is more concentrated and centrality is more prominent [35].

3.3. Next-Day Stock Pattern Prediction Based on KNN and SVM. We train the optimal prediction models based on KNN and SVM algorithms by the obtained network topology characteristic variables, and then predict next-day patterns of three single stock indexes using the testing data set.

3.3.1. Detail Prediction Procedure. Figure 3 shows the detail process of the pattern prediction experiment for each single stock index. It includes two major steps: the first step is model training, from which we can get the best models; and the second step is pattern prediction.

First, the correlation between the second day's stock patterns and the network characteristic variables of a single stock index in the training set is tested by the Kruskal-Wallis test. The network topological characteristic variables which are significantly correlated with the price pattern of each stock index will be the input variables for the stock index prediction.

The values of topological characteristic variables are normalized so that a smaller valued indicator does not be ignored

because of an indicator with larger value [26]. Formula (20) is used to standardize the variables [38]:

$$\hat{y}_i = \frac{y_i - y_{min}}{y_{max} - y_{min}}. \quad (20)$$

Next, in order to get a subset of each combination mode, we divide the training set into several training subsets according to the number of types of combination patterns (or combination pattern nodes) in the training set. Obviously, in each subset, the day's combination pattern is the same, but the next-day patterns of each stock index can be different. For each training subset, the next-day patterns of a single stock index are the classification variables, and the network topological characteristic variables are the input feature variables for the KNN and SVM algorithms. To prevent overfitting, the cross-validation and search method is used to determine the optimal parameters in this study.

Finally, on the testing data set, the next-day patterns of each stock index are predicted by the obtained models.

The optimal training models are found according to the recognized combination pattern, and the next-day stock patterns are predicted on the basis of the topological characteristic variables of the current corresponding 30-day network. The average market volatility and the standardization parameters used for the testing set were obtained through the training set.

3.3.2. Model Selection Criteria. Cross-validation is widely used for model selection because of its simplicity and universality, so we use cross-validation method and search

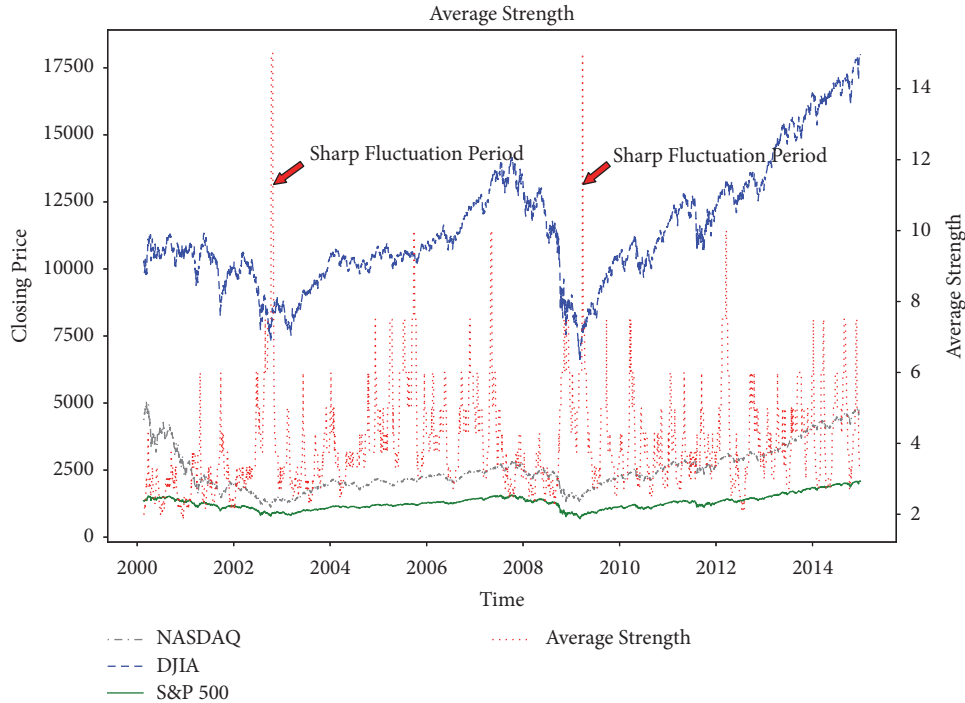


FIGURE 4: The evolution of the three stock indexes and average strength.

method to determine the optimal parameters in this study [39]. We cross-validated the K -parameters for KNN by trying all values of $K = \{1, 2, 3, \dots, 30\}$. To determine the optimal parameter values for SVM, we perform a grid search on $C = \{10^{-2}, 10^{-1}, 10^0, 10^1, 10^2\}$, $\sigma = \{10^{-2}, 10^{-1}, 10^0, 10^1, 10^2\}$ to identify the best combination.

Using the k -fold cross-validation method, for each combination of parameters, the training data set is divided into a subset with k equal parts, and $k-1$ parts of the data are used as the training data, while the other part is used as the verification data. In this way, the accuracy rates of k verification sets can be obtained after k iterations. Taking the average accuracy rate of the k verification sets, that is the verification score, as the criterion for parameter selection, the optimal combination of parameters can be found. In addition, leave-one-out (LOO) is another simple, efficient, and common cross-validation method. When using LOO, one sample is taken from the data set each time as a validation set, and the other samples are used as the training set. Thus, for a data set with n samples, we can get a total of n different test sets and their corresponding training sets. LOO is very suitable for model selection of small samples because only one sample is extracted from the training set at a time as a verification set so that fewer samples are wasted [40].

In this study, we use 3-fold cross-validation if the training subset contains more than 100 samples and use LOO cross-validation otherwise.

4. Empirical Results and Analysis

4.1. Data Processing. We used the closing prices of the S&P 500, NASDAQ, and DJIA from 1 January 2000 to 31

December 2014 as the sample data set. This resulted in 3769 daily records. The data were obtained from the Wind database, one of the most authoritative financial database in China (the Wind database can be downloaded from <https://www.wind.com.cn/>). First, the five-day return rate and five-day volatility of each stock index are calculated. The method outlined in the previous section is used to symbolize the stock index, and then we obtain the combination patterns for the three stock indexes each day.

A sliding window with a length of 30 days and a step of one day is used to divide the stock pattern time series into 3740 time periods. A directed weighted network is constructed for each period of stock price patterns, resulting in 3740 networks. There are 47 pattern nodes in all of the networks.

The method we construct the stock networks is original, so we used Python for coding. We used some functions in the Python 3.7 standard library including networkx, sklearn, pandas, and matplotlib for our analysis.

4.2. Analysis of Network Topological Characteristics. The average degree centrality, average intensity, average shortest path length, and closeness centrality of each network are calculated using formulas (15), (16), (17), and (19). The evolution of these four network topological characteristics is shown in Figures 4–7.

As can be seen from the figures, the points where average degree centrality, average intensity, and closeness centrality reach their peak value and the average shortest path length reaches its minimum value all correspond to periods when the overall market is fluctuating wildly. When

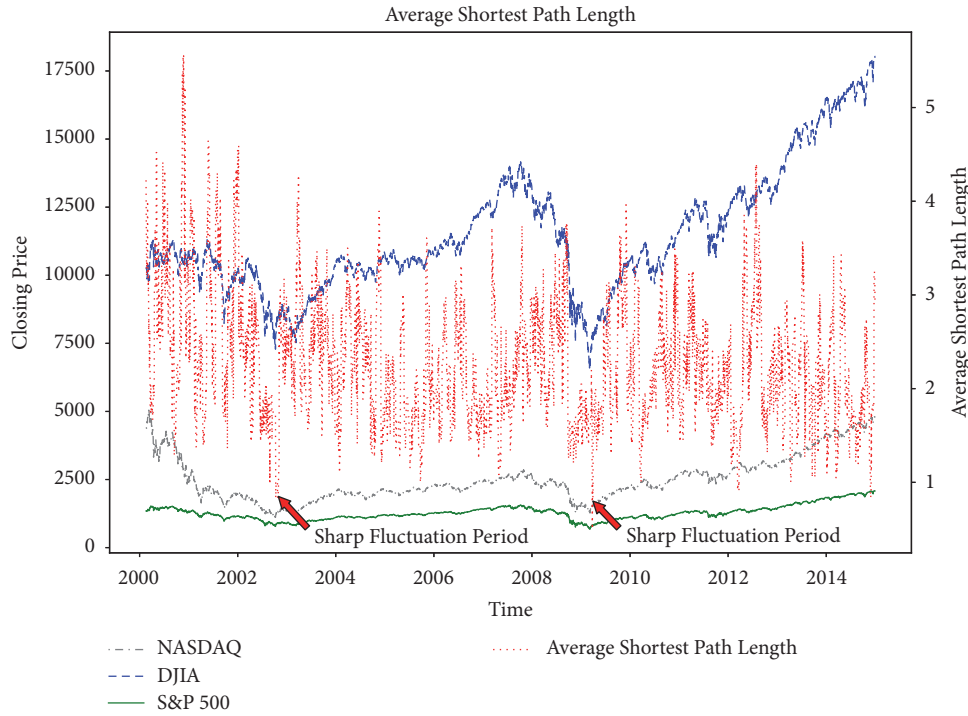


FIGURE 5: The evolution of the three stock indexes and average shortest path length.

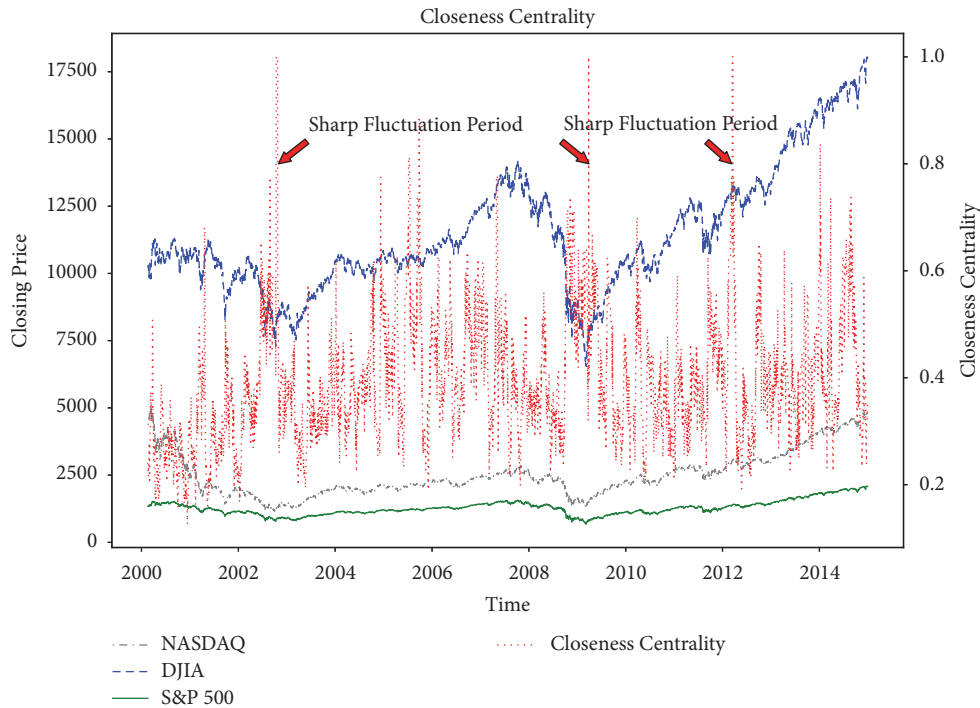


FIGURE 6: The evolution of the three stock indexes and closeness centrality.

the three stock indexes fell to their lowest levels in October 2002 and March 2009, the average degree centrality, average strength, and closeness centrality of the network reached their highest points, while the shortest path length reached its lowest point. These two periods correspond to the last phase

of the dot-com bubble crisis and the subprime mortgage crisis. In addition, the closeness centrality and the average degree centrality reached their maximum points again in March 2012, which corresponded with another long period of sharp fluctuations in the US stock market. The results

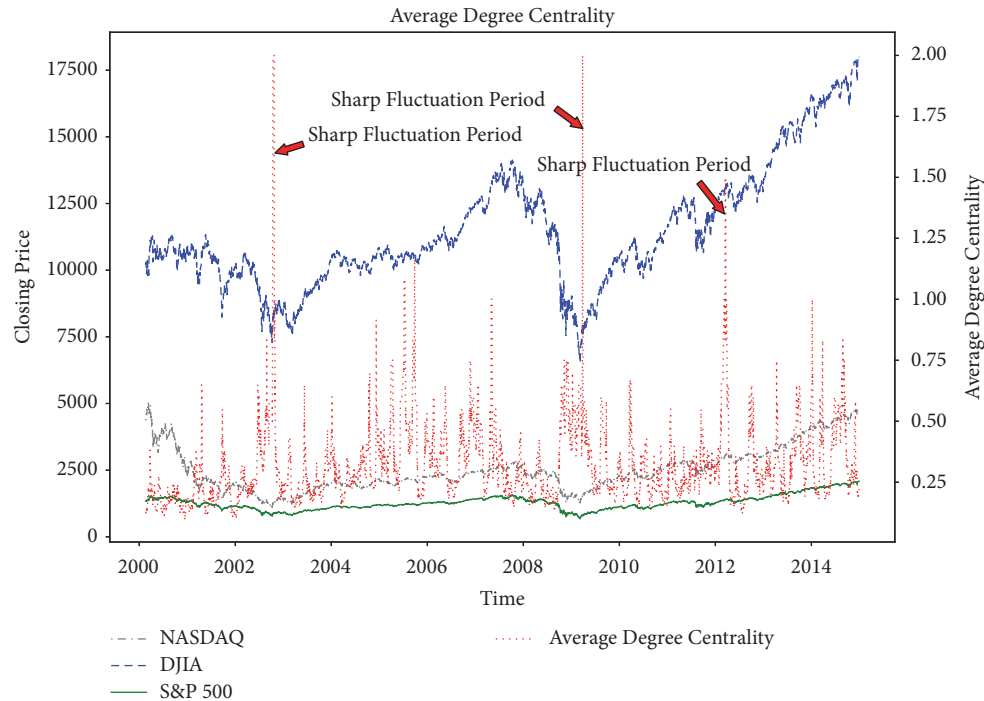


FIGURE 7: The evolution of the three stock indexes and average degree centrality.

show that these four network topological characteristics have a remarkable relationship to the anomaly of the indexes in the US stock market. During the sharp fluctuation periods, the comovement of three stock indexes is stronger, and the 30-day networks are simpler.

The maximum value of the average strength of the network reflects the fact that there are relatively fewer nodes in the stock pattern network, and the fluctuation modes of stocks are monotonous. It shows that in the month before the extreme values, the price fluctuations of the three stock indexes were synchronous, resulting in the relative simplicity of the volatility combination pattern. Before the three indexes reached their lowest levels in 2002 and 2009, they were basically in a state of substantial decline. The correlation and consistency of the indexes reached their maximum during this period, and so the average strength of the network reached its maximum, which is consistent with the findings of previous studies on complex networks using price time series [9, 41].

In addition, when the stock market is in a period of dramatic fluctuations, the node types in the network are monotonous, the stock network is constantly switching between several price models, and the edge density is larger, so the average degree centrality reaches its maximum. During this period, the nodes are more compact, the transformation between nodes is faster, fewer edges need to be passed, and the average shortest path of the network reaches its minimum. Although other modes may emerge during this period, the large fluctuation mode occupies the most important position, and the price patterns tend to shift between the main patterns so that reaching maximum closeness centrality. This conclusion is the same as that of Wang et al. [10].

4.3. Next-Day Stock Pattern Prediction Using KNN and SVM Algorithms. By analyzing the 30-day network topological characteristics corresponding to each trading day, we find that the extreme values of the network topological characteristics can reflect the periods of dramatic fluctuations in the system composed of the three stock indexes. KNN and SVM algorithms are used to predict the next-day patterns of each stock index when the combination patterns of the three stock indexes and the corresponding 30-day network topological characteristics for the current day are known.

Based on the theory of cross-validation [42], and in order to keep the year intact and ensure the continuity of the years, we used the closing prices of the S&P 500, NASDAQ, and DJIA from 1 January 2000 to 31 December 2014 as the training sample data set. The testing sample data set used the closing prices of the three indices from 1 January 2015 to 31 December 2017. The training set and the testing set contained 3769 and 755 records, respectively. Since the training sample set has 47 pattern nodes, we divided the training data into 47 training subsets.

4.3.1. Kruskal–Wallis Tests to Filter Variables. Following the methods used to select variables in existing studies, we used the Kruskal–Wallis test to filter the four network topological characteristic variables and next-day patterns of each stock index using the training samples [43]. The results are shown in Table 2.

It can be seen from Table 2 that the p-values of the variables and the next-day patterns of the various stock indexes are all less than or equal to 0.1 except for the next-day patterns of the DJIA, where closeness centrality is not significant. Therefore, when predicting the next-day patterns

TABLE 2: Kruskal–Wallis tests of the four network topological characteristic variables and next-day patterns of each stock index.

	S&P500 stock index	NASDAQ stock index	DJIA stock index
Average strength	43.0578* * * (0.001)	340.2039* * * (0.001)	29.8488* * * (0.001)
Average shortest path length	9.3241** (0.0253)	28.1597* * * (0.001)	7.6535* (0.0537)
Average degree centrality	25.4071* * * (0.001)	250.9374* * * (0.001)	13.0833* * * (0.0045)
Closeness centrality	8.4532** (0.0375)	190.0352* * * (0.001)	0.6204 (0.8917)

Note: figures in parentheses are p-values.

TABLE 3: Prediction accuracy of the optimal KNN and SVM models in relation to next-day patterns for the three stock indexes using the testing set.

Closeness	Algorithms	DJIA stock index	S&P500 stock index	NASDAQ stock index
Yes	KNN	74.83%	72.58%	72.45%
	SVM	74.97%	73.11%	74.57%
No	KNN	72.98%	70.59%	70.86%
	SVM	76.03%	72.45%	73.64%

of the DJIA, the closeness centrality is removed, leaving the three other variables as input variables for the KNN and SVM algorithms. When predicting the next-day patterns of the S&P 500 and the NASDAQ, all four network topological characteristic variables are retained as input variables.

4.3.2. *Predicting Stock Patterns Using KNN and SVM Algorithms.* The accuracy of prediction is defined as:

$$\text{Accuracy rate} = \frac{\text{The number of correct prediction}}{\text{Total number of sample set}} \times 100\% \quad (21)$$

We compare the predicted next-day patterns with the actual next-day patterns of the stock index. If they are the same on a given day, we can say that our prediction is correct. The proportion of the number of correctly predicted samples to the total number of samples is the accuracy rate. The accuracy rate is close to 1 means that the models yield more accurate predictions, whereas the accuracy rate is close to 0 means that the models are less accurate.

After obtaining the optimal models using KNN and SVM algorithms in relation to the training set using the cross-validation and search methods, the models are used to predict patterns using the testing set, and their performance is evaluated based on their prediction accuracy rates. Table 3 shows the prediction accuracy of the optimal models obtained using KNN and SVM algorithms for the three stock indexes.

From Table 3, it can be seen that KNN and SVM algorithms can identify appropriate models based on the training set using the cross-validation and search methods, with prediction accuracies in relation to the testing set of greater than 70%. However, generally, the prediction accuracy of SVM algorithms is higher than that of KNN algorithms. It is similar to the findings of previous studies on the two

algorithms; that is, the generalization ability of the SVM classification model is greater than that of the KNN model [1, 2]. To further illustrate the predictive effect of closeness on the three stock indexes, we compare the prediction accuracy rate in the cases of closeness and no closeness. We find an interesting result wherein the prediction model without closeness using SVM has the highest prediction accuracy rate when predicting the next-day patterns of the DJIA stock index. This result indicates that SVM is more accurate and sensitive than KNN. Closeness does not affect predictions regarding the DJIA, as the results of the Kruskal–Wallis test show.

Diether et al. examined short-selling in US stocks using SEC-mandated data for 2005 and found that short-selling activity was strongly positively correlated with previous five-day returns and volatility [44]. The five-day movement of stocks is also very important for short-term investment in stocks or funds in the real world. Thus, if the investor can forecast future five-day volatility patterns, more information can be obtained to support short-selling strategies. For instance, if the next five-day pattern is predicted to be S1 (sharp rise), the investor can execute a short-selling strategy the next day.

5. Conclusion

Based on the complex network method, this study analyzes the stock price fluctuation patterns of the three most important stock indexes for the US stock market. Unlike previous studies, this study uses the three stock indexes to build pattern networks for the system, rather than using a single stock index. From the analyses of the average strength, average shortest path length, average degree centrality, and closeness centrality of the price pattern network every 30 days, it is found that when the overall stock market is in a period of

dramatic fluctuations, the average strength, average degree centrality, and closeness centrality reach their maximum values, while the average shortest path length reaches its minimum value. This shows that price volatility pattern networks can reflect special periods on the stock market. In periods of dramatic fluctuations on the stock market, the comovement of various indexes is stronger, the edge density of the corresponding pattern network is greater, the conversion between price modes is faster, and the conversion area of nodes is more concentrated. It shows the validity of using price pattern network characteristics to identify special periods on the stock market. To a certain extent, they can reflect abnormal periods on the stock market from a macropoint of view. When the four indicators approach extreme values, investors should exercise caution.

The stock price network characteristic variables not only contain price change information for individual stocks, but also reflect the overall change characteristics of the market at the macrolevel. Therefore, another focus of this study is the use of the network characteristic variables as input variables for KNN and SVM algorithms to predict the next-day fluctuation patterns of individual stocks. Firstly, the Kruskal–Wallis test is used to test the next-day patterns of three stock indexes and four network characteristic variables, and we find that closeness does not affect predicting the next-day stock patterns of DJIA index. In the case of the combination price patterns for the current day, the network characteristic variables are used as the input variables for KNN and SVM algorithms to predict the next-day stock price patterns, and the accuracy of the two algorithms in relation to the testing set is compared. The results show that both the KNN and SVM algorithms display a high level of accuracy in predicting the next-day stock price patterns, with prediction accuracy of greater than 70% for all three stock indexes. However, the generalization ability of the SVM algorithm is greater than that of the KNN algorithm. Thus, it is possible to predict stock trends by using a proper classification algorithm and combining the structural characteristics of the multistock price network. This approach can generate more information for financial trading strategies in the real world and provides a new focus for future research into stock price prediction.

The application of the complex network method to the stock market is still in the developmental stage. Revealing the characteristics of stock price fluctuations by using complex networks is helpful in understanding the essence of stock price fluctuations and providing profit-making strategies. A more detailed examination of the correlation between financial crises and network topological properties is a worthy topic for further research. In addition, the combined use of machine learning and complex network methods to study the stock market deserves more in-depth discussion and diversified development.

Data Availability

The data used to support the findings of this study have been deposited in <https://www.kesci.com/home/dataset/5c82706ed635ff002ca24a19>.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported, in part, by the National Natural Science Foundation of China (Grants nos. 71371200, 71071167, and 71071168).

References

- [1] L. A. Teixeira and A. L. I. De Oliveira, “A method for automatic stock trading combining technical analysis and nearest neighbor classification,” *Expert Systems with Applications*, vol. 37, no. 10, pp. 6885–6890, 2010.
- [2] M. Ballings, D. Van Den Poel, N. Hespeels, and R. Gryp, “Evaluating multiple classifiers for stock price direction prediction,” *Expert Systems with Applications*, vol. 42, no. 20, pp. 7046–7056, 2015.
- [3] W. Huang, Y. Nakamori, and S.-Y. Wang, “Forecasting stock market movement direction with support vector machine,” *Computers & Operations Research*, vol. 32, no. 10, pp. 2513–2522, 2005.
- [4] Y.-W. Cheung, M. D. Chinn, and A. G. Pascual, “Empirical exchange rate models of the nineties: Are any fit to survive?” *Journal of International Money and Finance*, vol. 24, no. 7, pp. 1150–1175, 2005.
- [5] G. Armano, A. Murru, and F. Roli, “Stock market prediction by a mixture of genetic-neural experts,” *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 16, no. 5, pp. 501–526, 2002.
- [6] A. H. Moghaddam, M. H. Moghaddam, and M. Esfandyari, “Stock market index prediction using artificial neural network,” *Journal of Economics, Finance and Administrative Science*, vol. 21, no. 41, pp. 89–93, 2016.
- [7] R. Choudhry and K. Garg, *A Hybrid Machine Learning System for Stock Market Forecasting*, vol. 39, 2008.
- [8] Y. K. Kwon, S. S. Choi, and B. R. Moon, “Stock prediction based on financial correlation,” in *Proceedings of the Conference on Genetic Evolutionary Computation*, 2005.
- [9] L. Lacasa, V. Nicosia, and V. Latora, “Network structure of multivariate time series,” *Scientific Reports*, vol. 5, 2015.
- [10] M. Wang, Y. Chen, L. Tian, S. Jiang, Z. Tian, and R. Du, “Fluctuation behavior analysis of international crude oil and gasoline price based on complex network perspective,” *Applied Energy*, vol. 175, pp. 109–127, 2016.
- [11] B. M. Tabak, T. R. Serra, and D. O. Cajueiro, “Topological properties of stock market networks: the case of Brazil,” *Physica A: Statistical Mechanics and Its Applications*, vol. 389, no. 16, pp. 3240–3249, 2010.
- [12] L. Lacasa, B. Luque, F. Ballesteros, J. Luque, and J. C. Nuno, “From time series to complex networks: the visibility graph,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 13, pp. 4972–4975, 2008.
- [13] Z. Gao, Q. Cai, Y. Yang, W. Dang, and S. Zhang, “Multiscale limited penetrable horizontal visibility graph for analyzing nonlinear timeseries,” *Scientific Reports*, vol. 6, no. 1, Article ID 35622, 2016.
- [14] E. Zhuang, M. Small, and G. Feng, “Time series analysis of the developed financial markets’ integration using visibility graphs,”

- Physica A: Statistical Mechanics and its Applications*, vol. 410, pp. 483–495, 2014.
- [15] R. V. Donner, Y. Zou, J. F. Donges, N. Marwan, and J. Kurths, “Recurrence networks—a novel paradigm for nonlinear time series analysis,” *New Journal of Physics*, vol. 12, no. 3, Article ID 033025, 2010.
- [16] Y. Li, H. Cao, and Y. Tan, “Novel method of identifying time series based on network graphs,” *Complexity*, vol. 17, no. 1, pp. 13–34, 2011.
- [17] N. Marwan, J. F. Donges, Y. Zou, R. V. Donner, and J. Kurths, “Complex network approach for recurrence analysis of time series,” *Physics Letters A*, vol. 373, no. 46, pp. 4246–4254, 2009.
- [18] Y. Li, H. Cao, and Y. Tan, “A comparison of two methods for modeling large-scale data from time series as complex networks,” *AIP Advances*, vol. 1, no. 1, Article ID 012103, p. 509, 2011.
- [19] Y. Yang and H. Yang, “Complex network-based time series analysis,” *Physica A: Statistical Mechanics and Its Applications*, vol. 387, no. 5–6, pp. 1381–1386, 2008.
- [20] M. Wang, A. L. M. Vilela, L. Tian, H. Xu, and R. Du, “A new time series prediction method based on complex network theory,” in *Proceedings of the 5th IEEE International Conference on Big Data, (Big Data) 2017*, pp. 4170–4175, December 2017.
- [21] Y. Shimada, T. Kimura, and T. Ikeguchi, *Analysis of Chaotic Dynamics Using Measures of the Complex Network Theory*, Springer, Berlin, Germany, 2008.
- [22] X. Xu, J. Zhang, and M. Small, “Superfamily phenomena and motifs of networks induced from time series,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 50, pp. 19601–19605, 2008.
- [23] Z. Ming, W. Er-Hong, Z. Ming-Yuan, and M. Qing-Hao, “Directed weighted complex networks based on time series symbolic pattern representation,” *Acta Physica Sinica*, vol. 66, no. 21, Article ID 210502, 2017.
- [24] W.-Q. Huang, S. Yao, and X.-T. Zhuang, “A network dynamic model based on SSE composite index and trading volume fluctuation,” *Journal of Northeastern University*, vol. 31, no. 10, pp. 1516–1520, 2010.
- [25] S. H. Kim and S. H. Chun, “Graded forecasting using an array of bipolar predictions: Application of probabilistic neural networks to a stock market index,” *International Journal of Forecasting*, vol. 14, no. 3, pp. 323–337, 1998.
- [26] J. Patel, S. Shah, P. Thakkar, and K. Kotecha, “Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques,” *Expert Systems with Applications*, vol. 42, no. 1, pp. 259–268, 2015.
- [27] M.-C. Wu, S.-Y. Lin, and C.-H. Lin, “An effective application of decision tree to stock trading,” *Expert Systems with Applications*, vol. 31, no. 2, pp. 270–274, 2006.
- [28] M. V. Subha and S. T. Nambi, “Classification of stock index movement using k-nearest neighbours (k-NN) algorithm,” *WSEAS Transactions on Information Science and Applications*, vol. 9, no. 9, pp. 261–270, 2012.
- [29] C. G. Atkeson, A. W. Moore, and S. Schaal, “Locally weighted learning,” *Artificial Intelligence Review*, vol. 11, no. 1–5, pp. 11–73, 1997.
- [30] C.-J. Huang, D.-X. Yang, and Y.-T. Chuang, “Application of wrapper approach and composite classifier to the stock trend prediction,” *Expert Systems with Applications*, vol. 34, no. 4, pp. 2870–2878, 2008.
- [31] S. Mitra and T. Acharya, *Data Mining: Multimedia, Soft Computing, and Bioinformatics*, John Wiley & Sons, 2005.
- [32] V. Cherkassky, “The nature of statistical learning theory,” *Technometrics*, vol. 38, no. 4, pp. 409–409, 1996.
- [33] W.-M. Lin, C.-H. Wu, C.-H. Lin, and F.-S. Cheng, “Classification of multiple power quality disturbances using support vector machine and one-versus-one approach,” in *Proceedings of the 2006 International Conference on Power System Technology, POWERCON2006*, China, October 2006.
- [34] F. Li and J. Xiao, “How to get effective slide-window size in time series similarity search,” *Journal of Frontiers of Computer Science & Technology*, vol. 3, no. 1, pp. 105–112, 2009.
- [35] X. Sun, M. Small, Y. Zhao, and X. Xue, “Characterizing system dynamics with a weighted and directed network constructed from time series data,” *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 24, no. 2, Article ID 024402, 9 pages, 2014.
- [36] L. C. Freeman, “Centrality in social networks conceptual clarification,” *Social Networks*, vol. 1, no. 3, pp. 215–239, 1978–1979.
- [37] A. W. Wolfe, “Social network analysis: methods and applications,” *American Ethnologist*, vol. 24, no. 4, pp. 136–137, 1995.
- [38] K. Shin, T. S. Lee, and H. Kim, “An application of support vector machines in bankruptcy prediction model,” *Expert Systems with Applications*, vol. 28, no. 1, pp. 127–135, 2005.
- [39] S. Arlot and A. Celisse, “A survey of cross-validation procedures for model selection,” *Statistics Surveys*, vol. 4, pp. 40–79, 2010.
- [40] G. C. Cawley and N. L. C. Talbot, “Efficient leave-one-out cross-validation of kernel fisher discriminant classifiers,” *Pattern Recognition*, vol. 36, no. 11, pp. 2585–2592, 2003.
- [41] L. Xia, D. You, X. Jiang, and Q. Guobc, “Comparison between global financial crisis and local stock disaster on top of Chinese stock network,” *Physica A Statistical Mechanics Its Applications*, vol. 490, Article ID S0378437117307227, pp. 222–230, 2017.
- [42] Z. Zhou and Y. Yu, *Machine Learning and Its Application 2011*, Tsinghua University Press, Beijing, China, 2009.
- [43] X. Wang, H. Xue, and W. Jia, *Prediction Model of Stock's Rosing and Felling Based on BP Neural Network*, Value Engineering, 2010.
- [44] K. B. Diether, K.-H. Lee, and I. M. Werner, “Short-sale strategies and return predictability,” *Review of Financial Studies*, vol. 22, no. 2, pp. 575–607, 2009.

Research Article

Big Data Market Optimization Pricing Model Based on Data Quality

Jian Yang ¹, Chongchong Zhao,¹ and Chunxiao Xing²

¹School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China

²Research Institute of Information, Beijing National Research Center for Information Science and Technology, Department of Computer Science and Technology, Institute of Internet Industry, Tsinghua University, Beijing 100084, China

Correspondence should be addressed to Jian Yang; yangjian2015@xs.ustb.edu.cn

Received 3 February 2019; Accepted 7 April 2019; Published 23 April 2019

Guest Editor: Thiago Christiano Silva

Copyright © 2019 Jian Yang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

In recent years, data has become a special kind of information commodity and promoted the development of information commodity economy through distribution. With the development of big data, the data market emerged and provided convenience for data transactions. However, the issues of optimal pricing and data quality allocation in the big data market have not been fully studied yet. In this paper, we proposed a big data market pricing model based on data quality. We first analyzed the dimensional indicators that affect data quality, and a linear evaluation model was established. Then, from the perspective of data science, we analyzed the impact of quality level on big data analysis (i.e., machine learning algorithms) and defined the utility function of data quality. The experimental results in real data sets have shown the applicability of the proposed quality utility function. In addition, we formulated the profit maximization problem and gave theoretical analysis. Finally, the data market can maximize profits through the proposed model illustrated with numerical examples.

1. Introduction

With the rapid development of information technology, big data has become the core resource of all walks of life. Government departments, research institutions, IT companies, financial institutions, etc. have generated massive amounts of data during operations. In addition, due to the rise of mobile networks and smart terminals, a large proportion of people now have smart phones with sensors, which can easily collect data beyond the past possible range using GPS, cameras, microphones, etc. The storage and calculation of big data are no longer the sole purpose. By using data mining and machine learning to analyze data, it provides an opportunity to bring about breakthroughs in processing video, images, and speech [1]. Unfortunately, only a small amount of data is currently being fully utilized and its use is limited as well. The reuse of these data can create huge commercial value, which is the true meaning of big data. Therefore, in order to make profits and provide data utilization, data can be resold to other organizations [2].

Marketplaces are enablers for the exchange of data. Therefore, data trading has become an innovative business model that has driven the advent of DT (Data Technology) era. In this era, data has become an important asset for companies, from the exclusive internal data to the sharing between companies. However, due to the lack of standardized data sharing channels and unified transaction specifications, big data trading platforms and data markets have emerged as the times require in this context.

Nowadays, data products and related services are increasingly being provided to the online data market, which carries the data publisher's data and provides it to data consumers. Figure 1 [3] presents an intuitive description of the formation and flow of data products. Firstly, the initial seller is the original data provider. For example, Xignite [4] sells financial data, Gnip [5] publishes social network data, and Factual [6] deals with geographical data. Secondly, the data market provides a centralized management platform for data providers to upload, store, and sell data in order to support online transactions of the data. The current

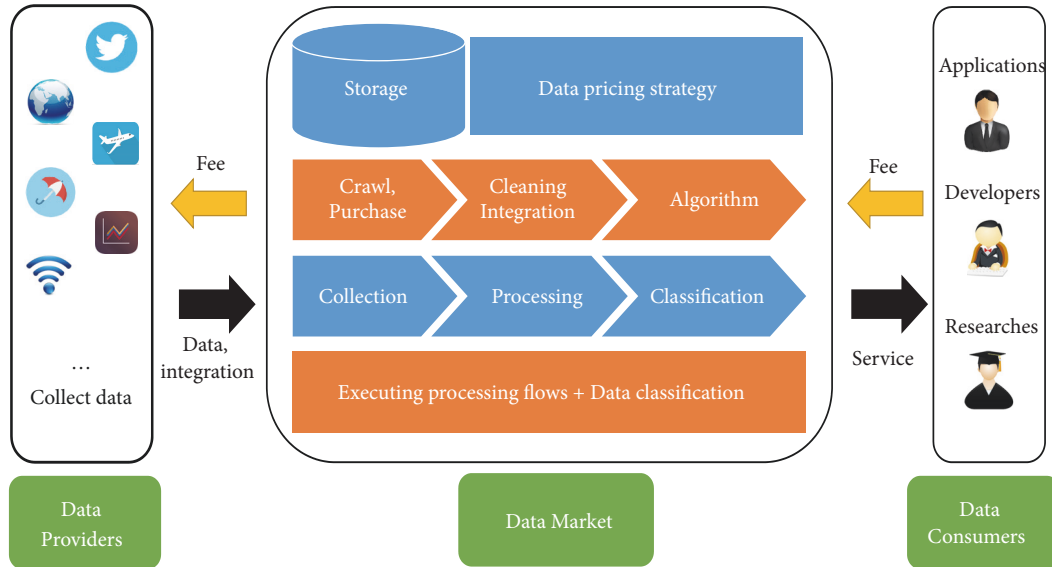


FIGURE 1: A typical big data market model.

data platforms include Factual, Infochimps [7], Xignite, and Windows Azure Data Marketplace [8]. The last is the terminal buyer, which is the consumer of big data products. There are generally three types of consumers who have demand for big data products, i.e., enterprises, government departments, and research institutions. These consumers need the data and corresponding services provided by the online market in order to innovate products, optimize decisions, or conduct research.

However, the big data market has not formed a unified pricing mechanism yet, and various pricing strategies are still not perfect; i.e., different data markets offer different pricing mechanisms. Currently, the major pricing mechanisms in the data market include subscription, bundling, and discrimination. However, the impact of data quality on the pricing mechanism has rarely been studied. Many literatures [9, 10] indicate that data quality is very important for the evaluation of data value. Hence, in this paper, we have proposed a pricing model based on quality utility to optimize data market pricing.

The key contributions of this paper can be summarized as follows:

- (i) We first summarized several dimensional indicators that affected data quality and established a linear model to calculate the quality scores. Based on this, a hierarchical division method of the square root of the quality score is proposed.
- (ii) We proposed a utility model based on the quality level and verified it with real-world datasets, using machine learning algorithms. The results have proved the applicability of this utility model.
- (iii) From the perspective of economics, we considered the consumers' willingness to pay and formulated an optimized pricing scheme based on the quality utility function. Numerical experiments have shown that

the owners of data platform can maximize profits by determining the quality level and subscription fee.

The remainder of this paper is organized as follows: Section 2 reviews the related work. Then, the dimension of data quality and the method of level division are presented in Section 3. Section 4 describes the utility function of data quality and the suitability of the model is verified by machine learning algorithm. Section 5 formulates the profit maximization problem and gives theoretical analysis. Section 6 presents and analyzes the numerical experimental results. Section 7 concludes the paper.

2. Literature Review

The valuation of intangible assets, such as cloud computing services [11–13] and network information services [14, 15], is not a new challenge for practitioners and researchers. Relevant scholars have done a lot of work on the pricing of information products and services.

Before studying data pricing, we first review the representative work of these methods. The information service market usually involves three commonly used pricing mechanisms:

- (1) Subscription mechanism: Windows Azure Data Market [8] is a decent example for subscription pricing scheme. Azure has monthly subscriptions of two types, limited and unlimited. Balasubramanian et al. [16] consider the difference between usage frequency associated with payment model and consumer psychological cost. They believe that the two pricing mechanisms of information products, i.e., fixed cost and pay-per-payment mechanisms, may affect the profit of the seller.
- (2) Bundling pricing: this strategy originates from capital data market, and it represents an aggregation technique [17]. In the capital data market, data vendors

bundle multiple types of products in accordance with certain strategies and allocate different prices for them to be selected by heterogeneous consumers. Niyato et al. [18], considering smart data pricing methods to solve the problem of IoT data management, adopt a binding strategy that allows multiple providers to form alliances and provide bundled services to attract more users and achieve higher revenues.

- (3) Version control pricing mechanism: the strategy is a widespread differentiation strategy used in information-product markets. Wei et al. [19] inspect the versioning strategy where consumers differ in individual tastes for quality. They found that if groups have mutually exclusive characteristics, they are the values associated with the shared features; then, versioning strategy is optimal. Li et al. [20] defined a nonlinear function to describe the “willingness to pay” and the utility of consumers with specific quality requirements and developed a hybrid steady state evolution algorithm. They observed that monopolies can obtain more profits by using multiversion strategy.

There are also some scholars who have studied the pricing model of data products from different perspectives. Koutris et al. [21] studied query-based pricing, and they designed a pricing algorithm that satisfies no-arbitrage and no-discount allowing the price of any query to be exported automatically. Shen et al. [22] proposed a big personal data pricing model based on tuple granularity. By investigating the data attributes that affect the value of data, this model is proposed to implement a positive rating and reverse pricing for big personal data. By dynamically adjusting the model parameters, the users can enjoy improved benefits. Yang et al. [23] studied the pricing model of personal privacy data. They proposed a framework to compensate for privacy loss. This method can compensate for privacy loss based on user’s preferences and allow users to control their data through financial means.

Through extensive review of the literature, we can conclude that existing data pricing literature either investigates published data pricing methods or studies new approaches that focus on relevance and privacy. Data quality is a key factor affecting data assessment and has been ignored so far.

In the entire data life cycle, such as data creation, transformation, transmission, and application, each stage may cause various data quality problems. Liu et al. [24] summarized the problems faced by current big data research in data collection, processing, and analysis, namely, the collection of unreal data, information incompleteness, consistency, and reliability. In [25], there are a total of 21 quality standards. Ding et al. [26] summarize relevant quality dimensions and review their applicability to the data market.

Data quality is characterized by multidimensionality and complexity. Therefore, in this paper, we consider an optimized pricing model based on data quality, hoping to

provide data platform owners with useful pricing decision recommendations.

3. Data Value Evaluation Based on Quality

When the data market owner wants to sell data at a reasonable price, the first thing to consider is to evaluate the value of data. On the one hand, data value can be measured by the size of data [27], on the other hand, it can be measured based on the quality of data. In this paper, we evaluate data from the perspective of data quality. First, we introduce different dimensions of data quality. Then, we establish a linear model based on these dimensions to evaluate data value. Finally, we adopt the square root mapping function and divide the quality level.

3.1. Dimensions of Data Quality. Data quality includes multiple dimensions. The measurement of dimensions will vary according to the type of data, so quality has to be evaluated using the criteria that the data has to comply with. In [28, 29], the applicability of the quality dimension to the data market has been reviewed, especially the concept of version control, i.e., the data seller creates different quality versions of the data product to suit the needs and tastes of heterogeneous consumers. Literature [30] summarizes seven quality standards, which were expressed as $Q_d = \{accuracy, completeness, redundancy, data\ volume, latency, response\ time, timeliness\}$. These quality dimensions allow continuous versioning. In other words, we can create any number of quality levels based on them. For simplicity, only three measures in Q_d are all scaled in interval $[0, 1]$ and will be demonstrated here in detail, i.e., *accuracy*, *completeness*, and *redundancy*. Table 1 [31] contains the metrics, report names, and description definitions for each quality attribute and lists the calculation formulas.

Several other quality dimensions also have their calculation methods. However, due to space limit, we omit them from the paper.

3.2. Data Quality Level Division. Creating a universal data quality assessment standard can be an arduous task for all types of data. Without loss of generality, a linear model is presented as below, but other options may exist.

$$\begin{aligned} \text{Qualityscore} &= w_1 * \text{accuracy} + w_2 * \text{completeness} \\ &+ \dots + w_n * \text{redundancy} \quad (1) \\ \text{s.t.} \quad w_1 + w_2 + \dots + w_n &= 1 \end{aligned}$$

where w_1, w_2, \dots, w_n are related weight factors, which can be set by users in practice.

We adopt the method of dividing the quality level in [30]. In this paper, the quality score is defined in interval $(0, 1)$, and we first scale it to the sector of the appropriate function domain $[s_{min}, s_{max}]$, e.g., $[0, 100]$. Then, since the square root function can produce more reasonable level intervals, we adopt the square root of the quality score to rank on the basis of the previous step. For instance, a domain of $[0, 100]$ and quality levels of $Q_l = 10$, as done in this paper,

TABLE 1: Metric definitions, description, and calculation.

Attributes	Metric	Description	Variables	Formula
Accuracy	Proportion of accurate cells	Indicate the proportion cells in a data source that has correct value according to the domain and the type of information of the data source.	n_{ce} : Number of cells with errors n_{cl} : Number of cells	$pac=1-\frac{n_{ce}}{n_{cl}}$
Completeness	Proportion of complete cells	Indicate the proportion of complete cells in a dataset. It means the cells that are not empty and have a meaningful value assigned (i.e., a value coherent with the domain of the column).	n_r : Number of rows n_c : Number of columns i_c : Number of incomplete cells n_{cl} : Number of cells	$n_{cl} = n_r * n_c$ $pcc=1-\frac{i_c}{n_{cl}}$
Redundancy	Proportion of duplicate records	Redundancy expresses the proportion of duplicate records in the data source. Since this factor is the cost-indicator, we convert it to the benefit-indicator.	n_r : Number of rows red : Number of duplicate records	$pdc=1-\frac{red}{n_r}$

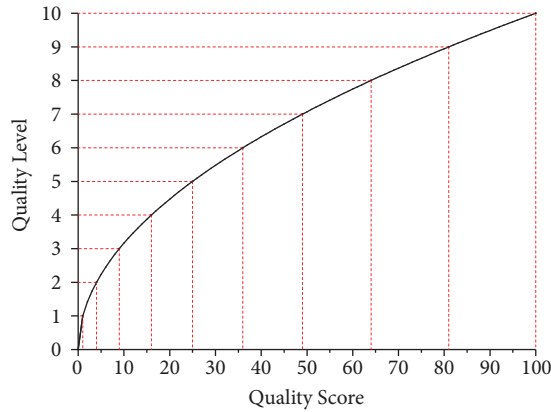


FIGURE 2: Mapping of quality scores and levels.

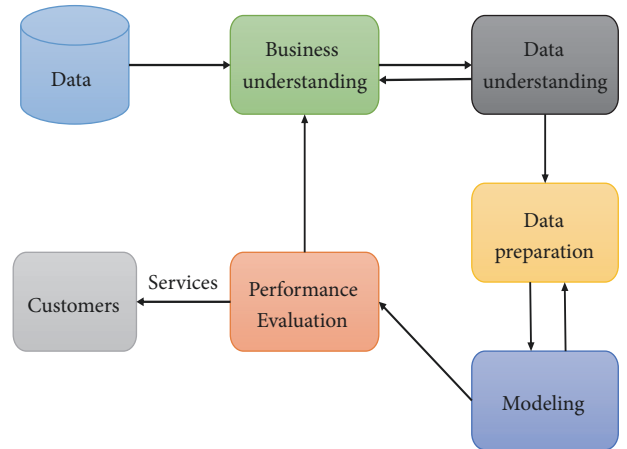


FIGURE 3: Big data business intelligence service.

is that examples are more illustrative. Figure 2 presents such a mapping relationship.

4. The Utility of Data Quality

4.1. Utility Functions. In current big data business applications, it is usually big data sets that adopt model-based methods to extract knowledge and information to solve complex business applications. Figure 3 shows the process of big data business intelligence.

It can be seen that data plays an important role in the entire business analysis. The quality of data directly determines the accuracy of the machine learning model [32] and ultimately affects business decisions. In order to explain this certain phenomenon, the usefulness of quality must be measured on a new scale. Therefore, the usefulness of quality is $U(q)$, which is the utility of quality. According to the experience of machine learning and data mining, under the condition of the same amount of data, the higher quality information is input into the classifier, the better the

classification effect will be. Therefore, this utility function can be considered as the quality of the model. For example, the utility is the accuracy of classifying input into a discrete-value output.

We suppose that a utility function $U(q)$ has the following three basic properties:

- (1) $U(q)$ is nonnegative.
- (2) $U(q)$ is an increasing function of q .
- (3) $U(q)$ is a concave function of q .

Usually we assume that the function $U(q)$ is nonnegative and twice differentiable; then, (2) and (3) state that $U'(q) > 0$ and $U''(q) < 0$.

The first attribute is rational as quality utility cannot be negative. The second attribute is the obvious requirement that the higher the quality, the better. Several reasons are given for the third property. One way to justify it is to require that the marginal utility $U'(q)$ is a decreasing function [33] of data quality q .

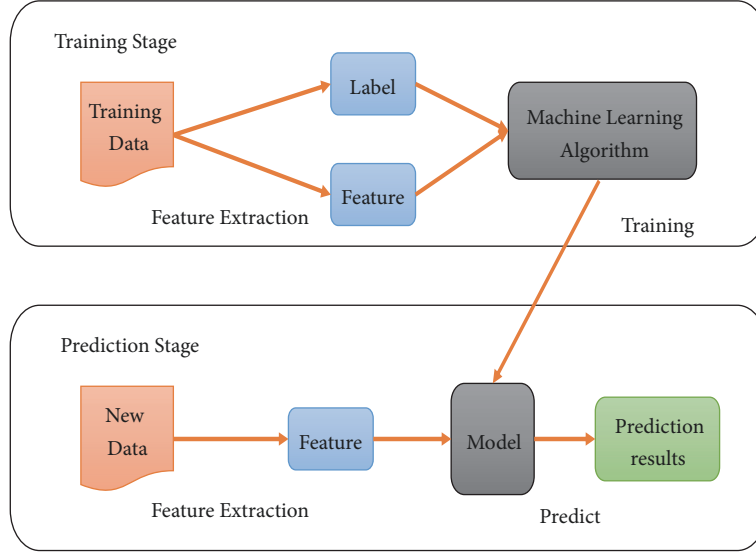


FIGURE 4: A basic machine learning workflow.

4.2. *Estimating Utility Functions.* In order to determine the utility function of data quality in big data analysis, we consider the study from the perspective of classification-based machine learning.

Next, we describe the process of classification. For a data set D , it can be expressed as a $k \times l$ matrix \bar{X} , where each row corresponds to an item, and the first l elements of each row correspond to the l property values of the item. A machine learning task can be divided into two phases, as shown in Figure 4.

(i) *Training Stage.* The data is subjected to feature extraction to generate data features and prediction targets (*Label*) and then trained by machine learning algorithms to generate the model.

(ii) *Predicting Stage.* Input testing dataset: after feature extraction, produce data features, using the trained model to make predictions and finally producing prediction results.

As shown in (2). A is a label column, which is the real category attribute. The last column is the category predicted by the classifier, denoted by Pr . Evaluating a classifier can be judged by minimizing the error between A and Pr , i.e., $\min \sum_{i=1}^k \|A - Pr\|^2$.

$$\bar{X} = \begin{matrix} & X_1 & X_2 & \cdots & X_l & A & Pr \\ m_1 & \left(\begin{array}{cccccc} x_{11} & x_{12} & \cdots & x_{1l} & \check{y}_1 & \hat{y}_1 \\ x_{21} & x_{22} & \cdots & x_{2l} & \check{y}_2 & \hat{y}_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ x_{k1} & x_{k2} & \cdots & x_{kl} & \check{y}_k & \hat{y}_k \end{array} \right) & \end{matrix} \quad (2)$$

Suppose that the classification accuracy for each item m_i is defined as a . But beyond that, we also assume the utility, i.e., accuracy $U = ACC$. To estimate the utility function, we use data of different quality levels

during the model training. Specifically, the experimental point $(q_{s1}, a_1), \dots, (q_{sj}, a_j), \dots, (q_{sk}, a_k)$ is a nondecreasing sequence, where q_{sj} is the quality level of the corresponding data and satisfies $q_{sj} \leq q_{s(j+1)}$. These points are then used to find a set of optimal parameters of the utility function $U(q_s; \beta)$ by nonlinear least squares, where β is an optimal parameter. The optimal parameter of the utility function $U(q_s; \beta)$ by minimizing the sum of square errors is as follows:

$$\min \sum_{j=1}^k \|ACC_j - U(q_{sj}; \beta)\|^2 \quad (3)$$

In this paper, for simplicity, we consider the following exponential-based utility function:

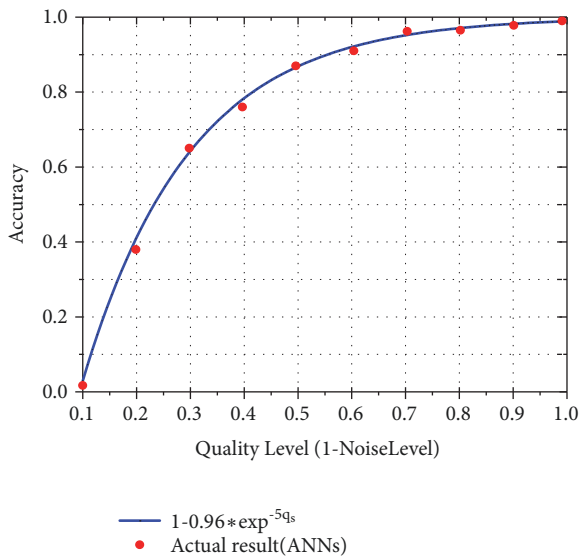
$$\begin{aligned} U(q_s; \beta = [\beta_1, \beta_2, \beta_3]) &= \beta_1 - \beta_2 \exp(\beta_3 q_s) \\ \text{s.t. } \beta_1, \beta_2 &> 0 \\ \text{and } \beta_3 &< 0. \end{aligned} \quad (4)$$

where q_s is the quality level and β_1 , β_2 , and β_3 are the curve fitting parameters of the utility function to real-world experiments, i.e., the ground truth. In order to find a utility function that satisfies the corresponding condition, we can adjust the parameter β , so that the sum of the squared errors between the experimental and the estimated points is minimized.

4.3. *Experimental Evaluation Based on Real Datasets.* In order to prove the rationality of the proposed utility function, we use a real dataset called MNIST [34], which contains a variety of hand-written digital pictures and contains the labels for each picture. We use Artificial Neural Networks (ANNs) model for classification training. ANNs use nonlinear mathematical equations to successively develop meaningful relationships between input and output variables through a learning process. Specifically, we applied convolutional neural network (*cnn*) for classification training.

TABLE 2: Frequently used notations.

Notation	Description
q_s	The quality level of the data products and services
p	Subscription fees for data products and services
WTP	Customers' Willingness to Pay
η	Customer sensitivity to quality level
$\mathcal{E}(q_s, p)$	Profit resulting from the separate sales of the data product and service under p and q_s
$U(q_s, \beta)$	Data utility with curve fitting parameter p and quality level q_s
N	The number of customers willing to buy a data product or service
c	The unit price of the data quality
$\mathcal{L}(\cdot)$	Lagrangian of the profit function $U(q_s, \beta)$

FIGURE 5: Accuracy trends under different quality levels ($\beta_1 = 1$, $\beta_2 = 0.96$, $\beta_3 = -5$).

Due to the multidimensionality and complexity of data quality, it would be a difficult task if all quality dimensions were taken into account to classify quality levels. Our goal is to illustrate the effect of different quality levels of a given data on model classification capabilities. For simplicity, and in order to reflect our motivation, in the experimental design stage, we draw on the experience of the concept of signal-noise ratio (SNR) [35] in the electronic information field. Specifically, we use the method of adding noise to the label data to express the effect of different quality levels on the accuracy of the model. In the experiment, we assume that the original MNIST training set and testing set labels are all noise-free. Use the following steps to add noise to the label:

- (i) Select K samples from N total samples according to the given noise ratio NoiseLevel, $K = N * \text{NoiseLevel}$.
- (ii) For each sample of the selected K samples, replace its original label with a random number between 0 and 9 except the original label.

The quality level is the inverse image of the noise level. For simplicity, Figure 5 shows the trend of accuracy at different

levels of quality. Obviously, as the quality level increases, the accuracy also increases, and the higher the quality level, the smaller the increase in accuracy. The accuracy of growth is getting smaller and smaller. Further, the proposed utility function can well fit the actual accuracy result and rationalize the concave function. It also facilitates the derivation of optimal pricing, which will be described in the next section.

5. Optimal Pricing

In this section, we first analyzed consumers' willingness to pay from the perspective of consumer behavior. Then, we introduced the profit maximization model with data quality level. Finally, the closed-form solutions of the subscription fee and quality level were derived and proved to be globally optimal. The key notations and description used throughout the paper were defined in Table 2.

5.1. Customers' Willingness to Pay. Every consumer in the market has personal preferences and interests. They make purchasing decisions based on their own needs, preferences, and prices by a self-selection process. This self-selection is described by a consumer's Willingness To Pay (WTP) [36]. WTP refers to the price that a customer is willing to pay in order to purchase a certain number of data products or services. This price is also referred to the customer's reservation price. In other words, they are willing to pay the highest price of the product. We assume that the data platform knows the customer's willingness to pay obeys the probability distribution, and the data platform is faced with a choice dilemma, i.e., loss of customer because of high price or consumer surplus due to low price. Each arriving consumer has a specific subjective price for a certain product, i.e., reservation price, and only if the consumer's reservation price is greater than the value of the product, the customer will purchase it.

We assume customers' sensitivities of quality level by $\eta = \{\eta_1, \eta_2, \dots, \eta_M\}$, which is randomly distributed from 0 to 1. Note that the higher the quality of data provided, the more the willingness of customers to pay for the data product, which is

$$\frac{\partial WTP}{\partial U(q_s)} > 0 \quad (5)$$

where $U(q_s)$ is the data quality utility function mentioned in Section 4. Customers who would like to obtain the best experience need to pay more. Assume the WTP function is linear, which is

$$WTP = \eta U(q_s) \quad (6)$$

5.2. Profit Function of Data Platform. In Section 1, we described a typical big data market model. The data platform purchases raw data from the data publishers and pays for the data providers. The data platform needs to process, convert, and store the collected data, or to establish application-level services (e.g., business analysis, visualization). This results in fixed costs (purchased from raw data) and variable costs (data processing, deep processing), which are collectively referred to costs in this paper. The data platform can set the subscription fee based on the quality level of the provided data and service to determine its profit maximization. The data consumer decides whether or not to purchase according to their willingness to pay and consumption. We assume that the probability density of all customers' willingness to pay is $f(p)$, and its cumulative distribution function is $F(p)$, which indicates the probability that consumers' willingness to pay is less than the value of products, i.e., the probability that the customer is unwilling to purchase the product. Then, the expected profit of the data platform is computed as follows:

$$\begin{aligned} \mathcal{E}(q_s, p) &= pNPr - cq_s = pN \int_0^{WTP} f(w) dw - cq_s \\ &= pN \left(\int_0^p f(w) dw + \int_p^{WTP} f(w) dw \right) \\ &- cq_s = pN (F(p) + WTP - p) - cq_s \end{aligned} \quad (7)$$

where N is the number of potential customers, p is the subscription fee of the data product, q_s is the data quality level, and c is the data cost of the unit quality purchased from the data provider. Profit $\mathcal{E}(\cdot)$ is the difference between subscription revenue and total data cost. The costs of the service (such as calculation cost) are ignored.

In the above equation, $F(p)$ is equal to 0, and substituting (6) into (7) we have

$$\mathcal{E}(q_s, p) = pN (\eta U(q_s) - p) - cq_s \quad (8)$$

where, without loss of generality, we assume that $\eta=1$, and (4) and (8) are merged and organized as follows:

$$\mathcal{E}(q_s, p) = pN (\beta_1 - \beta_2 \exp(-\beta_3 q_s) - p) - cq_s \quad (9)$$

The profit maximization problem can be formulated as follows:

$$\begin{aligned} &\text{maximize} \quad \mathcal{E}(q_s, p) \\ &\text{s.t.} \quad C_1: q_s \geq 0; \\ &\quad \quad C_2: p \geq 0 \end{aligned} \quad (10)$$

The goal of (10) is to maximize the profitability of the data platform by jointly optimizing q_s and p . For constraints C_1

and C_2 , they ensure nonnegative solutions of q_s and p . Next, we will provide closed-form solutions (\bar{q}_s, \bar{p}) to this profit maximization problem and prove their global optimality.

5.3. Optimal Pricing and Quality Level. We use Karush-Kuhn-Tucker (KKT) [37] conditions to optimize the profitability of the data platform. The KKT condition is an important idea for solving Lagrangian duality problem. It is widely used in operations research, convex and nonconvex optimization, machine learning, and other fields. Based on (10), we describe the Lagrangian dual problem as follows:

$$\begin{aligned} \mathcal{L}(q_s, p, \lambda_1, \lambda_2) &= \mathcal{E}(q_s, p) + \lambda_1 q_s + \lambda_2 p \\ \text{s.t.} \quad \lambda_1 &\geq 0, \\ \lambda_2 &\geq 0 \end{aligned} \quad (11)$$

where λ_1 and λ_2 are called Lagrange multipliers and they are related to constraints C_1 and C_2 , respectively.

Proposition 1. *The closed-form solutions of \bar{q}_s and \bar{p} exist. Equation (10) has the following two roots:*

$$\bar{p} = \frac{\beta_1 \pm \sqrt{\beta_1^2 + 8c/N\beta_3}}{4} \quad (12)$$

and

$$\bar{q}_s = \frac{\ln\left(\left(\beta_1 \pm \sqrt{\beta_1^2 - 8c/N\beta_3}\right)/2\beta_2\right)}{\beta_3} \quad (13)$$

where $\lambda_1 = 0$ and $\lambda_2 = 0$.

Proof. To get this result, we first need to find (11) the first derivative of q_s and p . Then set both derivatives to zero and set the constraint to $(\lambda_1 = \lambda_2 = 0)$. In this way, a closed-form solution can be derived by a set of equations consisting of (14) and (15).

$$\frac{\partial \mathcal{L}(\cdot)}{\partial q_s} = -Np\beta_2\beta_3 \exp(\beta_3 q_s) - c + \lambda_1 = 0 \quad (14)$$

$$\frac{\partial \mathcal{L}(\cdot)}{\partial p} = N(\beta_1 - \beta_2 \exp(\beta_3 q_s) - 2p) + \lambda_2 = 0 \quad (15)$$

□

Next, we can consider two special cases where the data quality level q_s is fixed or the subscription fee p is fixed. The former corresponds to a situation where the data product has a fixed quality level, and the data platform owner only optimizes the subscription fee. In contrast, the latter corresponds to a fixed subscription fee and the data platform owner only optimizes the quality level of the data. We have the following proposition.

Proposition 2. *On the one hand, if q_s is fixed, the solution \bar{p} of the problem in (10) is globally optimal. On the other hand, if p is fixed, the solution \bar{q}_s of the problem in (10) is globally optimal.*

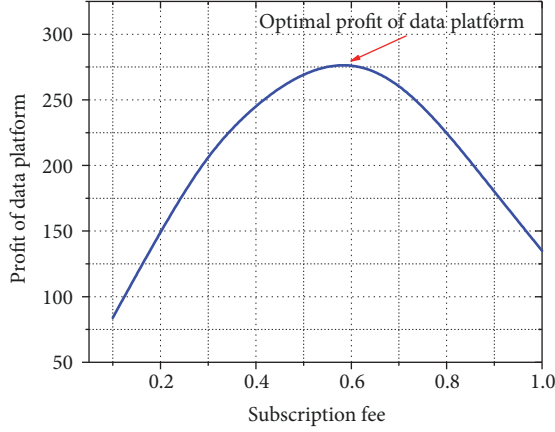


FIGURE 6: Data platform profit under different subscription fees.

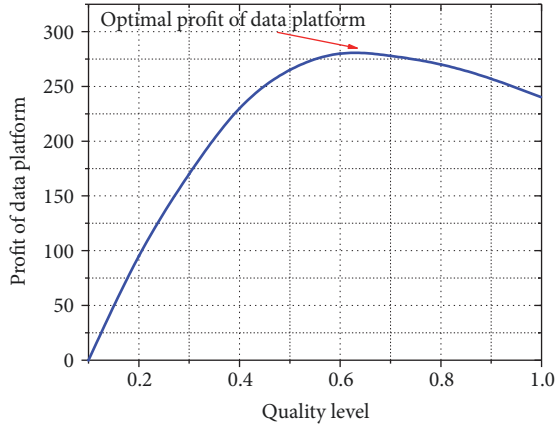


FIGURE 7: Data platform profits under different quality levels.

Proof. We solve the second derivatives of $\mathcal{G}(q_s, p)$ for q_s , p , respectively.

$$\frac{\partial^2 \mathcal{G}(q_s, p)}{\partial q_s^2} = -pN\beta_2\beta_3^2 \exp(\beta_3 q_s) < 0 \quad (16)$$

$$\frac{\partial^2 \mathcal{G}(q_s, p)}{\partial p^2} = -2N < 0 \quad (17)$$

which are nonpositive. Therefore, the solutions of the special cases are globally optimal. \square

6. Numerical Experiment

In this section, we consider using the previous utility function $U(q_{sj}; \beta)$ to obtain the numerical results of the optimal pricing scheme. From this, we can further provide data platform owners with useful decision strategies. We adopt the fitted parameters as shown in Figure 5. In addition, we assume that the number of consumers is 1000. For verification purpose, we standardize the data quality level from 0 to 1.

Figures 6 and 7 show the profit of the data platform under parameters β , p , q_s . In Figure 6, we set the fixed data quality level $q_s = 0.6$ and, at the same time, change

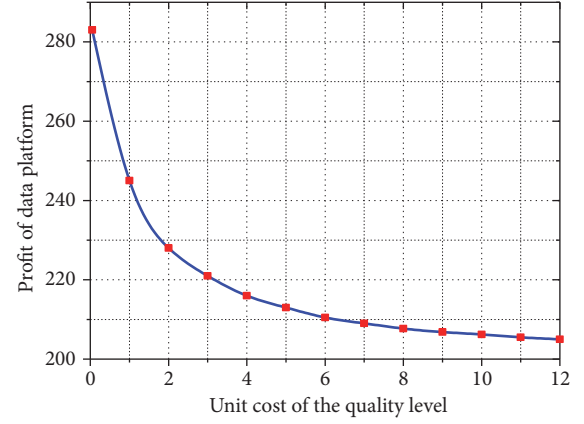


FIGURE 8: Data platform profit under different quality level costs.

subscription fee. Obviously, when the price of data is low, it will stimulate consumer spending and bring profit growth to the data platform. When the price of data is high, it will reduce the profit of the data platform. The possible influence factor is that too high price affects the willingness to pay of consumers, resulting in the loss of data platform profits. Obviously, the optimal subscription price to maximize profits can be calculated by (10).

In Figure 7, we fix $c = 1$ to investigate the effect of different levels of data quality on platform profits. Obviously, when the quality level is lower, the utility of the data and optimized subscription fee are also lower, resulting in less profit for the data platform. However, if the data quality level is high, the cost of the data platform will also increase (i.e., the data platform needs to pay more for the data publisher), which will lead to lower profits. The curve in Figure 8 shows the result. The profit of the data platform decreases as the data price of per unit quality increases. Obviously, the maximum profit can be achieved when applying the best requested data quality.

7. Conclusions

In this paper, we proposed a data pricing and profit maximization model based on data quality levels. We first constructed a linear model of the quality score based on the data quality dimension and used the square root to divide the quality level. Then we established a quality level utility model and verified the applicability of the model with machine learning algorithms. Finally, we proposed an optimized pricing mechanism allowing data platform owners to optimize quality levels and subscription fees to maximize profits.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This research was supported in part by National Nature Science Foundation of China (Grant no. 91646202) and National Key R&D Program of China (SQ2018YFB140235).

References

- [1] S. A. Fricker and Y. V. Maksimov, "Pricing of data products in data marketplaces," in *Proceedings of the International Conference of Software Business*, vol. 304, pp. 49–66, 2017.
- [2] Y. Jiao, P. Wang, D. Niyato, M. Abu Alsheikh, and S. Feng, "Profit maximization auction and data management in big data markets," in *Proceedings of the 2017 IEEE Wireless Communications and Networking Conference, WCNC*, pp. 1–6, San Francisco, CA, USA, 2017.
- [3] A. Muschalle, F. Stahl, A. Laser, and G. Vossen, "Pricing approaches for data markets," in *Proceedings of the Workshop Business Intelligence for the Real Time Enterprise*, pp. 129–144, 2012.
- [4] Xignite, <http://www.xignite.com/>.
- [5] Gnip, <http://support.gnip.com/>.
- [6] Factual, <https://www.factual.com/>.
- [7] Infochimps, <http://www.infochimps.com/>.
- [8] Microsoft windows azure marketplace, <https://azuremarketplace.microsoft.com/en-us/marketplace/>.
- [9] H.-T. Moges, V. Vlasselaer Van, W. Lemahieu, and B. Baensens, "Determining the use of data quality metadata (DQM) for decision making purposes and its impact on decision outcomes - An exploratory study," *Decision Support Systems*, vol. 83, pp. 32–46, 2016.
- [10] M. Barnabishvili, T. Ulrichs, and R. Waldherr, "Data on the descriptive overview and the quality assessment details of 12 qualitative research papers," *Data in Brief*, vol. 8, pp. 1059–1068, 2016.
- [11] M. Al-Roomi, S. Al-Ebrahim, S. Buqrais, and I. Ahmad, "Cloud computing pricing models: a survey," *International Journal of Grid and Distributed Computing*, vol. 6, no. 5, pp. 93–106, 2013.
- [12] F. Teng and F. Magoulès, "Resource Pricing and Equilibrium Allocation Policy in Cloud Computing," in *Proceedings of the 2010 IEEE 10th International Conference on Computer and Information Technology (CIT)*, pp. 195–202, Bradford, United Kingdom, June 2010.
- [13] H. Shah-Mansouri, V. W. S. Wong, and R. Schober, "Joint optimal pricing and task scheduling in mobile cloud computing systems," *IEEE Transactions on Wireless Communications*, vol. 16, no. 8, pp. 5218–5232, 2017.
- [14] X. Liu, M. Dong, K. Ota, P. Hung, and A. Liu, "Service pricing decision in cyber-physical systems: insights from game theory," *IEEE Transactions on Services Computing*, vol. 9, no. 2, pp. 186–198, 2016.
- [15] D. Niyato, X. Lu, P. Wang, D. I. Kim, and Z. Han, "Economics of internet of things: an information market approach," *IEEE Wireless Communications*, vol. 23, pp. 136–145, 2016.
- [16] S. Balasubramanian, S. Bhattacharya, and V. V. Krishnan, "Pricing information goods: a strategic analysis of the selling and pay-per-use mechanisms," *Marketing Science*, vol. 34, no. 2, pp. 218–234, 2015.
- [17] Y. Bakos and E. Brynjolfsson, "Aggregation and disaggregation of information goods: Implications for bundling, site licensing, and micropayment systems," *Lectures in E-Commerce*, pp. 103–122, 2001.
- [18] D. Niyato, D. T. Hoang, N. C. Luong, P. Wang, D. I. Kim, and Z. Han, "Smart data pricing models for the internet of things: a bundling strategy approach," *IEEE Network*, vol. 30, no. 2, pp. 18–25.
- [19] X. Wei and B. R. Nault, "Monopoly versioning of information goods when consumers have group tastes," *Production Engineering Research and Development*, vol. 23, no. 6, pp. 1067–1081, 2014.
- [20] M. Li, H. Feng, F. Chen, and J. Kou, "Optimal versioning strategy for information products with behavior-based utility function of heterogeneous customers," *Computers & Operations Research*, vol. 40, no. 10, pp. 2374–2386, 2013.
- [21] P. Koutris, P. Upadhyaya, M. Balazinska, B. Howe, and D. Suciu, "Query-based data pricing," in *Proceedings of the Symposium on Principles of Database Systems*, vol. 62, pp. 167–178, 2012.
- [22] Y. Shen, B. Guo, Y. Shen, X. Duan, X. Dong, and H. Zhang, "A pricing model for Big Personal Data," *Tsinghua Science and Technology*, vol. 21, no. 5, pp. 482–490, 2016.
- [23] J. Yang and C. Xing, "Personal data market optimization pricing model based on privacy level," *Information*, vol. 10, no. 4, p. 123, 2019.
- [24] J. Liu, J. Li, W. Li, and J. Wu, "Rethinking big data: A review on the data quality and usage issues," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 115, pp. 134–142, 2016.
- [25] F. Stahl and G. Vossen, "Data quality scores for pricing on data marketplaces," in *Asian Conference on Intelligent Information and Database Systems*, vol. 9621, pp. 215–224, 2016.
- [26] X. Ding, H. Wang, D. Zhang, J. Li, and H. Gao, "A fair data market system with data quality evaluation and repairing recommendation," in *Asia-Pacific Web Conference*, vol. 9313, pp. 855–858, 2015.
- [27] D. Niyato, M. A. Alsheikh, P. Wang, D. I. Kim, and Z. Han, "Market model and optimal pricing scheme of big data and internet of things (IoT)," in *Proceedings of the 2016 IEEE International Conference on Communications, ICC 2016*, pp. 1–6, Kuala Lumpur, Malaysia, 2016.
- [28] F. Stahl, *High-quality Web information provisioning and quality-based data pricing [Ph.D. thesis]*, University of Münster, 2015.
- [29] H. Yu and M. Zhang, "Data pricing strategy based on data quality," *Computers & Industrial Engineering*, vol. 112, supplement 1, pp. 1–10, 2017.
- [30] F. Stahl and G. Vossen, "Fair knapsack pricing for data marketplaces," in *Advances in Databases and Information Systems*, vol. 9809, pp. 46–59, 2016.
- [31] J. Yang and C. Xing, "Data source selection based on an improved greedy genetic algorithm," *Symmetry*, vol. 11, no. 2, p. 273, 2019.
- [32] R. Blake and P. Mangiameli, "The effects and interactions of data quality and problem complexity on classification," *Journal of Data and Information Quality*, vol. 2, no. 2, pp. 1–28, 2011.
- [33] J. Greene and J. Baron, "Intuitions about declining marginal utility," *Journal of Behavioral Decision Making*, vol. 14, no. 3, pp. 243–255, 2001.
- [34] The mnist database, <http://yann.lecun.com/exdb/mnist/>.
- [35] R. Kieser, P. Reynisson, and T. J. Mulligan, "Definition of signal-to-noise ratio and its critical role in split-beam measurements," *ICES Journal of Marine Science*, vol. 62, no. 1, pp. 123–130, 2005.

- [36] K. Wertenbroch and B. Skiera, "Measuring consumers' willingness to pay at the point of purchase," *Journal of Marketing Research*, vol. 39, no. 2, pp. 228–241, 2002.
- [37] G. Gordon and R. Tibshirani, "Karush–Kuhn–Tucker conditions," *Optimization*, pp. 1–26, 2012, <https://www.cs.cmu.edu/ggordon/10725-F12/slides/16-kkt.pdf>.

Research Article

Pricing Strategies in Dual-Channel Supply Chain with a Fair Caring Retailer

Lufeng Dai , Xifu Wang, Xiaoguang Liu, and Lai Wei

School of Traffic and Transportation, Beijing Jiaotong University, Beijing, China

Correspondence should be addressed to Lufeng Dai; 15114213@bjtu.edu.cn

Received 11 September 2018; Revised 2 December 2018; Accepted 28 January 2019; Published 18 April 2019

Guest Editor: Ahmet Sensoy

Copyright © 2019 Lufeng Dai et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Manufacturers add online direct channels that inevitably engage in channel competition with offline retail channels. Since price is an important factor in consumers' choice of purchasing channel, pricing strategy has become a popular topic for research on dual-channel competition and coordination. In contrast to previous research on pricing strategies based on the full rationality of members, we focus on the impact of retailers' fairness concerns on pricing strategies. In this study, the hybrid dual-channel supply chain consists of one manufacturer with a direct channel who acts as the leader and a retailer who acts as the follower. First, we use the Stackelberg game approach to determine the equilibrium pricing strategy for a fair caring retailer. Simultaneously, we consider a centralized dual-channel supply chain as the benchmark for a comparative analysis of the efficiency of a decentralized supply chain. Furthermore, we study pricing strategies when the retailer has fairness concerns and determine the complete equilibrium solutions for different ranges of the parameters representing cross-price sensitivity and fairness. Finally, through numerical experiments, the pricing strategies, the profit and utility of the manufacturer and retailer, and the channel efficiency of the supply chain are compared and analysed for two scenarios. We find that fairness concerns reduce the manufacturer's profits, while for the most part, the retailers' profit can be improved; however, the supply chain cannot achieve complete coordination.

1. Introduction

In recent years, the number of people who shop online has grown. Increasingly, manufacturers such as Lenovo, Dell, and Nike have added direct channels to increase their profitability. When a manufacturer sells through a traditional retailer and has a direct channel to consumers, it is called a dual-channel distribution system. In a dual-channel distribution system, the manufacturer and its retailers sell essentially the same products. Compared with traditional retail channels, online direct sales channels have lower operating costs, and consumers are given choices that provide them with more convenience and price discount (Takahashi K [1]). Supply-demand and competitive relationships coexist between the manufacturer and retailer after an online direct marketing channel opens. Competition may also lead to conflicts between the two channels in terms of cross-channel price and operations.

With the popularity of online shopping, the gaps between channels in some dual supply chains are narrowing as

consumers have gradually adapted to the heterogeneity of channels and shopped more rationally. According to Accenture's Chinese consumer research report for 2018[2], preference of consumers for shopping online or going to stores is almost equal, particularly, smart digital consumers paying more attention to price comparison. China-ASEAN Mobile Internet Industry Alliance released a survey report on comparison of online and offline shopping behaviors of consumers for 2018 [3]. The report shows that the considered factors for consumers to choose online or offline shopping are basically the same. Especially when it comes to the products that consumers often buy online, such as cosmetics, clothing, electronic products, FMCG, etc., people tend to pay more attention to the price of products. The narrowing of channel differences means that different channels will cover the same customer groups and make the problem of channel price competition become more prominent.

A large number of investigations and studies have confirmed that firms, similar to individuals, have fairness preferences when they are treated unfairly in business (Rabin M [4];

Fehr E [5]; Kahneman D [6]; Kumar N [7]) and have shown that fairness concerns have a significant impact on decision-making. Compared with the traditional supply chain, the dual-channel supply chain easily occur unfair cooperation due to channel competition and other factors, which will lead to conflicts of interest and even the breakdown of cooperation. In order to increase sales, high-end liquor companies such as Yibin Wuliangye Group Company Ltd. have added online direct channels and offered discounted product on it (Dong Zhao [8]). The good market prospects of Chinese liquor make the liquor enterprises raise the wholesale price. Meanwhile, the liquor enterprises were trying to stabilize the offline retail price and limit the minimum selling price to maintain the brand image (Dong Zhao [8]; Chen Xing [9]). The impact of online direct selling and the passive situation of offline retail made some retailers feel that they have been treated unfairly and privately took the way of price-off promotions which destroyed the offline price system and damaged the interest of Wuliangye. Similar unfair phenomenon also happened in the cooperation between Gome and Gree Electric Appliances (Xuefei Zhong [10]). How should manufacturers coordinate channel conflicts by price strategies considering retailers' fairness concern.

This paper try to apply a fairness preference and the cross-price effect to the dual-channel supply chain decision-making model to obtain results and insights more in line with actual management practice. In this study, we consider a dual-channel supply chain where a manufacturer acts as the leader and a fair caring retailer is the follower in Stackelberg game and approach the complete equilibrium solutions of the supply chain members.

The remainder of this paper is organized as follows. In Section 2, we review dual-channel supply chain literatures related to pricing strategy and fairness. In Section 3, we develop the structure and demand functions of a dual-channel supply chain and discuss the model assumptions and describe the parameters. In Section 4, we discuss the equilibrium pricing strategies of the members when the retailer has no fairness concerns. In Section 5, we analyse the equilibrium pricing strategies of the manufacturer and a fair caring retailer. In Section 6, a numerical experiment is used to analyse the influence of fairness concerns and cross-price sensitivity on member decisions and supply chain performance. We conclude by presenting the equilibrium results and suggesting directions for future research in Section 7. All of the proofs are provided in Appendix.

2. Literature Review

Our paper is related to two streams of research: pricing strategy and fairness concern in the dual-channel supply chain.

2.1. Pricing Strategy Researches in Dual-Channel Supply Chain. Chiang et al. [11] showed that a vertically integrated online channel allows the manufacturer to constrain its retailer's pricing behavior in a dual-channel supply chain. Guo Yajun [12] showed that adding an online direct channel can expand

the market, but it may also exacerbate channel conflicts and depress the retail price, leading to a loss in retailers' profit. Cattani et al. [13] studied the situation in which manufacturers adopt different pricing strategies in order to alleviate channel conflicts and found that reducing wholesale prices could alleviate the double margin effect and improve supply chain performance (Qing Fang [14]). Considering the dual-channel supply chain led by retailers and manufacturers, respectively, and making a comparative analysis of the optimal price decision, the dominant party will make the wholesale price beneficial to maximize its profit. Yan R [15] constructed the channel demand model based on consumer utility and studied the pricing strategy considering consumers have same price sensitivity of different channels. This paper found that channel conflict can be alleviated by price strategy after manufacturers encourage retailers to improve retail services. Based on the same channel price sensitivity of consumers, Tian J F [16] studied the pricing strategy when manufacturers develop retail service. With the development of dual-channel supply chain, some researchers have found that consumers have gradually adapted to the heterogeneity of channels, especially the products that consumers often buy online. Therefore, price comparison has become the focus of consumer attention. Xu et al. [17] analysed the price comparison behavior of consumers and its impact on decision-making and profit of supply chain members and found that the retailers and supplier are all more willing to avoid the existence of price comparison with the objective of profit maximization. Shen et al. [18] researched on the pricing strategy considering price comparison behavior and designed the corresponding coordination mechanism. To reduce channel conflict, Bo Li [19] considered a consistent pricing strategy in the two channels, which means that the price in the direct channel is equal to that in the retail channel. Zhang F [20, 21] established the dynamic price game model and analysed the impact of price adjustment on the profit of supply chain members. Excessive price adjustment is often detrimental to their own interests but will make the other side to get more profits. Many scholars (e.g., [22–25]) have studied the price strategies for different channel structures and different product strategies.

The papers mentioned above have shown that price strategies play an important role in allocating channel profits and coordinating channel conflict. In dual-channel supply chain, price competition and channel conflicts may make members pay more attention to the distribution of profits; this paper make attempts to integrate members' fair preferences into pricing studies.

2.2. Fairness Concern Researches in Dual-Channel Supply Chain. Apart from the single-channel supply chain field, few scholars are engaged in applying fairness preferences to the dual-channel supply chain in their research (You Q et al. [26]; Guangxing Wei et al. [27]), among which Tengfei Nie and Shaofu Du [28], Qinghua Li and Bo Li [29], and Fang Z et al. [20] are the most representative. Reference [28] studied the application of a quantity discount contract in a dyadic supply chain consisting of one supplier and two retailers with no cross-price influence between channels. Retailers also

focus on both horizontal and vertical fairness. This article determines the pricing strategies of the members of a supply chain when the fairness parameter differs. Further, it also introduces other coordination mechanisms to prove that the quantity discount contract cannot fully coordinate the supply chain. Reference [29] considered that the retailer provides value-added services, and they study the pricing decisions of the supply chain members for two scenarios: one in which the retailer has fairness concerns and a second in which it does not. The partial equilibrium solution of the channel quota is given in this article, but the complete equilibrium solution is not discussed. Reference [20] investigated two noncooperative dynamic game models: a Stackelberg game model and a vertical Nash game model. The paper used numerical experiments to analyse the influence of the retailer fairness preference on the dynamic behavior of supply chain members. The FS fairness model is simplified from using a piecewise function to using a continuous function to discuss how the retailer's behavior related to its fairness concerns influences member decisions and utility (Fujing Xu et al. [30]; Lei Wang et al. [31]; Bo Li et al. [32]).

In the aforementioned articles, there is a lack of attention to fairness concerns and cross-price sensitivity. Considering the influence of fairness concerns on strategy, only the local equilibrium solution of the members is inferred, and the complete process of the member's game cannot be fully understood based on this analysis. In practice, when the market environment and the level of fairness concerns of the members change, there are multiple equilibriums, which means that the members will adopt different strategies. Therefore, the complete equilibrium solutions and the corresponding management significance will become a focus of this paper.

3. Problem Statement

3.1. Model Assumption and Notation. w : per unit wholesale price of the manufacturer.

p_r : per unit offline retail price of the retailer.

p_e : per unit online direct price of the manufacturer.

w^n : the superscript n takes the values of d^* and f^{**} , which denote the optimal wholesale pricing strategies with and without fairness concerns.

p_i^n : the superscript n takes the values of c^* , d^* , and f^{**} , which denote the optimal strategies under the centralized and decentralized supply chain without fairness and the strategies considering fairness, respectively. The subscript i takes the values of r and e .

a : the potential market demand of the channel.

c : the manufacturer's marginal cost per unit.

θ : the cross-price sensitivity between channels.

α : a parameter reflecting the per unit difference in the payoffs of the manufacturer and the retailer when the retailer encounters disadvantageous unfairness.

d_r : the demand function of the offline retail channel.

d_e : the demand function of the online direct channel.

π_i : the subscript i takes the values of c , d , m , and r , which denote the total profit of both the centralized and

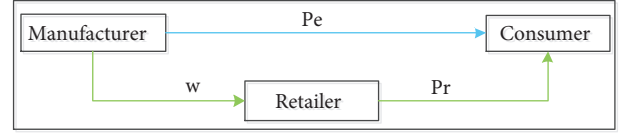


FIGURE 1: The dual-channel supply chain structure.

decentralized supply chains, the manufacturer's profit, and the retailer's profits without fairness concerns, respectively.

π_i^f : the superscript f denotes the scenario with fairness concerns; the subscript i takes the values of c , m , and r , which denote the total profit of the supply chain, the manufacturer's profit, and the retailer's profit, respectively.

The following assumptions are made in our model. (1) The manufacturer and the retailer only sell one kind of product. (2) The potential share of both the online and offline channels is the same. (3) Information is symmetrical between the manufacture and the retailer. (4) The retailer is in a weak position.

3.2. The Structure of the Dual-Channel Chain and Channel Demand Function. We consider a representative dual-channel structure in our study (see Figure 1). The supply chain consists of manufacturer M and retailer R . The manufacturer creates infinitely divisible and homogeneous products and sells through both traditional offline retailers and online direct sales channels. We establish a Stackelberg game model to describe the problem between a rational manufacturer and a retailer with a fairness concern. The process of the game is as follows: the manufacturer, as the initiator of the game, first determines the online direct selling price p_e and the wholesale price w ; the retailer acts as the follower and then sets the offline retail price p_r .

Linear demand functions are used to characterize channel demand and have been adopted in studies (Yue X [33]; Huang S [34]), and the corresponding demand functions to the manufacturer and the retailer are described as follows:

$$d_r = a_r - b_r p_r + \theta p_e, \quad (1)$$

$$d_e = a_e - b_e p_e + \theta p_r \quad (2)$$

The differences in channel characteristics are mainly reflected in channel price elasticity b and basic market demand of channel a . Subscript r and e represent the offline and online channels, respectively. Channel price elasticity depicts consumers' sensitivity to channel price. Basic market demand of channel reflects consumers' loyalty to the channel. When channel differences decrease, market characteristics corresponding to the original differentiated channels are gradually converging. Some scholars (Yan R [15]; Guangye Xu [35]) consider consumers have the same price sensitivity to different channels, but the channel loyalty is different. Channel loyalty of consumers is changing with the shopping habit. Research on China's digital consumers released by McKinsey Greater China in 2017 states that over 90% of consumers compare online and offline channels when buying consumer electronics (Wei Wang [36]). According to a white

paper on big data and online cosmetics consumption in 2016, although the growth rate of e-commerce has been slowing since 2012, the proportion of online and offline consumption is expected to be evenly divided into 2018 (Niuli [37]). Therefore, this paper considers a situation where the potential share of the channels is the same. Without loss of generality, we set parameter b_r and b_e to 1, as were done in the studies ([23]; Liu M [38]). θ is the coefficient of cross-price sensitivity. The equation $0 < \theta < 1$ reflects the own-price effects, which are greater than the cross-price effects.

Obviously, it is necessary to impose additional inequity constraints on the parameters to guarantee the operation of the dual channels: (i) $p_r \geq w$, $p_e \geq w$. If $p_e < w$, then the retailer will find a less expensive source from the direct channel. (ii) $d_e \geq 0$ and $d_r \geq 0$, which ensures that every channel has sales. (iii) $w \geq c$.

4. Pricing Strategy of Members When the Retailer Has No Fairness Concerns

4.1. Equilibrium Analysis of the Centralized Supply Chain. To examine the efficiency of the decentralized supply chain both with and without fairness concerns, we consider a centralized dual-channel supply chain as a benchmark, where the manufacturer and the retailer are regarded as a vertically integrated supply chain system. The members make decisions to maximize the overall profit of the supply chain, and the wholesale price w is no longer the decision variable in the centralized supply chain. The problem of the supply chain members is given as follows:

$$\max_{p_r, p_e} \pi_c = d_r(p_r - c) + d_e(p_e - c) \quad (3)$$

$$\begin{aligned} s.t \quad & p_r \geq c, \\ & p_e \geq c, \\ & d_e \geq 0, \\ & d_r \geq 0. \end{aligned} \quad (4)$$

By simultaneously solving the first-order conditions of the equations above for p_r , p_e , it can be shown that the Hessian matrix H is negative definite. The optimal channel price p_r , p_e and the total profit of the supply chain are obtained.

$$\begin{aligned} p_e^{c*} = p_r^{c*} &= \frac{a}{2(1-\theta)} - \frac{c}{2}, \\ \pi_c &= \frac{(\theta c + a - c)^2}{2(1-\theta)}. \end{aligned} \quad (5)$$

In an integrated supply chain, the decision maker sets a uniform retail price for both online and offline sales to avoid channel competition. Obviously, with an increase in the cross-price sensitivity coefficient, the channel price increases, and the overall profit of the supply chain increases. In reality, enterprises often adopt the same price for the dual-channel, such as Suning, but a higher level of channel management is required in this situation (Chun Yuan et al. [39]).

4.2. Equilibrium Analysis of the Decentralized Supply Chain. In this section, we consider a decentralized dual-channel supply chain based on the assumption that neither party in the supply chain has fairness concerns and that both make decisions to maximize their individual profits.

4.2.1. The Retailer's Problem. Given the manufacturer's network direct selling price p_e and wholesale price w , according to the previous game, the profit of the retailer is maximized as follows:

$$\max_{p_r} \pi_r = d_r(p_r - w). \quad (6)$$

The response function for the retailer can be described as follows:

$$p_r = \frac{a + w + \theta p_e}{2}. \quad (7)$$

4.2.2. The Manufacturer's Problem. The manufacturer's decision problem can be described as follows:

$$\max_{p_e, w} \pi_m = d_r(w - c) + d_e(p_e - c) \quad (8)$$

$$\begin{aligned} s.t \quad & p_r = \frac{a + w + \theta p_e}{2}, \\ & p_e \geq w, \\ & d_e \geq 0, \\ & d_r \geq 0, \\ & w > c. \end{aligned} \quad (9)$$

Therefore, the Hessian matrix of the manufacturer's profit function

$$H_{\pi_m} \begin{bmatrix} p_e \\ w \end{bmatrix} = \begin{bmatrix} \theta^2 - 2 & \theta \\ \theta & -1 \end{bmatrix} \quad (10)$$

is negatively definite. The manufacturer's profit function is a concave function of p_e and w , and the decision problem is a convex optimization problem. Thus, a unique equilibrium solution exists. Therefore, we can deduce the optimal decision as follows:

$$p_e^{d*} = w^{d*} = \frac{a}{2(1-\theta)} - \frac{c}{2} \quad (11)$$

It is easy to prove (11) and satisfy (9). The optimal retail price is given by bringing (11) into (7):

$$p_r^* = \frac{(3-\theta)a}{4(1-\theta)} + \frac{(1+\theta)c}{4} \quad (12)$$

By bringing (11) and (12) into (6) and (8), we can obtain the profit of the manufacturer, the profit of the retailer, and the total profit of the supply chain as follows:

$$\begin{aligned}\pi_m &= \frac{(\theta + 3)(\theta c + a - c)^2}{8(1 - \theta)}, \\ \pi_r &= \frac{(\theta c + a - c)^2}{16}, \\ \pi_d &= \frac{(\theta c + a - c)^2(\theta + 7)}{16(1 - \theta)}.\end{aligned}\quad (13)$$

By analysing the inferred strategies and profits, we find the following: the channel efficiency ($\pi_d/\pi_c = (\theta + 7)/8$) of the decentralized supply chain is an increasing function of the cross-price coefficient θ , which indicates that the double marginalization will be weakened when θ increases. Similarly, we can draw the same conclusion from the equilibrium pricing strategy. The cross-price sensitivity coefficient has a positive impact on the channel price. As θ increases (see (11) and (12)), both the manufacturer and the retailer will use a higher pricing strategy to improve its profits, which increases the overall profit of the supply chain.

5. Pricing Strategy of Members When the Retailer Has Fairness Concerns

The supply chain cannot be coordinated when the retailer does not have a fairness concern. However, it is necessary to determine how the strategy changes when we consider the effect of fairness concerns and whether channel efficiency could be improved. These issues are discussed in the following section.

5.1. The Retailer's Problem. Because we consider the fairness concerns of the retailer in the distribution of profits, we must establish a model for fairness concerns. Some may argue that a more general model that includes both aversion to disadvantageous inequality and aversion to advantageous inequality (for example, Fehr and Schmidt [5] and Charness

& Rabin [40]) is more desirable. However, a preference for advantageous inequality is much less prominent (Loewenstein, Thompson, and Bazerman [41] did not find it in their experiment Ho & Su [42]). Furthermore, we assume that the retailer's fairness reference is the manufacturer's profit $\pi_{s,r}$ instead of $\gamma\pi_{s,r}$ ($\gamma > 0$) because the general setting will not produce substantially different or more insightful results than a simple setting with $\gamma = 1$ (Pavlov & Katok [43]; Tengfei Nie and Shaofu Du [28]). The utility function of the retailer can be written as follows:

$$U_r = \pi_r - \alpha(\pi_{m,r} - \pi_r)^+, \quad (14)$$

where $\pi_r = d_r(p_r - w)$ denotes the retailer's monetary payoff, $\pi_{m,r} = d_r(w - c)$ denotes the profits of the manufacturer's offline channel, and α denotes the level of the retailer's fairness concern about the distribution of offline profits, as retailers pay more attention to the profits made by manufacturers from offline channels and compare them to the profits made from the online channel. If the retailer's monetary profit is lower than the equitable profit $\pi_{m,r} - \pi_r \geq 0$, disadvantageous inequality occurs. By contrast, if $\pi_{m,r} - \pi_r < 0$, then $U_r = \pi_r$, which indicates that the retailer's utility function is equal to the profit function.

When the retailer faces disadvantage inequality, its decision problem is

$$\max_{p_r} U_r = (1 + \alpha)d_r(p_r - w) - d_r(w - c) \quad (15)$$

$$s.t \quad p_r \leq 2w - c \quad (16)$$

The decision of the retailer is

$$p_r^f = \begin{cases} \frac{w(2\alpha + 1)}{2(\alpha + 1)} + \frac{\theta p_e}{2} - \frac{a + \alpha(a - c)}{2(\alpha + 1)} & \text{if } p_e < K_1 w - D_1 \\ 2w - c & \text{if } p_e \geq K_1 w - D_1 \end{cases} \quad (17)$$

where $K_1 = (2\alpha + 3)/\theta(\alpha + 1)$; $D_1 = (\alpha + 2)c/\theta(\alpha + 1) + a/\theta$. The corresponding utility of the retailer is

$$U_r = \begin{cases} \frac{(b(\alpha + 1)p_e - (2\alpha + 1)w + (\alpha + 1)a + \alpha c)^2}{4(\alpha + 1)} & \text{if } p_e < K_1 w - D_1 \\ (w - c)(\theta p_e - 2w + a + c) & \text{if } p_e \geq K_1 w - D_1 \end{cases} \quad (18)$$

When the retailer is faced with advantageous inequality, its decision problem is

$$\max_{p_r} U_r = (1 + \alpha)d_r(p_r - w) - d_r(w - c) \quad (19)$$

$$s.t \quad p_r > 2w - c \quad (20)$$

The decision of the retailer is

$$p_r^f = \begin{cases} \frac{\theta p_e + w + a}{2} & \text{if } p_e > K_2 w - D_2 \\ 2w - c & \text{if } p_e \leq K_2 w - D_2 \end{cases} \quad (21)$$

where $K_2 = 3/\theta$; $D_2 = (a + 2c)/\theta$.

The corresponding utility of the retailer is

U_r

$$= \begin{cases} \frac{1}{4}(a-w+\theta p_e)^2 & \text{if } p_e > K_2 w - D_1 \\ (w-c)(\theta p_e - 2w + a + c) & \text{if } p_e \leq K_2 w - D_1. \end{cases} \quad (22)$$

$$p_r^f = \begin{cases} \frac{\theta p_e + w + a}{2} & \text{if } p_e > K_2 w - D_2 \\ 2w - c & \text{if } K_1 w - D_1 \leq p_e \leq K_2 w - D_2 \\ \frac{w(2\alpha + 1)}{2(\alpha + 1)} + \frac{\theta p_e}{2} - \frac{a + \alpha(a - c)}{2(\alpha + 1)} & \text{if } p_e < K_1 w - D_1 \end{cases} \quad (23)$$

$$U_r = \begin{cases} \frac{1}{4}(a-w+\theta p_e)^2 & \text{if } p_e > K_2 w - D_1 \\ (w-c)(\theta p_e - 2w + a + c) & \text{if } K_1 w - D_1 \leq p_e \leq K_2 w - D_1 \\ \frac{(b(\alpha + 1)p_e - (2\alpha + 1)w + (\alpha + 1)a + \alpha c)^2}{4(\alpha + 1)} & \text{if } p_e < K_1 w - D_1. \end{cases} \quad (24)$$

5.2. *The Manufacturer's Problem.* We divide the feasible region of the manufacturer's strategies to obtain the equilibrium solutions. By substituting (12) into $d_e \geq 0$, $d_r \geq 0$, and summarizing other conditions ($p_e \geq w$, $w > c$, and the fairness boundary condition), we can confirm the feasible region as shown in Figure 2.

The feasible region consists of R1, R2, and R3. R1 and R3 denote the feasible region of the manufacturer's strategy when the retailer faces both disadvantageous inequality and advantageous inequality. Therefore, R2 denotes the feasible region when the retailer obtains a fair distribution of the profits. The expressions for the boundary conditions are summarized in Table 1.

It is easy to prove that R2 and R3 satisfy the constraint ($d_r > 0$). Considering the response functions of the different regions, we can solve the partial equilibrium strategies of the manufacturer accordingly. Then, we can obtain the optional solutions (w^{f**} , p_e^{f**}) by comparing the partial equilibrium strategies of the different regions.

In R1, we denote $(w_1^f, p_{e,1}^f)$ as the partial equilibrium strategies and π_{m1}^f as the optimal profit of the manufacturer. The optimal model is as follows:

$$\max_{p_e, w} \pi_{m1}^f = d_e(p_e - c) + d_r(w - c). \quad (25)$$

$$\begin{aligned} \text{s.t. } p_r &= \frac{w(2\alpha + 1)}{2(\alpha + 1)} + \frac{\theta p_e}{2} + \frac{a + \alpha(a - c)}{2(\alpha + 1)}, \\ p_e &< K_1 w - D_1, \\ p_e &\geq K_3 w - D_3, \\ p_e &> K_4 w - D_4, \\ p_e &< K_5 w - D_5. \end{aligned} \quad (26)$$

Proposition 1. *The decision and corresponding utility of the retailer can be summarized as follows:*

In R2, we denote $(w_2^f, p_{e,2}^f)$ as the partial equilibrium strategies and π_{m2}^f as their optimal profit. The optimal model is as follows:

$$\max_{p_e, w} \pi_{m2}^f = d_e(p_e - c) + d_r(w - c) \quad (27)$$

$$\begin{aligned} \text{s.t. } p_r &= 2w - c, \\ p_e &\geq K_1 w - D_1, \\ p_e &\leq K_2 w - D_2, \\ p_e &> K_3 w - D_3, \\ p_e &\leq K_6 w - D_6. \end{aligned} \quad (28)$$

In R3, we denote $(w_3^f, p_{e,3}^f)$ as the partial equilibrium strategies, and π_{m3}^f denotes the optimal profit for the manufacturer as follows:

$$\max_{p_e, w} \pi_{m3}^f = d_e(p_e - c) + d_r(w - c). \quad (29)$$

$$\begin{aligned} \text{s.t. } p_r &= \frac{\theta p_e + w + a}{2}, \\ p_e &\geq K_2 w - D_2, \\ p_e &\geq K_3 w - D_3, \\ p_e &\leq K_7 w - D_7, \\ w &> c. \end{aligned} \quad (30)$$

Therefore, the optimal profit from the decision-making problem for the manufacturer is $\max \pi_m^{f*} = \max\{\pi_{m1}^f, \pi_{m2}^f, \pi_{m3}^f\}$. Thus, we denote (w^{f**}, p_e^{f**}) as the global optimal solution, which is referred to as an equilibrium strategy in the following. By substituting (w^{f**}, p_e^{f**}) into (23), we can deduce the optimal retail price p_r^f . Furthermore,

TABLE 1: Boundary conditions of the feasible region.

	Expressions for boundary conditions	Meaning
L1	$p_e = \frac{(2\alpha + 3)}{\theta(\alpha + 1)}w - \frac{(\alpha + 2)c}{\theta(\alpha + 1)} - \frac{a}{\theta}$	Fairness condition
L2	$p_e = \frac{3}{\theta} - \frac{a + 2c}{\theta}$	Fairness condition
L3	$p_e = w$	Foundation boundary
L4	$p_e = \frac{(2\alpha + 1)w}{\theta(\alpha + 1)} - \frac{a}{\theta} - \frac{\alpha c}{\theta(\alpha + 1)}$	R1: $d_r = 0$
L5	$p_e = \frac{(2\alpha + 1)\theta w}{(2 - \theta^2)(\alpha + 1)} + \frac{(2 + \theta)a}{2 - \theta^2} - \frac{\alpha\theta c}{(2 - \theta^2)(\alpha + 1)}$	R1: $d_e = 0$
L6	$p_e = \frac{(3\alpha + 2)\theta w}{2(2 - \theta^2)(\alpha + 1)} + \frac{(\alpha + 1)(\theta + 2)a - ((\alpha + 1)(\theta^2 + \theta - 2) + \alpha\theta)c}{2(2 - \theta^2)(\alpha + 1)}$	R2: $d_e = 0$
L7	$p_e = \frac{\theta w}{2 - \theta^2} + \frac{(\theta + 2)a}{2 - \theta^2}$	R3: $d_e = 0$

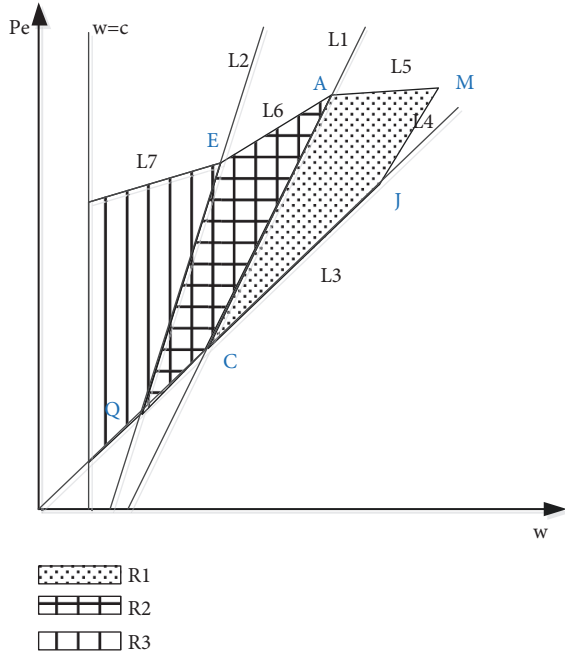


FIGURE 2: The feasible region of the manufacturer's strategies.

TABLE 2: The partial equilibrium strategies of the manufacturer in R1.

α	$(0, \alpha_1]$		$(\alpha_1, 1)$		
θ	$(0, \bar{\theta}_3]$	$(\bar{\theta}_3, 1)$	$(0, \bar{\theta}_2]$	$(\bar{\theta}_2, \bar{\theta}_3]$	$(\bar{\theta}_3, 1)$
w_1^f	w_1^{f*}	$\bar{w}_{1,4}^{f*}$	$\bar{w}_{1,3}^{f*}$	w_1^{f*}	$\bar{w}_{1,4}^{f*}$
$p_{e,1}^f$	$p_{e,1}^{f*}$	$\bar{p}_{e1,4}^{f*}$	$\bar{p}_{e1,3}^{f*}$	$p_{e,1}^{f*}$	$\bar{p}_{e1,4}^{f*}$

we can obtain the optimal profit and utility of the retailer as well as the profit of the manufacturer.

First, we discuss the pricing strategies of R1, R2, and R3 (hereinafter referred to as "partial equilibrium strategies"). $(w_i^{f*}, p_{e,i}^{f*})$, $i = 1, 2, 3$, denotes the extreme point in region

TABLE 3: The partial equilibrium strategies of the manufacturer in R2.

θ	$(0, \bar{\theta}_6]$	$(\bar{\theta}_6, \bar{\theta}_5]$	$(\bar{\theta}_5, 1)$
w_2^f	$\bar{w}_{2,2}^{f*}$	w_2^{f*}	$\bar{w}_{2,1}^{f*}$
$p_{e,2}^f$	$\bar{p}_{e2,2}^{f*}$	$p_{e,2}^{f*}$	$\bar{p}_{e2,1}^{f*}$

i . $(\bar{w}_{i,j}^{f*}, \bar{p}_{e,i,j}^{f*})$ denotes the maximum point on boundary j in region i .

Lemma 2. In R1, the partial equilibrium strategies of the manufacturer are shown in Table 2.

Appendix A provides the proof of Lemma 2, the analytical expressions of all partial equilibrium strategies and the thresholds of the parameters.

Similarly, we provide the optional solutions for R2 and R3 in Lemmas 3 and 4.

Lemma 3. In R2, the partial equilibrium strategies of the manufacturer are shown in Table 3.

Appendix B provides the proof of Lemma 3, the analytical expressions of all partial equilibrium strategies, and the thresholds of the parameters.

Lemma 4. In R3, the partial equilibrium strategy of the manufacturer is shown as follows:

$$(w_3^f, p_{e,3}^f) = (\bar{w}_{3,2}^{f*}, \bar{p}_{e3,2}^{f*}) = (\bar{w}_{2,2}^{f*}, \bar{p}_{e2,2}^{f*}) \quad (31)$$

The partial equilibrium strategy of the manufacturer in R3 $(\bar{w}_{3,2}^{f*}, \bar{p}_{e3,2}^{f*})$ is equal to the partial $(0 < \theta \leq \bar{\theta}_6)$ solution of R2 $(\bar{w}_{2,2}^{f*}, \bar{p}_{e2,2}^{f*})$, which proves that there is no partial equilibrium strategy in R3. We can also obtain the same conclusion through the following analysis. The retailer will decide $p_r^f = 2w - c$ rather than $p_r^f = (\theta p_e + w + a)/2$ because of the fairness concern when the manufacturer adopts the optimal pricing strategy $(w_3^{f*}, p_{e,3}^{f*})$, which reduces the profit of the manufacturer. Thus, the manufacturer will adopt strategy

TABLE 4: The complete equilibrium strategy of the manufacturer.

θ	Parameter range		Equilibrium strategy		
	α	θ	α	w_e^{f**}	p_e^{f**}
$0 < \theta \leq \bar{\theta}_6$	$0 < \alpha \leq \alpha_1$	$0 < \theta \leq \bar{\theta}_7$	$0 < \alpha \leq \alpha_2$	w_1^{f*}	$P_{e,1}^{f*}$
		$\bar{\theta}_6 < \theta \leq \bar{\theta}_7$	$\alpha_2 < \alpha \leq \alpha_1$	$\bar{w}_{2,2}^{f*}$	$\bar{P}_{e,2,2}^{f*}$
	$\alpha_1 < \alpha < 1$	$0 < \theta \leq \bar{\theta}_2$	$0 < \alpha \leq \alpha_1$	w_1^{f*}	$P_{e,1}^{f*}$
		$\bar{\theta}_2 < \theta \leq \bar{\theta}_7$	$\alpha_1 < \alpha < 1$	$\bar{w}_{2,2}^{f*}$	$\bar{P}_{e,2,2}^{f*}$
		$\bar{\theta}_7 < \theta \leq \bar{\theta}_6$	$\alpha_1 < \alpha < 1$	$\bar{w}_{2,2}^{f*}$	$\bar{P}_{e,2,2}^{f*}$
$\bar{\theta}_6 < \theta \leq \bar{\theta}_5$	$0 < \alpha < 1$	$\bar{\theta}_6 < \theta \leq \bar{\theta}_8$	$0 < \alpha \leq \alpha_3$	w_1^{f*}	$P_{e,1}^{f*}$
		$\bar{\theta}_8 < \theta \leq \bar{\theta}_5$	$\alpha_3 < \alpha < 1$	w_2^{f*}	$P_{e,2}^{f*}$
		$\bar{\theta}_5 < \theta \leq \bar{\theta}_3$	$0 < \alpha < 1$	w_1^{f*}	$P_{e,1}^{f*}$
$\bar{\theta}_5 < \theta \leq \bar{\theta}_3$	$0 < \alpha < 1$	$\bar{\theta}_5 < \theta \leq \bar{\theta}_3$	$0 < \alpha < 1$	w_1^{f*}	$P_{e,1}^{f*}$
$\bar{\theta}_3 < \theta < 1$	$0 < \alpha < 1$	$\bar{\theta}_3 < \theta < 1$	$0 < \alpha < 1$	$\bar{w}_{1,4}^{f*}$	$\bar{P}_{e,1,4}^{f*}$

$(\bar{w}_{2,2}^{f*}, \bar{P}_{e,2,2}^{f*})$ rather than $(w_3^{f*}, p_{e,3}^{f*})$ after considering the situation. This phenomenon also reflects the diversity of the impact of fairness concerns on the member's decision-making.

Appendix C provides the proof of Lemma 4.

Proposition 5. *By comparing the partial equilibrium strategies of R1-R3, we can determine the complete equilibrium strategies, as shown in Table 4.*

Appendix D provides the proof of Proposition 5, the analytical expressions of all partial equilibrium strategies, and the thresholds of the parameters.

Note that θ and α codivide the manufacturer's final decision space in Table 4. The complete equilibrium solution is composed of four pricing strategies of the manufacturer and the retailer. Additionally, there are some remarkable phenomena. (1) The manufacturer does not adopt the pricing strategy in R3, which is illustrated in Lemma 4. (2) When the cross-price sensitivity coefficient exceeds a threshold $\bar{\theta}_3$, the pricing strategy of the manufacturer makes the offline sales volume too small, and the manufacturer may choose to cancel the offline retail channel at this time.

The expressions of the optimal solutions are complex; therefore, the analysis of Proposition 5 is supported by numerical examples in the next section. In addition, all of the thresholds for θ and α are analytic expressions; therefore, the influence of the parameters on the members' decisions and supply chain efficiency is analysed by selecting parameters that are representative of a real-life situation.

6. Numerical Analysis

In this section, using numerical experiments, we provide additional management implications to prove the propositions discussed above. The analysis is conducted as follows: we analyse the impacts of the retailer's fairness concerns and cross-price sensitivity on the pricing strategies and the profits and utility of the two members in different settings. In

particular, we focus on the impact of the influencing factor on the variations in channel efficiency when the members change their strategies.

As this paper mainly studies the influence of fairness concerns on dual-channel decision-making, several reference values are set for the cross-price sensitivity coefficient, and an interval simulation is conducted for the fairness concern, where α varies from 0 to 1. We employ data based on a comparison of previous studies ([23, 29]). The cross-price sensitivity coefficient is set to 0.3 and 0.5, which means that the coefficient is normal and high, respectively. The other basic parameters in the experiments are set as follows: $a=1$ and $c=0.3$. The constraint problem is no longer considered, as the complete equilibrium solution satisfies the constraint conditions in the proof. The experimental results are shown in Figures 3–7.

Observation 6 (change in equilibrium strategy).

(1) *Effects of θ and α on Pricing Strategy.* As α changes, the manufacturer develops two strategies considering a fair caring retailer, as shown in Figures 3 and 4. For convenience, $(p_{r,1}^{f*}, w_1^{f*}, p_{e,1}^{f*})$ is referred to as equilibrium strategy 1, and $(p_{r,1}^{f*}, w_2^{f*}, p_{e,2}^{f*})$ is referred to as equilibrium strategy 2. The two scenarios used for the supply chain are simple; in scenario 1, the retailer does not have fairness concerns, and in scenario 2, the retailer does.

The decision analyses of the retailer and manufacturer are shown as follows. (i) In scenario 2, when the cross-price sensitivity coefficient is normal, $\theta = 0.3$, and the level of the retailer's fairness concern is less. The manufacturer will set a high wholesale price and direct channel price to reduce the profits of an ambitious retailer in a traditional retail channel, which results in a disadvantageous inequality. As α increases, meaning that the retailer's sense of fairness grows stronger, the manufacturer sets a lower w_1^* and $p_{e,1}^*$, and the retailer will set a higher retail price $p_{r,1}^{f*}$. When the parameters exceed the threshold $\bar{\alpha}$ (0.57), the manufacturer will adopt equilibrium strategy 2, which consists of a lower w

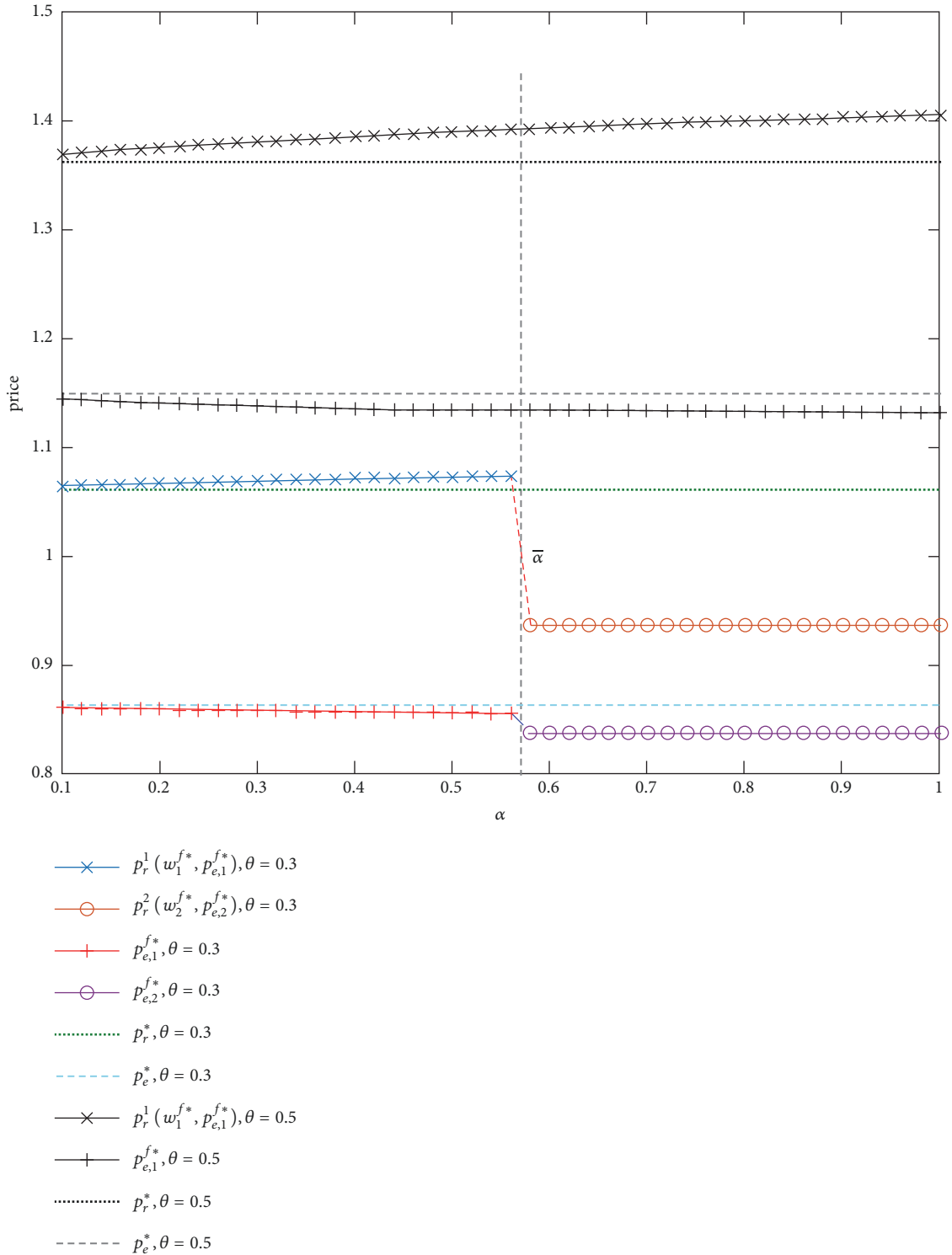
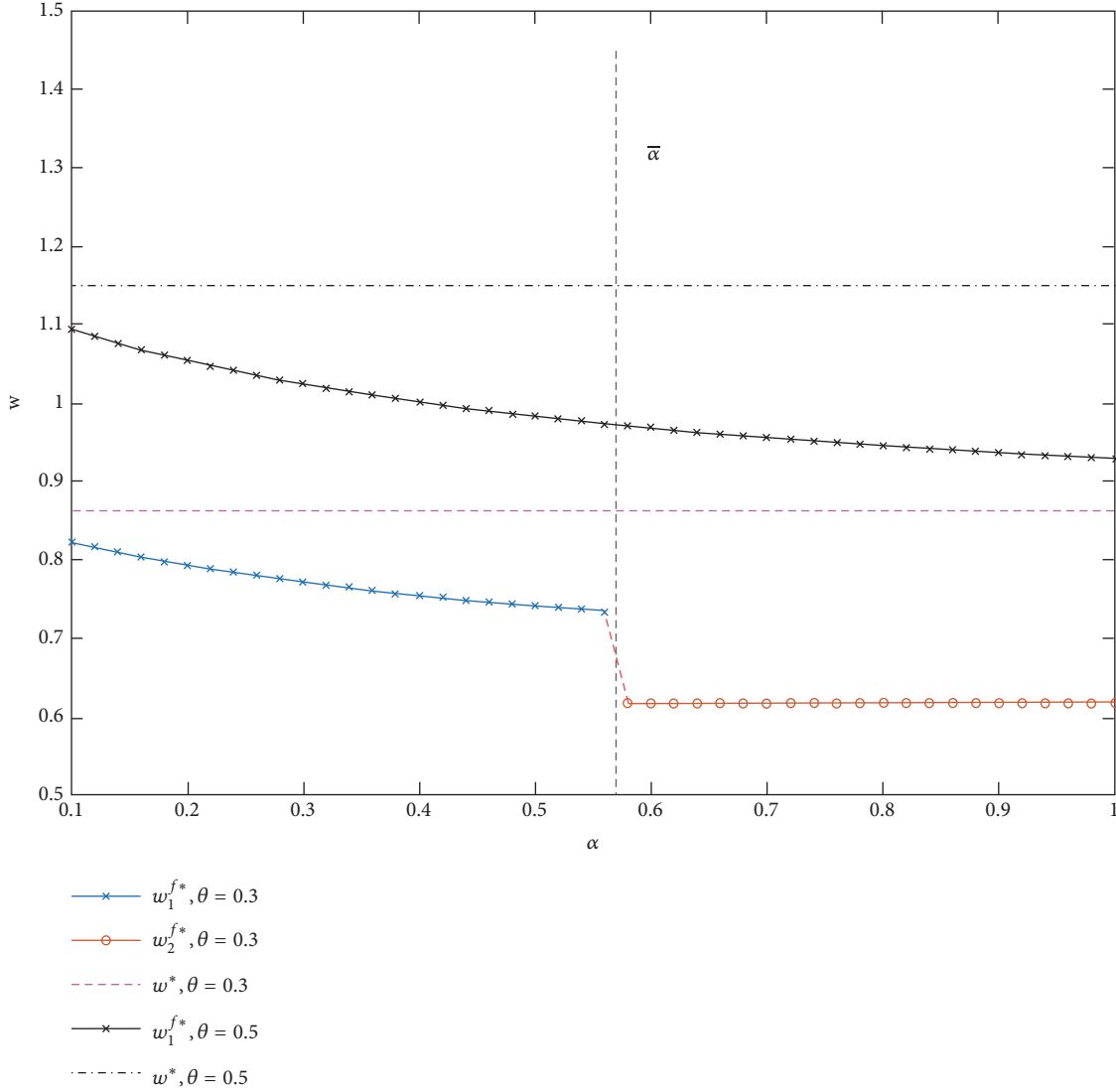


FIGURE 3: Impact of θ and α on online and offline prices.

and p_e , to achieve channel fairness. (ii) In scenario 2, when the cross-price sensitivity coefficient is high, $\theta = 0.5$, the price strategies of the retailer and manufacturer are all higher than before ($\theta = 0.5$). Another interesting phenomenon is observed. As the cross-sensitivity coefficient increases, the manufacturer becomes more likely to always adopt one

strategy without considering the disadvantageous inequality, which leads the retailer to care more about fairness. This phenomenon reflects the diversity of strategies used by members, which conflicts with the conclusions of Qinghua Li [29] and proves Proposition 5. (iii) The wholesale price and online direct price in scenario 2 are always lower than those

FIGURE 4: Impact of θ and α on w .

for scenario 1, and the gaps slowly increase as α increases. Compared to the manufacturer, the offline retail price set by the retailer in scenario 2 is lower than that in scenario 1 only when the fairness channel exists. (iv) Retailers with a stronger fairness concern are more likely to enter a neutral state or obtain a fairness result.

Observation 7 (changes in the profits and utility of the members).

(1) *Effects of θ and α on the Manufacturer's Profit.* We can deduce some information by observing Figure 5. (i) In Figure 5, $\theta = 0.3$, and the manufacturer adopts equilibrium strategy 1, as the profit of this strategy is higher than that of equilibrium strategy 1. As α increases, the gap between the two strategies decreases. Therefore, the manufacturer will change its strategy if the level of the fairness concern of the retailer exceeds a threshold $\bar{\alpha}$. A comparison of the profit for the manufacturer's two strategies is proof of the previous

analysis in Observation 6. While the cross-price sensitivity coefficient is high, when $\theta=0.5$, the manufacturer will always adopt equilibrium strategy 1 even though the corresponding profit decreases as α increases. This phenomenon occurs because the manufacturer would rather have a disadvantageous inequality existing than preserve a fairness channel that could hurt his interest. (ii) Due to retailer's behavior related to his fairness concern, the manufacturer's profit is always less than in scenario 1. (iii) Cross-price sensitivity has a positive effect on the manufacturer.

(2) *Effects of θ and α on the Retailer's Profit and Utility.* Figure 6 shows the utility and profit of the retailer. Similarly, the analysis has some implications. (i) Compared to scenario 1, the profit and utility of the retailer increase because of his fairness concern, which differs from that of the manufacturer. (ii) The retailer's utility ($\theta = 0.3$) is more than that for ($\theta = 0.5$) as α increases. If $\alpha > \bar{\alpha}$, the fairness channel exists, and both the utility and profit increase considerably. It can be seen

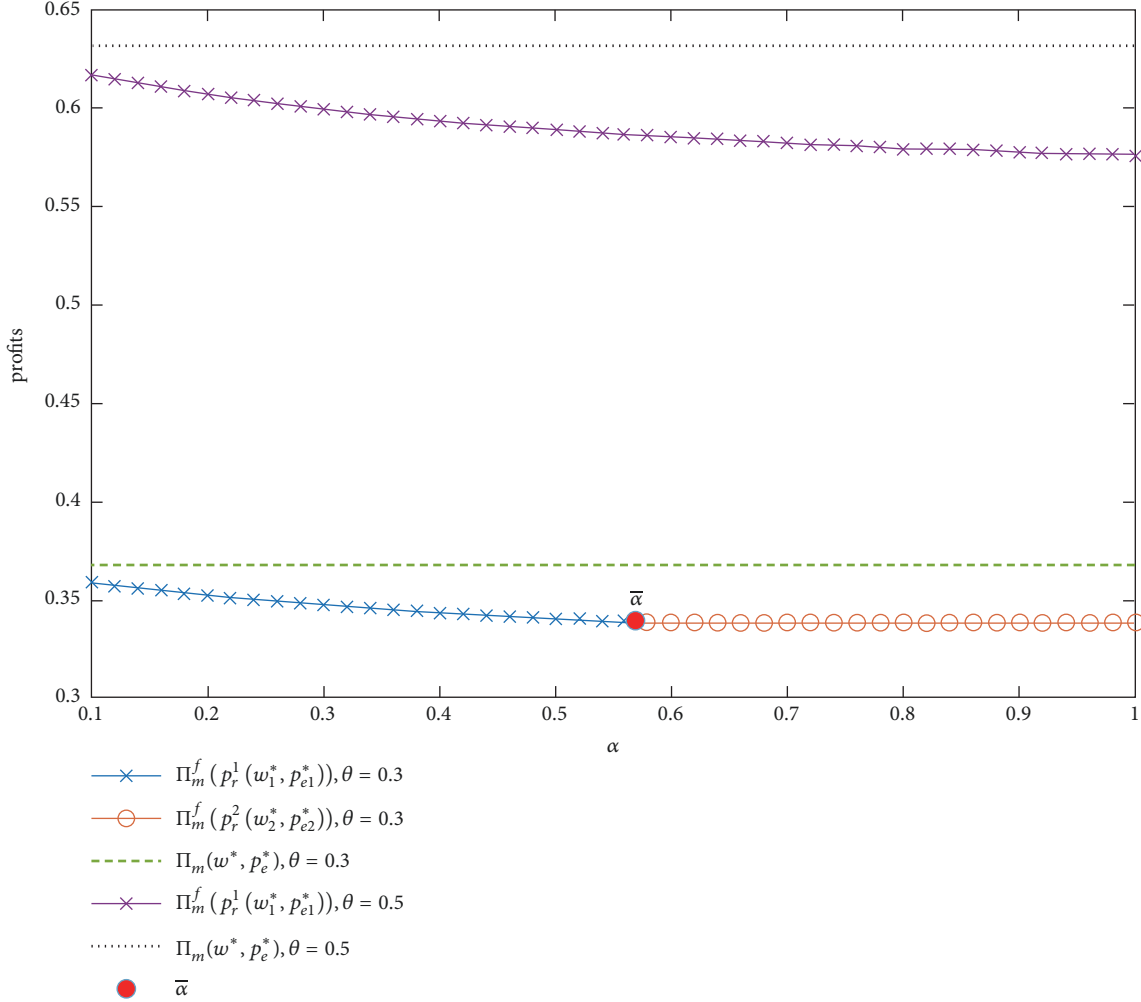


FIGURE 5: Impact of θ and α on π_m^f for the two scenarios.

that fairness concerns have a greater effect on the retailers than cross-price sensitivity.

Observation 8 (changes in channel efficiency). The fairness concern of the retailer will affect the decisions and profits of the members. To further illustrate the impact of the retailer's fairness concerns on double marginalization, we use $(\pi_r^{f*} + \pi_m^{f*})/\pi_c^{c*}$ to present the channel efficiencies, as done in [29], where π_c^{c*} represents the total profit of the centralized supply chain. Figure 7 illustrates how the channel efficiencies change as α increases and θ changes.

Figure 7 provides valuable information. (i) An increase in α widens the gap between the two scenarios when the retailers' fairness concerns do not reach a certain level $\bar{\alpha}$ ($\alpha < \bar{\alpha}$). The intuitive explanations for this result are as follows. As α increases, the manufacturer reduces both the wholesale prices and the network direct prices, while the retailers raise retail prices to boost their profits, which leads to double marginalization. (ii) We discuss the situation in which the level of the retailers' fairness concerns exceeds the threshold ($\alpha > \bar{\alpha}$). In this case, cross-price sensitivity is

normal. The manufacturer adopts a lower pricing strategy because he is focused on the retailer's fairness concerns. This adjustment results in a fairness channel. Then, the retailer obtains a greater profit even after setting a lower price. The performance of the supply chain is obviously enhanced and tends towards Pareto optimality. Simultaneously, cross-price sensitivity is high. Channel efficiency decreases as α increases. This change can be explained by the fact that the manufacturer's strategy considers his own interests, and he did not provide a fair distribution for the retailer. (iii) The supply chain cannot be coordinated by a constant wholesale price in scenarios 1 and 2.

7. Concluding Remarks

In this paper, we considered a dual-channel supply chain that includes one manufacturer and one retailer. A supply-demand relationship and a competitive relationship will coexist between the manufacturer and retailer after an online direct marketing channel opens. Based on this complex relationship, we investigated the considered model with two

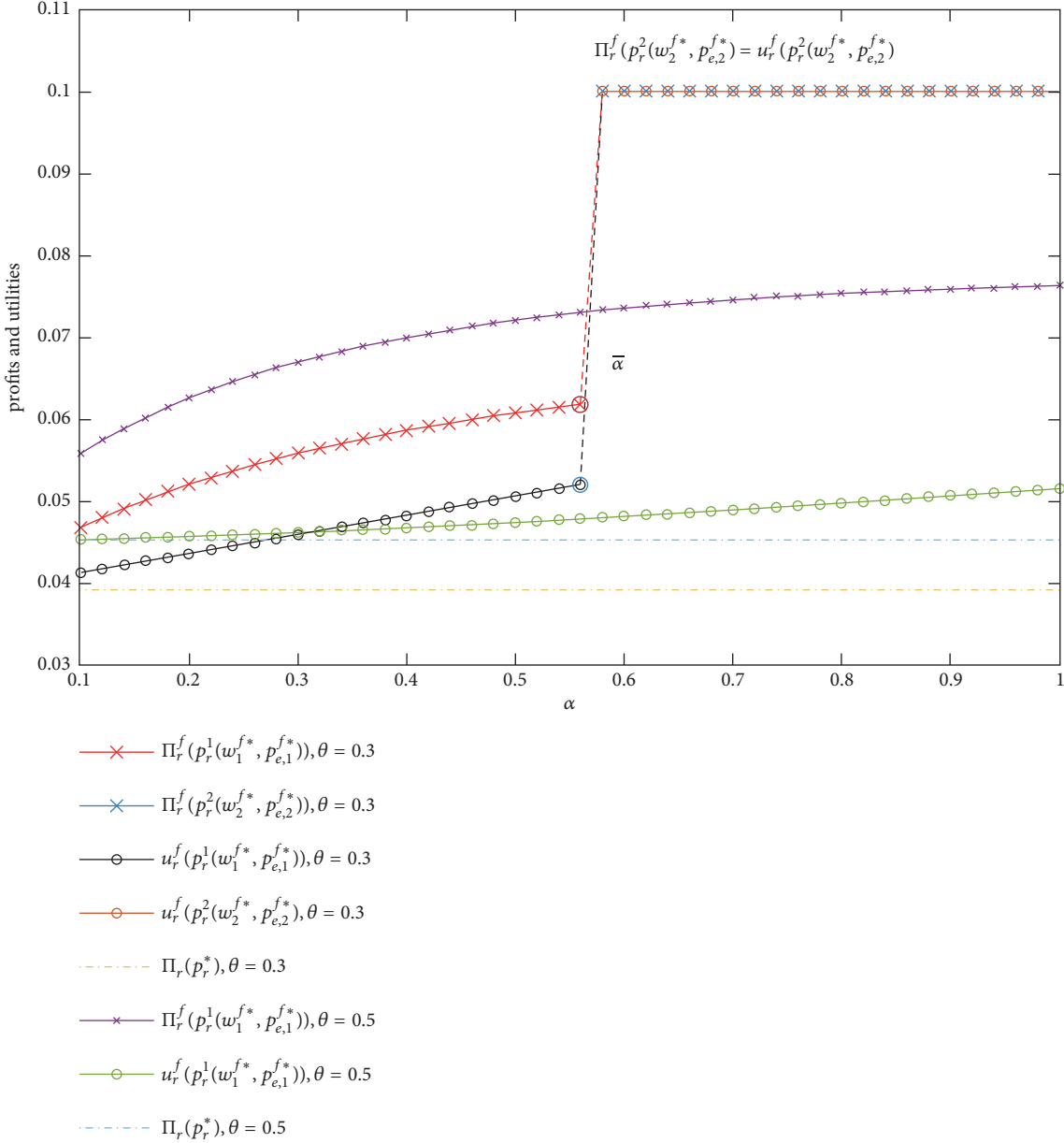


FIGURE 6: Impact of θ and α on π_r^f and u_r^f for the two scenarios.

scenarios to represent whether the retailer has fairness concerns. Simultaneously, the cross-price sensitivity coefficient, which affects the price competition between dual channels, is emphasized in the analysis of pricing decisions. By theoretical derivation, we get the following conclusions and management enlightenments as follows.

When the manufacturer faces a rational retailer, the optimal pricing strategy of the decentralized supply chain members are increasing functions of the cross-price coefficient θ . Therefore, the enterprise managers can set a relatively high retail price, and its negative effect on channel demand will be weakened when all channel consumers pay more attention to price comparison between channels. The result is more conducive to improving the profits of supply chain.

The channel efficiency of the decentralized supply chain is an increasing function of the cross-price coefficient θ , which indicates that the double marginalization will be weakened when θ increases.

Retailers' fair concern behavior has an impact on price strategy. We integrate fair concern into the study of price strategy. Different from previous studies, this paper deduce a complete set of pricing decisions when manufacturers balance channel fairness and self-interest. θ and α codetermine the manufacturer's final decision which reflect that rational manufacturers should take into account both the cross-price impact between channels and retailers' fair behavior when formulating pricing strategies. The decision set reflects that when the cross-price influence coefficient is fixed, the

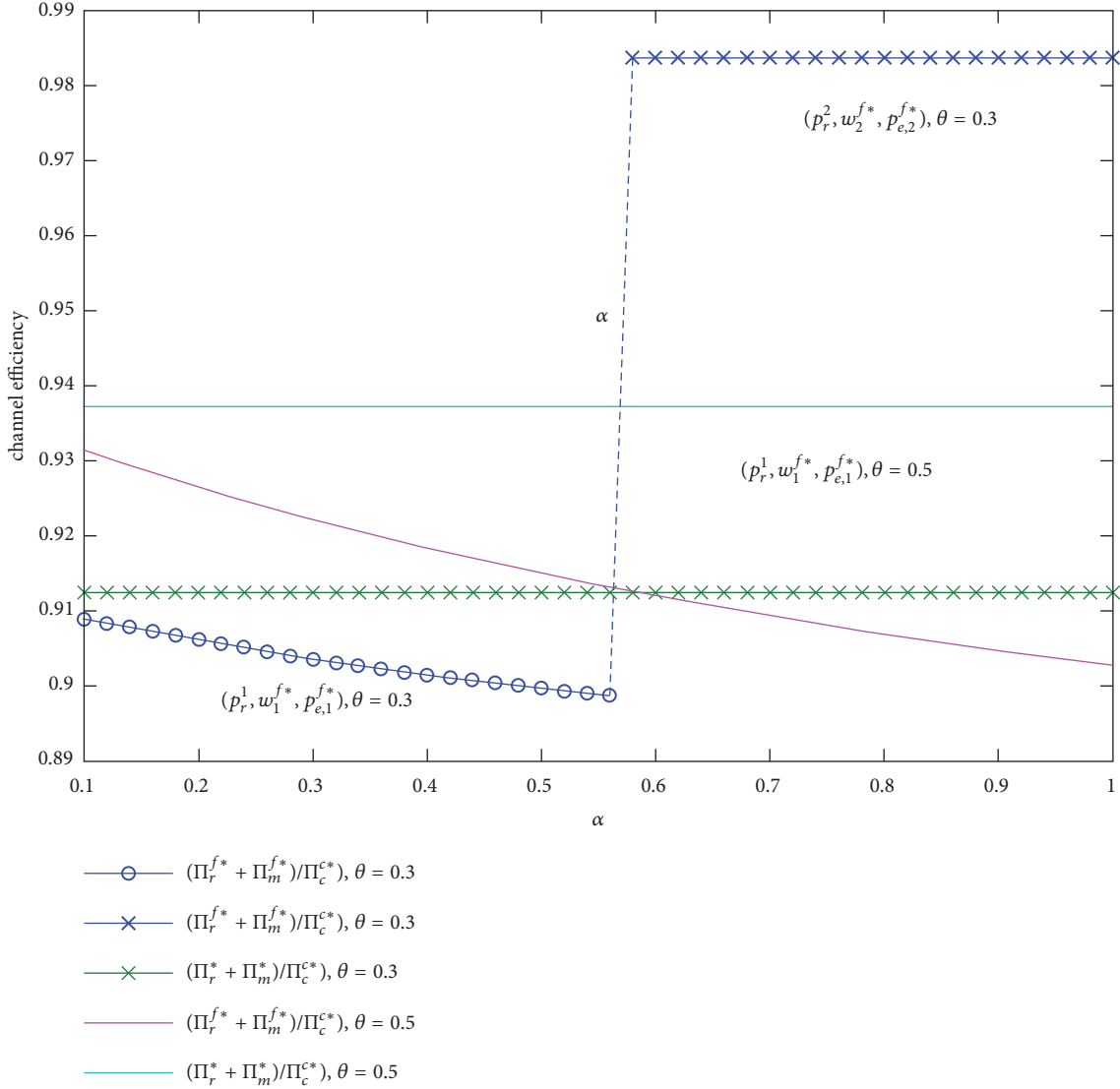


FIGURE 7: Impact of θ and α on channel efficiencies for the two scenarios.

retailers with stronger fair concern are more likely to impel the manufacturer to make a relatively fair price strategy.

Furthermore, by numerical experiments, the thesis and propositions are verified. Some findings and the corresponding management implications are given as follows.

When the cross-price influence between channels is lower, retailers that are more concerned about fairness impel manufacturers set a lower wholesale price and online direct price which is conducive to channel coordination. This phenomenon is consistent with previous inference. When the cross-price influence between channels is higher, manufacturers establish higher price strategies which are more beneficial for their own interests. Considering the retailer's fairness concerns, the cross-price effect is not always conducive to supply chain coordination, which is different from the situation facing rational retailers. Therefore, when confronting a retailer with strong fairness concerns, it is more suitable for the manufacturer to cooperate when the

cross-price impact between channels is relatively low. In contrast, among those retailers less concerned about fairness, rational retailers would be better partners.

In order to further explore the article's conclusions on the application value of supply chain management. A natural extension of our research would be to test our model predictions. For example, our analysis predicts that the manufacturer will obtain less profit when he provides a fair distribution for the retailer in the offline retail channel. In the home furnishings industry of China, the offline retailers feel that they have been unfairly treated based on the loss of profits due to direct channels. Some brands such as KUKA propose giving 10-15% of their profits to dealers [44]. To enhance offline marketing, Vivo surrenders a greater portion of the profits to offline mobile phones retailers, while some brands do not give enough profits, which results in low sales enthusiasm on the part of retailers [45, 46].

The demand functions are constructed for the dual-channel supply chain with homogeneous channel. The implication of this paper has limitations to some extent. Not all product consumers are primarily concerned with price when choosing shopping channel. Furthermore, some manufacturers adopt a heterogeneous product strategy to effectively alleviate the channel conflict. Therefore, considering the heterogeneity of channels and conducting a systematic statistical analysis of our model is a significant topic for future research.

Appendix

A. Proof of Lemma 2

In region 1, the manufacturer's problem is as follows:

$$\begin{aligned}
 \max \quad & \pi_{m1}^f = d_e(p_e - c) + d_r(w - c) \\
 \text{s.t.} \quad & p_r = \frac{w(2\alpha + 1)}{2(\alpha + 1)} + \frac{\theta p_e}{2} + \frac{a + \alpha(a - c)}{2(\alpha + 1)}, \\
 & p_e < K_1 w - D_1, \\
 & p_e \geq K_3 w - D_3, \\
 & p_e > K_4 w - D_4, \\
 & p_e < K_5 w - D_5.
 \end{aligned} \tag{A.1}$$

Therefore, the Hessian matrix of the manufacturer's profit function can be calculated.

$$H_{\pi_{m1}} \begin{bmatrix} p_e \\ w \end{bmatrix} = \begin{bmatrix} \theta^2 - 2 & \frac{\theta(3\alpha + 2)}{2(\alpha + 1)} \\ \frac{\theta(3\alpha + 2)}{2(\alpha + 1)} & \frac{-2\alpha - 1}{\alpha + 1} \end{bmatrix} \tag{A.2}$$

There is a threshold for the cross-price sensitivity coefficient, $\bar{\theta}_1 = 2\sqrt{2(2\alpha^2 + 3\alpha + 1)/(17\alpha^2 + 24\alpha + 8)}$. When $\theta \in (0, \bar{\theta}_1)$, the Hessian matrix is negative definite. The manufacturer's profit function is a concave function of p_e and w , and the decision problem is a convex optimization problem. Thus, a unique extreme point exists. If the extreme point is in the feasible region, then the extreme point is the partial equilibrium solution for the manufacturer. In contrast, the partial equilibrium solution is on the boundary of RI. If $\bar{\theta}_1 \leq \theta < 1$, the Hessian matrix is neither negative definite nor positive, which can be easily proved. In this case, there is no extreme point in the feasible region,

which suggests that the optimal solution in RI is on the boundary.

The process of deduction is as follows:

(1) For inferring the complete partial equilibrium solutions in RI, some basic work should be solved, including the extreme point, the optimal value of the boundary and the corresponding parameter range. This part is given in Appendix A.1.

(2) By comparing the results of Appendix A.1, we could deduce the partial equilibrium solution and the corresponding parameter range when the Hessian matrix is negative definite ($0 < \theta < \bar{\theta}_1$). This part is given in Appendix A.2.

(3) When $\bar{\theta}_1 \leq \theta < 1$, we should determine the optimal value of each boundary of RI and the corresponding parameter range. By comparing the results of Appendix A.1, we can obtain the corresponding partial equilibrium solutions. This part is given in Appendix A.3.

A.1. Basic Work. (1) The optimal solution on boundary L1 and the corresponding parameter range are as follows.

The manufacturer's problem is as follows:

$$\begin{aligned}
 \max_{p_e, w} \quad & \pi_{m1}^f = d_e(p_e - c) + d_r(w - c) \\
 \text{s.t.} \quad & p_r = \frac{w(2\alpha + 1)}{2(\alpha + 1)} + \frac{\theta p_e}{2} + \frac{a + \alpha(a - c)}{2(\alpha + 1)} \\
 & p_e = K_1 w - D_1, \\
 & w_C \leq w \leq w_A
 \end{aligned} \tag{A.3}$$

The optimal profit of the manufacturer on L1 is

$$\begin{aligned}
 \bar{w}_{1,1}^{f*} &= \frac{Y1a + Y2c}{X}, \\
 \bar{p}_{e1,1}^{f*} &= \frac{Y3a + Y4c}{X},
 \end{aligned} \tag{A.4}$$

where

$$\begin{aligned}
 Y1 &= -2\theta^2\alpha^2 - 4\theta^2\alpha + 2\theta\alpha^2 - 2\theta^2 + 5\theta\alpha + 4\alpha^2 \\
 &\quad + 3\theta + 10\alpha + 6, \\
 Y2 &= Y2 - (\alpha + 1)^2\theta^3 - (2\alpha^2 + 9\alpha + 7)\theta^2 + 4\alpha + 6, \\
 Y3 &= 4\theta\alpha^2 + 12\theta\alpha + 4\alpha^2 + 8\theta + 12\alpha + 9, \\
 Y4 &= -4\theta^2\alpha^2 - 10\theta^2\alpha - 6\theta^2 + 4\alpha^2 + \theta + 12\alpha + 9.
 \end{aligned} \tag{A.5}$$

When $0 < \theta < 1$, we have $w_C < \bar{w}_{1,1}^{f*} < w_A$. Then, the corresponding profit is

$$\pi_{m1}^f(\bar{w}_{1,1}^{f*}, \bar{p}_{e1,1}^{f*}) = \frac{-\left(4(\alpha + 1)^2\theta^2 + 8(\alpha^2 + 3\alpha + 2)\theta + 4\alpha^2 + 16\alpha + 13\right)(\theta c + a - c)^2}{4(4\alpha^2 + 11\alpha + 7)\theta^2 - 4(2\alpha + 3)^2}. \tag{A.6}$$

(2)The optimal solution on boundary L3 and the corresponding parameter range are as follows:

$$\begin{aligned} \max_{p_e, w} \quad & \pi_{m1}^f = d_e (p_e - c) + d_r (w - c) \\ \text{s.t} \quad & p_r^f = \frac{w(2\alpha + 1)}{2(\alpha + 1)} + \frac{\theta p_e}{2} + \frac{a + \alpha(a - c)}{2(\alpha + 1)} \\ & p_e = K_3 w - D_3, \end{aligned}$$

$$w_C \leq w \leq w_J \quad (\text{A.7})$$

$(\bar{w}_{1,3}, \bar{p}_{e,1,3})$ denotes the decision of the manufacturer on this boundary, and the remainder of the boundaries follow this method. By solving the problem, we can obtain the following:

$$\text{If } 0 < \theta \leq \bar{\theta}_9$$

$$\bar{w}_{1,3} = \bar{w}_{1,3}^{f*} = \frac{c(\alpha + 1)\theta^2 + ((4c - a)\alpha - a + 2c)\theta - (3a + 5c)\alpha - 3(a + c)}{2((\alpha + 1)\theta + 4\alpha + 3)(\theta - 1)}, \quad (\text{A.8})$$

$$\bar{p}_{e1,3} = \bar{p}_{e1,3}^{f*} = \bar{w}_{1,3}^{f*}.$$

If $\bar{\theta}_9 < \theta < 1$

$$\bar{w}_{1,3} = w_J,$$

$$\bar{p}_{e1,3} = p_{e,J}, \quad (\text{A.9})$$

$$\bar{\theta}_9 = \frac{-5\alpha - 2 + \sqrt{33\alpha^2 + 40\alpha + 16}}{2(\alpha + 1)}.$$

(3)The optimal solution on boundary L4 and the corresponding parameter range are as follows:

$$\begin{aligned} \max_{p_e, w} \quad & \pi_{m1}^f = d_e (p_e - c) + d_r (w - c) \\ \text{s.t} \quad & p_r^f = \frac{w(2\alpha + 1)}{2(\alpha + 1)} + \frac{\theta p_e}{2} + \frac{a + \alpha(a - c)}{2(\alpha + 1)} \quad (\text{A.10}) \\ & p_e = K_4 w - D_4, \\ & w_J \leq w < w_M \end{aligned}$$

(4)The optimal solution on boundary L5 and the corresponding parameter range are as follows:

$$\begin{aligned} \max_{p_e, w} \quad & \pi_{m1}^f = d_e (p_e - c) + d_r (w - c) \\ \text{s.t} \quad & p_r^f = \frac{w(2\alpha + 1)}{2(\alpha + 1)} + \frac{\theta p_e}{2} + \frac{a + \alpha(a - c)}{2(\alpha + 1)} \quad (\text{A.13}) \\ & p_e = K_5 w - D_5, \end{aligned}$$

$$w_A \leq w < w_M$$

By solving the problem, we can obtain the following:

$$\text{If } 0 < \theta^2 \leq 1 - 1/2\alpha,$$

$$\bar{w}_{1,5} = w_A, \quad (\text{A.14})$$

$$\bar{p}_{e1,5} = p_{e,A}.$$

By solving the problem, we can obtain the following:

$$\text{If } 0 < \theta < 1/(\alpha + 1),$$

$$\bar{w}_{1,4} = w_J, \quad (\text{A.11})$$

$$\bar{p}_{e1,4} = p_{e,J}.$$

If $1/(\alpha + 1) \leq \theta < 1$

$$\begin{aligned} \bar{w}_{1,4} &= \bar{w}_{1,4}^{f*} \\ &= \frac{((\alpha + 1)\theta^2 + (\alpha - 1)\theta - 2\alpha)c + (\alpha + 1)(\theta - 2)a}{(4\alpha + 2)(\theta - 1)}, \quad (\text{A.12}) \end{aligned}$$

$$\bar{p}_{e1,4} = \bar{p}_{e1,4}^{f*} = \frac{a + c - \theta c}{2(1 - \theta)}.$$

If $1 - 1/2\alpha < \theta^2 \leq 1,$

$$\begin{aligned} \bar{w}_{1,5} &= \bar{w}_{1,5}^{f*} = \frac{(3\alpha + 1)(\theta - 1)c - (\alpha + 1)a}{(4\alpha + 2)(\theta - 1)}, \\ \bar{p}_{e1,7} &= \bar{p}_{e1,5}^{f*} = \frac{(\theta - \theta^2)c + (4 - 2\theta^2 - \theta)}{2(\theta - 1)(\theta^2 - 2)}. \end{aligned} \quad (\text{A.15})$$

(5)The extreme point of the manufacturer's profit function is as follows.

By solving the equations above for w and p_e , we can obtain the extreme point as follows:

$$w_1^{f*} = \frac{-((\theta^2 + 6\theta + 4)\alpha^2 + (\theta^2 + 10\theta + 8)\alpha + 4\theta + 4)a}{X_1} + \frac{((-\theta^3 + 12\theta^2 + 2\theta - 12)\alpha^2 + (-\theta^3 + 15\theta^2 + 2\theta - 16)\alpha + 4\theta^2 - 4)c}{X_1}, \quad (\text{A.16})$$

$$p_{e,1}^{f*} = \frac{-((7\theta + 8)\alpha^2 + (11\theta + 12)\alpha + 4\theta + 4)a}{X_1} + \frac{((10\theta^2 - \theta - 8)\alpha^2 + (13\theta^2 - \theta - 12)\alpha + 4\theta^2 - 4)c}{X_1},$$

where $X_1 = 17\theta^2\alpha^2 + 24\theta^2\alpha + 8\theta^2 - 16\alpha^2 - 24\alpha - 8$.

A.2. Inferring the Partial Equilibrium Solution When $0 < \theta < \bar{\theta}_1$. We substitute w_1^{f*} and $p_{e,1}^{f*}$ into all the boundaries. “+” denotes the extreme point that satisfies the corresponding constraint; similarly, “-” indicates that the constraint is not satisfied. The results of all the constraints are summarized in Table 5.

Analysis 1 (if $\{0 < \alpha \leq \alpha_1\} \cap \{0 < \theta \leq \bar{\theta}_3\} \cup \{\alpha_1 < \alpha < 1\} \cap \{\bar{\theta}_2 < \theta \leq \bar{\theta}_3\}$). The extreme point satisfies all the constraints. Thus, the corresponding partial equilibrium solution in this situation is

$$(w_1^f, p_{e,1}^f) = (w_1^{f*}, p_{e,1}^{f*}). \quad (\text{A.17})$$

Analysis 2 (if $\{\alpha_1 < \alpha < 1\} \cap \{0 < \theta \leq \bar{\theta}_2\}$). Based on Table 5, only constraint (L3) is not satisfied, which implies that the partial equilibrium solution is on the boundary of L3. Combined with the results for L3 of Appendix A.1, we can

prove that $\bar{\theta}_2 < \bar{\theta}_9$. Therefore, we can infer the corresponding partial equilibrium solution in this situation:

$$(w_1^f, p_{e,1}^f) = (\bar{w}_{1,3}^{f*}, \bar{p}_{e,1,3}^{f*}). \quad (\text{A.18})$$

Analysis 3 (if $\{0 < \alpha < 1\} \cap \{\bar{\theta}_3 < \theta < \bar{\theta}_1\}$). Based on Table 5, only constraint (L3) is not satisfied, which implies that the partial equilibrium solution is on the boundary of L3. Combined with the result of L3 of Appendix A.1, we can prove that $1/(\alpha + 1) < \bar{\theta}_3$. Therefore, we can infer the corresponding partial equilibrium solution in this situation:

$$(w_1^f, p_{e,1}^f) = (\bar{w}_{1,4}^{f*}, \bar{p}_{e,1,4}^{f*}). \quad (\text{A.19})$$

A.3. Inferring the Partial Equilibrium Solution When $\theta \in [\bar{\theta}_1, 1)$

Analysis 4. In this case, there are no extreme points in the feasible region, which suggests that the optimal solution in R1 is on the boundary. We need to infer and compare the solutions of L1, L3, L4, and L5.

(1) For all $\alpha \in (0, 1)$, it is easy to prove that $\bar{\theta}_1 > (-5\alpha - 2 + \sqrt{33\alpha^2 + 40\alpha + 16})/2(\alpha + 1)$, $\bar{\theta}_1 > 1/(\alpha + 1)$, and $1 - 1/2\alpha < (\bar{\theta}_1)^2$. Then, we can infer the solution on each boundary, which is shown in (2)-(5).

(2)

$$\pi_{m1}(\bar{w}_{1,1}, \bar{p}_{e,1,1}) = \pi_{m1}(\bar{w}_{1,1}^{f*}, \bar{p}_{e,1,1}^{f*}) = \frac{-(4(\alpha + 1)^2\theta^2 + 8(\alpha^2 + 3\alpha + 2)\theta + 4\alpha^2 + 16\alpha + 13)(\theta c + a - c)^2}{4(4\alpha^2 + 11\alpha + 7)\theta^2 - 4(2\alpha + 3)^2}. \quad (\text{A.20})$$

(3)

$$\pi_{m1}(\bar{w}_{1,3}, \bar{p}_{e,1,3}) = \pi_{m1}(\bar{w}_J, \bar{p}_{e,J}). \quad (\text{A.21})$$

(4)

$$\begin{aligned} \pi_{m1}(\bar{w}_{1,4}, \bar{p}_{e,1,4}) &= \pi_{m1}(\bar{w}_{1,4}^{f*}, \bar{p}_{e,1,4}^{f*}) \\ &= \frac{(\theta + 1)(\theta c + a - c)^2}{4(1 - \theta)}. \end{aligned} \quad (\text{A.22})$$

(5)

$$\begin{aligned} \pi_{m1}(\bar{w}_{1,5}, \bar{p}_{e,1,5}) &= \pi_{m1}(\bar{w}_{1,5}^{f*}, \bar{p}_{e,1,5}^{f*}) \\ &= \frac{(\alpha + 1)(\theta + 1)(\theta c + a - c)^2}{4(1 - \theta)(2 - \theta^2)(2\alpha + 1)}. \end{aligned} \quad (\text{A.23})$$

Combined with (2)-(5), we can calculate it further.

TABLE 5

α	$(0, \alpha_1]$		$(\alpha_1, 1)$		
	$(0, \bar{\theta}_3]$	$(\bar{\theta}_3, \bar{\theta}_1)$	$(0, \bar{\theta}_2]$	$(\bar{\theta}_2, \bar{\theta}_3]$	$(\bar{\theta}_3, \bar{\theta}_1)$
L1	+	+	+	+	+
L3	+	+	-	+	+
L4	+	-	+	+	-
L5	+	+	+	+	+

$\bar{\theta}_2 = (-2\alpha^2 - 7\alpha - 4 + \sqrt{100\alpha^4 + 204\alpha^3 + 113\alpha^2 + 8\alpha})/4(3\alpha^2 + 4\alpha + 1)$,
 $\bar{\theta}_3 = 2(\alpha + 1)/(3\alpha + 2)$, $\bar{\theta}_1 = 2\sqrt{2(2\alpha^2 + 3\alpha + 1)/(17\alpha^2 + 24\alpha + 8)}$, and
 $\alpha_1 = 1/2$.

(6)

$$\begin{aligned} \pi_{m1}(\bar{w}_{1,3}, \bar{p}_{e1,3}) &= \pi_{m1}(\bar{w}_J, \bar{p}_{e,J}) \\ &\leq \pi_m(\bar{w}_{1,4}^{f*}, \bar{p}_{e1,4}^{f*}). \end{aligned} \quad (\text{A.24})$$

(7)

$$\begin{aligned} \pi_{m1}(\bar{w}_{1,4}^{f*}, \bar{p}_{e1,4}^{f*}) - \pi_{m1}(\bar{w}_{1,5}^{f*}, \bar{p}_{e1,5}^{f*}) \\ = \frac{(2\theta^2\alpha + \theta^2 - 3\alpha - 1)(\theta + 1)(\theta c + a - c)^2}{4(1 - \theta)(2 - \theta^2)(2\alpha + 1)} \end{aligned} \quad (\text{A.25})$$

> 0.

(8)

$$\begin{aligned} \pi_{m1}(\bar{w}_{1,4}^{f*}, \bar{p}_{e1,4}^{f*}) - \pi_{m1}(\bar{w}_{1,1}^{f*}, \bar{p}_{e1,1}^{f*}) \\ = \frac{((\alpha + 1)(3\theta^3 - 5\theta^2 - 4\theta + 4) - 2\theta)(\theta c + a - c)^2}{4(1 - \theta)((4\alpha^2 + 11\alpha + 7)\theta^2 - (2\alpha + 3)^2)} \end{aligned} \quad (\text{A.26})$$

> 0.

Combined with (6)-(8), we can confirm that the partial equilibrium solution is

$$(w_1^f, p_{e,1}^f) = (\bar{w}_{1,4}^{f*}, \bar{p}_{e1,4}^{f*}). \quad (\text{A.27})$$

By summarizing Analyses 1-4, we can obtain the partial equilibrium solution set in region R1, which is shown in Table 2.

B. Proof of Lemma 3

The problem of the manufacturer is

$$\begin{aligned} \max \quad & \pi_{m2}^f = d_e(p_e - c) + d_r(w - c) \\ \text{s.t.} \quad & p_r^f = 2w - c \\ & p_e > K_1w - D_1, \\ & p_e < K_3w - D_3, \\ & p_e > K_4w - D_4, \\ & w > c \end{aligned} \quad (\text{B.1})$$

There is a threshold for the cross-price sensitivity coefficient, $\bar{\theta}_4 = 2\sqrt{2}/3$. When $\theta \in (0, \bar{\theta}_4)$, the Hessian matrix is negatively definite. If $\bar{\theta}_4 \leq \theta < 1$, then the Hessian matrix is neither negative definite nor positive definite. Similarly, we use a method such as Lemma 2 to deduce the conclusion.

B.1. Basic Work. (1)The result is the same as Appendix A.1.

(2)The optimal solution on boundary L2 and the corresponding parameter range are as follows:

$$\begin{aligned} \bar{w}_{2,2}^{f*} &= \frac{(2\theta^2 - 3\theta - 6)a}{14\theta^2 - 18} + \frac{(2\theta^3 + 9\theta^2 - 3\theta - 12)c}{14\theta^2 - 18}, \\ \bar{p}_{e2,2}^{f*} &= \frac{-(8\theta + 9)a}{14\theta^2 - 18} + \frac{(6\theta^2 - \theta - 9)c}{14\theta^2 - 18}. \end{aligned} \quad (\text{B.2})$$

For all $\theta \in (0, 1)$, $\alpha \in (0, 1)$, $w_Q < \bar{w}_{2,2}^{f*} < w_E$, $\alpha \in (0, 1)$.

(3)The optimal solution on boundary L3 and the corresponding parameter range are as follows:

$$\begin{aligned} (\bar{w}_{2,3}, \bar{p}_{e2,3}) &= \begin{cases} (\bar{w}_{2,3}^{f*}, \bar{p}_{e2,3}^{f*}) & \text{if } 0 < \theta \leq \frac{\alpha}{2(\alpha + 1)} \\ (w_C, p_{e,C}) & \text{if } \frac{\alpha}{2(\alpha + 1)} < \theta < 1 \end{cases} \\ \bar{w}_{2,3}^* &= \frac{a}{3 - 3\theta} + \frac{2}{3}c, \\ \bar{p}_{e2,3}^* &= \frac{a}{3 - 3\theta} + \frac{2}{3}c. \end{aligned} \quad (\text{B.3})$$

(4)The optimal solution on boundary L6 and the corresponding parameter range are as follows:

$$\begin{aligned} (\bar{w}_{2,6}, \bar{p}_{e2,6}) &= \begin{cases} (\bar{w}_{2,6}^{f*}, \bar{p}_{e2,6}^{f*}) & \text{if } 0 < \theta \leq \frac{\sqrt{2(\alpha + 1)(2\alpha + 1)}}{2(\alpha + 1)} \\ (w_A, p_{e,A}) & \text{if } \frac{\sqrt{2(\alpha + 1)(2\alpha + 1)}}{2(\alpha + 1)} < \theta < 1 \end{cases} \end{aligned} \quad (\text{B.4})$$

$$\bar{w}_{2,6}^{f*} = \frac{a}{4 - 4\theta} + \frac{3}{4}c,$$

$$\bar{p}_{e2,6}^{f*} = \frac{(2 - \theta)a}{2 - 2\theta} + \frac{\theta}{2}c.$$

(5)The extreme point of the manufacturer's profit function is as follows:

$$\begin{aligned} w_2^{f*} &= \frac{(3\theta + 2)a}{8 - 9\theta^2} - \frac{(6\theta^2 + \theta - 6)c}{8 - 9\theta^2}; \\ p_{e2}^{f*} &= \frac{(3\theta + 4)a}{8 - 9\theta^2} - \frac{(6\theta^2 - \theta - 4)c}{8 - 9\theta^2}. \end{aligned} \quad (\text{B.5})$$

B.2. Inferring the Partial Equilibrium Solution When $0 < \theta < \bar{\theta}_4$.

We substitute (w_2^{f*}, p_{e2}^{f*}) into each boundary. “+” indicates that the extreme point satisfies the corresponding constraint; similarly, “-” indicates that the constraint is not satisfied. The results of all the constraints are summarized in Table 6.

TABLE 6

θ	$(0, \bar{\theta}_6]$	$(\bar{\theta}_6, \bar{\theta}_5]$	$(\bar{\theta}_5, \bar{\theta}_4)$
L1	+	+	-
L2	-	+	+
L3	+	+	+
L6	+	+	+

$\bar{\theta}_4 = 2\sqrt{2}/3$, $\bar{\theta}_5 = (-2\alpha - 5 + \sqrt{100\alpha^2 + 164\alpha + 73})/12(\alpha + 1)$, and $\bar{\theta}_6 = (-5 + \sqrt{73})/12$.

We analyse the equilibrium solution of different parameter ranges and use $(w_2^f, p_{e,2}^f)$ to denote the corresponding partial equilibrium solutions in R2.

Analysis 1 (if $\{0 < \theta \leq \bar{\theta}_6\} \cap \{0 < \alpha < 1\}$). Based on Table 6, only constraint (L2) is not satisfied, which implies that the partial equilibrium solution is on the boundary of L3. Combined with the results for (2) of Appendix B.1, we can infer the corresponding partial equilibrium solution in this situation:

$$(w_2^f, p_{e,2}^f) = (\bar{w}_{2,2}^{f*}, \bar{p}_{e,2,2}^{f*}) \quad (\text{B.6})$$

Analysis 2 (if $\{\bar{\theta}_6 < \theta \leq \bar{\theta}_5\} \cap \{0 < \alpha < 1\}$). Based on Table 5, all the constraints are satisfied, which implies that the extreme point is the partial equilibrium solution.

$$(w_2^f, p_{e,2}^f) = (w_2^{f*}, p_{e,2}^{f*}). \quad (\text{B.7})$$

Analysis 3 (if $\{\bar{\theta}_5 < \theta < \bar{\theta}_4\} \cap \{0 < \alpha < 1\}$). Based on Table 6, only constraint (L1) is not satisfied, which implies that the partial equilibrium solution is on the boundary of

$$\pi_{m,2}(\bar{w}_{2,1}^*, \bar{p}_{e,2,1}^*) - \pi_{m,2}(\bar{w}_{2,2}^*, \bar{p}_{e,2,2}^*) = \frac{(\theta c + a - c)^2 \alpha ((\alpha + 1)(12\theta^4 - 8\theta^3 - 44\theta^2 - 8\theta + 16) - 11\theta^2 - 16\theta - 4)}{112((\alpha^2 + (11/4)\alpha + 7/4)\theta^2 - (\alpha + 3/2)^2)(9/7 - \theta^2)} > 0. \quad (\text{B.12})$$

Combined with (1)-(4), the partial equilibrium solution is

$$(w_2^f, p_{e,2}^f) = (\bar{w}_{2,1}^{f*}, \bar{p}_{e,2,1}^{f*}). \quad (\text{B.13})$$

To summarize Analyses 1–4, we can obtain the partial equilibrium solution set in region R2, which are shown in Table 3.

C. Proof of Lemma 4

The manufacturer's problem in R3 is

$$\begin{aligned} \max_{p_e, w} \quad & \pi_{m3}^f = d_e(p_e - c) + d_r(w - c) \\ \text{s.t.} \quad & p_r = \frac{\theta p_e + w + a}{2}, \\ & p_e \geq K_2 w - D_2, \end{aligned}$$

L1. Combined with the result of Appendix B.1 (1), we can infer the corresponding partial equilibrium solution in this situation:

$$(w_2^f, p_{e,2}^f) = (\bar{w}_{2,1}^{f*}, \bar{p}_{e,2,1}^{f*}) \quad (\text{B.8})$$

B.3. Inferring the Partial Equilibrium Solution When $0 < \theta < \bar{\theta}_4$

Analysis 4 (if $\{\bar{\theta}_4 \leq \theta < 1\} \cap \{0 < \alpha < 1\}$). Similarly, with R1, we should determine and compare the optimal solutions for L1, L2, L3, and L4.

(1)

$$\alpha \in (0, 1),$$

$$\frac{2\sqrt{2}}{3} > \frac{\sqrt{2}(\alpha + 1)(2\alpha + 1)}{2(\alpha + 1)}, \quad (\text{B.9})$$

$$\frac{2\sqrt{2}}{3} > \frac{\alpha}{2(\alpha + 1)}.$$

(2)

$$\begin{aligned} \pi_{m,2}(\bar{w}_{2,3}, \bar{p}_{e,2,3}) &= \pi_{m,2}(w_C, p_{e,C}) \\ &< \pi_{m,2}(\bar{w}_{2,3}^*, \bar{p}_{e,2,3}^*). \end{aligned} \quad (\text{B.10})$$

(3)

$$\begin{aligned} \pi_{m,2}(\bar{w}_{2,6}, \bar{p}_{e,2,6}) &= \pi_{m,2}(w_A, p_{e,A}) \\ &< \pi_{m,2}(\bar{w}_{2,3}^*, \bar{p}_{e,2,3}^*). \end{aligned} \quad (\text{B.11})$$

(4)

$$p_e \geq K_3 w - D_3,$$

$$p_e \leq K_7 w - D_7,$$

$$w > c$$

(C.1)

The Hessian matrix $H_{\pi_m} [\frac{p_e}{w}]$ is negative definite. The manufacturer's profit function is a concave function of p_e and w , and the decision problem is a convex optimization problem. Thus, a unique extreme point exists as follows:

$$p_{e,3}^{f*} = w_3^{f*} = \frac{a}{2(1 - \theta)} + \frac{c}{2} \quad (\text{C.2})$$

Analysis 1. If $(w_3^{f*}, p_{e,3}^{f*})$ is brought into each boundary, then only L2 is not satisfied. Thus, we can deduce that the partial equilibrium solution of R3 is on L2. The optimal point for the

L2 of R2 and R3 is the same. Combined with Appendix B.1, the partial equilibrium solution is

$$(w_3^f, p_{e,3}^f) = (\bar{w}_{3,2}^{f*}, \bar{p}_{e,3,2}^{f*}) = (\bar{w}_{2,2}^{f*}, \bar{p}_{e,2,2}^{f*}). \quad (C.3)$$

D. Proof of Proposition 5

We have obtained the partial equilibrium solution set of R1, R2, and R3. The manufacturer would compare the solutions for the different regions that are in the same parameter range and choose the optimum solution as the equilibrium pricing strategy. The process of inference is as follows:

(1) By comparing the different thresholds in Lemmas 2–4, we can divide the parameter ranges that the solution needs to compare. This part is given in Appendix D.1.

(2) Based on the results of Appendix D.1, we determine the complete equilibrium solution. This part is given in Appendices D.2 and D.3.

D.1. Divide the Parameter Ranges That the Solution Needs to Compare. Combined with Lemmas 2, 3, and 4; the relevant thresholds are summarized as follows:

$$\begin{aligned} \bar{\theta}_1 &= 2\sqrt{\frac{2(2\alpha^2 + 3\alpha + 1)}{17\alpha^2 + 24\alpha + 8}}, \\ \bar{\theta}_2 &= \frac{-2\alpha^2 - 7\alpha - 4 + \sqrt{100\alpha^4 + 204\alpha^3 + 113\alpha^2 + 8\alpha}}{4(3\alpha^2 + 4\alpha + 1)}, \\ \bar{\theta}_3 &= \frac{2(\alpha + 1)}{3\alpha + 2}, \\ \bar{\theta}_4 &= \frac{2\sqrt{2}}{3}, \\ \bar{\theta}_5 &= \frac{-2\alpha - 5 + \sqrt{100\alpha^2 + 164\alpha + 73}}{12(\alpha + 1)}, \\ \bar{\theta}_6 &= \frac{-5 + \sqrt{73}}{12}, \\ \alpha_1 &= \frac{1}{2}. \end{aligned} \quad (D.1)$$

$$\max(\bar{\theta}_3, \bar{\theta}_4) \begin{cases} \bar{\theta}_3 & 0 < \alpha \leq \frac{3 - 2\sqrt{2}}{3(\sqrt{2} - 1)} \\ \bar{\theta}_4 & \frac{3 - 2\sqrt{2}}{3(\sqrt{2} - 1)} < \alpha < 1; \end{cases} \quad (D.2)$$

$$\max(\bar{\theta}_2, 0) \begin{cases} 0 & 0 < \alpha \leq \alpha_1 \\ \bar{\theta}_2 & \alpha_1 < \alpha < 1. \end{cases}$$

(2) $\bar{\theta}_2 < \bar{\theta}_6 < \bar{\theta}_5 < \bar{\theta}_1 < 1$; $\bar{\theta}_5 < \bar{\theta}_4 < \bar{\theta}_1$; $\bar{\theta}_5 < \bar{\theta}_4 < \bar{\theta}_3$. Combined with (1) and (2), we can conclude

TABLE 7

region	θ				
	$(0, \bar{\theta}_6]$	$(\bar{\theta}_6, \bar{\theta}_5]$	$(\bar{\theta}_5, \bar{\theta}_4]$	$(\bar{\theta}_4, \bar{\theta}_3)$	$[\bar{\theta}_3, 1)$
R1	$w_1^{f*}, p_{e,1}^{f*}$	$w_1^{f*}, p_{e,1}^{f*}$	$w_1^{f*}, p_{e,1}^{f*}$	$w_1^{f*}, p_{e,1}^{f*}$	$\bar{w}_{1,4}^{f*}, \bar{p}_{e,1,4}^{f*}$
R2	$\bar{w}_{2,2}^{f*}, \bar{p}_{e,2,2}^{f*}$	$w_2^{f*}, p_{e,2}^{f*}$	$\bar{w}_{2,1}^{f*}, \bar{p}_{e,2,1}^{f*}$	$\bar{w}_{2,1}^{f*}, \bar{p}_{e,2,1}^{f*}$	$\bar{w}_{2,1}^{f*}, \bar{p}_{e,2,1}^{f*}$
R3	$\bar{w}_{3,2}^{f*}, \bar{p}_{e,3,2}^{f*}$	$\bar{w}_{3,2}^{f*}, \bar{p}_{e,3,2}^{f*}$	$\bar{w}_{3,2}^{f*}, \bar{p}_{e,3,2}^{f*}$	$\bar{w}_{3,2}^{f*}, \bar{p}_{e,3,2}^{f*}$	$\bar{w}_{3,2}^{f*}, \bar{p}_{e,3,2}^{f*}$

TABLE 8

region	θ				
	$(0, \bar{\theta}_6]$	$(\bar{\theta}_6, \bar{\theta}_5]$	$(\bar{\theta}_5, \bar{\theta}_3)$	$(\bar{\theta}_3, \bar{\theta}_4)$	$[\bar{\theta}_4, 1)$
R1	$w_1^{f*}, p_{e,1}^{f*}$	$w_1^{f*}, p_{e,1}^{f*}$	$w_1^{f*}, p_{e,1}^{f*}$	$\bar{w}_{1,4}^{f*}, \bar{p}_{e,1,4}^{f*}$	$\bar{w}_{1,4}^{f*}, \bar{p}_{e,1,4}^{f*}$
R2	$\bar{w}_{2,2}^{f*}, \bar{p}_{e,2,2}^{f*}$	$w_2^{f*}, p_{e,2}^{f*}$	$\bar{w}_{2,1}^{f*}, \bar{p}_{e,2,1}^{f*}$	$\bar{w}_{2,1}^{f*}, \bar{p}_{e,2,1}^{f*}$	$\bar{w}_{2,1}^{f*}, \bar{p}_{e,2,1}^{f*}$
R3	$\bar{w}_{3,2}^{f*}, \bar{p}_{e,3,2}^{f*}$	$\bar{w}_{3,2}^{f*}, \bar{p}_{e,3,2}^{f*}$	$\bar{w}_{3,2}^{f*}, \bar{p}_{e,3,2}^{f*}$	$\bar{w}_{3,2}^{f*}, \bar{p}_{e,3,2}^{f*}$	$\bar{w}_{3,2}^{f*}, \bar{p}_{e,3,2}^{f*}$

(3) if $0 < \alpha \leq (3 - 2\sqrt{2})/3(\sqrt{2} - 1) \implies 0 < \bar{\theta}_6 < \bar{\theta}_5 < \bar{\theta}_4 < \bar{\theta}_3 < \bar{\theta}_1 < 1$.

The problem of the manufacturer's decision is shown in Table 7:

(4) if $(3 - 2\sqrt{2})/3(\sqrt{2} - 1) < \alpha \leq \alpha_1 \implies 0 < \bar{\theta}_6 < \bar{\theta}_5 < \bar{\theta}_3 < \bar{\theta}_4 < \bar{\theta}_1 < 1$.

The problem of the manufacturer's decision is shown in Table 8:

(5) if $\alpha_1 < \alpha < 1 \implies 0 < \bar{\theta}_2 < \bar{\theta}_6 < \bar{\theta}_5 < \bar{\theta}_3 < \bar{\theta}_4 < \bar{\theta}_1 < 1$.

The problem of the manufacturer's decision is shown in Table 9.

D.2. R3 and R2: Comparison of the Partial Equilibrium Solutions of R3 and R2

Analysis 1 (if $\{0 < \theta \leq \bar{\theta}_6\} \cap \{0 < \alpha < 1\}$). As $\bar{w}_{3,2}^{f*} = \bar{w}_{2,2}^{f*}$, $\bar{p}_{e,3,2}^{f*} = \bar{p}_{e,2,2}^{f*}$, the partial equilibrium solutions of R3 and R2 are the same.

Analysis 2 (if $\{\bar{\theta}_6 < \theta \leq \bar{\theta}_5\} \cap \{0 < \alpha < 1\}$). It is easy to prove that $\pi_{m,2}(w_2^{f*}, p_{e,2}^{f*}) - \pi_{m,3}(\bar{w}_{3,2}^{f*}, \bar{p}_{e,3,2}^{f*}) > 0$. Thus, the partial equilibrium solutions of R2 are better than those for R3.

Analysis 3 (if $\{\bar{\theta}_5 < \theta < 1\} \cap \{0 < \alpha < 1\}$).

$$\begin{aligned} &\pi_{m,2}(\bar{w}_{2,1}^{f*}, \bar{p}_{e,2,1}^{f*}) - \pi_{m,2}(\bar{w}_{3,2}^{f*}, \bar{p}_{e,3,2}^{f*}) \\ &= \frac{(\theta c + a - c)^2 \alpha (Y5\alpha + Y6)}{112((\alpha^2 + (11/4)\alpha + 7/4)\theta^2 - (\alpha + 3/2)^2)(9/7 - \theta^2)} \end{aligned} \quad (D.3)$$

where $Y5 = 12\theta^2 - 8\theta^3 - 44\theta^2 - 8\theta + 16$; $Y6 = 12\theta^2 - 8\theta^3 - 55\theta^2 - 24\theta + 12$.

TABLE 9

region	θ					
	$(0, \bar{\theta}_2]$	$(\bar{\theta}_2, \bar{\theta}_6]$	$(\bar{\theta}_6, \bar{\theta}_5]$	$(\bar{\theta}_5, \bar{\theta}_3)$	$(\bar{\theta}_3, \bar{\theta}_4)$	$(\bar{\theta}_4, 1)$
R1	$\bar{w}_{1,1}^{f*}, \bar{p}_{e,1,1}^{f*}$	$w_1^{f*}, p_{e,1}^{f*}$	$w_1^{f*}, p_{e,1}^{f*}$	$w_1^{f*}, p_{e,1}^{f*}$	$\bar{w}_{1,4}^{f*}, \bar{p}_{e,1,4}^{f*}$	$\bar{w}_{1,4}^{f*}, \bar{p}_{e,1,4}^{f*}$
R2	$\bar{w}_{2,2}^{f*}, \bar{p}_{e,2,2}^{f*}$	$\bar{w}_{2,2}^{f*}, \bar{p}_{e,2,2}^{f*}$	$w_2^{f*}, p_{e,2}^{f*}$	$\bar{w}_{2,1}^{f*}, \bar{p}_{e,2,1}^{f*}$	$\bar{w}_{2,1}^{f*}, \bar{p}_{e,2,1}^{f*}$	$\bar{w}_{2,1}^{f*}, \bar{p}_{e,2,1}^{f*}$
R3	$\bar{w}_{3,2}^{f*}, \bar{p}_{e,3,2}^{f*}$	$\bar{w}_{3,2}^{f*}, \bar{p}_{e,3,2}^{f*}$	$\bar{w}_{3,2}^{f*}, \bar{p}_{e,3,2}^{f*}$	$\bar{w}_{3,2}^{f*}, \bar{p}_{e,3,2}^{f*}$	$\bar{w}_{3,2}^{f*}, \bar{p}_{e,3,2}^{f*}$	$\bar{w}_{3,2}^{f*}, \bar{p}_{e,3,2}^{f*}$

TABLE 10

region	θ			
	$(0, \bar{\theta}_6]$	$(\bar{\theta}_6, \bar{\theta}_5]$	$(\bar{\theta}_5, \bar{\theta}_3]$	$(\bar{\theta}_3, 1)$
R1	$w_1^{f*}, p_{e,1}^{f*}$	$w_1^{f*}, p_{e,1}^{f*}$	$w_1^{f*}, p_{e,1}^{f*}$	$\bar{w}_{1,4}^{f*}, \bar{p}_{e,1,4}^{f*}$
R2	$\bar{w}_{2,2}^{f*}, \bar{p}_{e,2,2}^{f*}$	$w_2^{f*}, p_{e,2}^{f*}$	$\bar{w}_{2,1}^{f*}, \bar{p}_{e,2,1}^{f*}$	$\bar{w}_{2,1}^{f*}, \bar{p}_{e,2,1}^{f*}$

TABLE 11

region	θ				
	$(0, \bar{\theta}_2]$	$(\bar{\theta}_2, \bar{\theta}_6]$	$(\bar{\theta}_6, \bar{\theta}_5]$	$(\bar{\theta}_5, \bar{\theta}_3)$	$(\bar{\theta}_3, 1)$
R1	$\bar{w}_{1,1}^{f*}, \bar{p}_{e,1,1}^{f*}$	$w_1^{f*}, p_{e,1}^{f*}$	$w_1^{f*}, p_{e,1}^{f*}$	$w_1^{f*}, p_{e,1}^{f*}$	$\bar{w}_{1,4}^{f*}, \bar{p}_{e,1,4}^{f*}$
R2	$\bar{w}_{2,2}^{f*}, \bar{p}_{e,2,2}^{f*}$	$\bar{w}_{2,2}^{f*}, \bar{p}_{e,2,2}^{f*}$	$w_2^{f*}, p_{e,2}^{f*}$	$\bar{w}_{2,1}^{f*}, \bar{p}_{e,2,1}^{f*}$	$\bar{w}_{2,1}^{f*}, \bar{p}_{e,2,1}^{f*}$

For all $\bar{\theta}_5 < \theta < 1$ and $0 < \alpha < 1$, we can prove that $(\alpha^2 + (11/4)\alpha + 7/4)\theta^2 - (\alpha + 3/2)^2 < 0$, $Y1\alpha + Y2 < 0$. Therefore, we conclude that $\pi_{m,2}(\bar{w}_{2,1}^{f*}, \bar{p}_{e,2,1}^{f*}) > \pi_{m,2}(\bar{w}_{3,2}^{f*}, \bar{p}_{e,3,2}^{f*})$. The manufacturer would choose the solution of R2.

Analysis 4. By combining Analysis 1–3, we find that the partial equilibrium solution of R2 is always more profitable than R3. Combined with Appendix D.1 (4)–(6), the problem of the manufacturer's decision can be simplified and is shown in Tables 10 and 11.

(1) If $0 < \alpha \leq \alpha_1$. See Table 10.

(2) If $\alpha_1 < \alpha < 1$. See Table 11.

$$\alpha^+ = \frac{-2(\theta + 1)(3\theta^3 + 16\theta^2 + 9\theta - 6 + \sqrt{-3\theta^6 - 4\theta^5 + 57\theta^4 + 2\theta^3 - 164\theta^2 - 48\theta + 64})}{12\theta^4 + 76\theta^3 + 89\theta^2 - 4\theta - 28},$$

$$\alpha^- = \frac{-2(\theta + 1)(3\theta^3 + 16\theta^2 + 9\theta - 6 - \sqrt{-3\theta^6 - 4\theta^5 + 57\theta^4 + 2\theta^3 - 164\theta^2 - 48\theta + 64})}{12\theta^4 + 76\theta^3 + 89\theta^2 - 4\theta - 28}.$$

(1) If $0 < \alpha \leq \alpha^+$, $\pi_{m,1}(w_1^{f*}, p_{e,1}^{f*}) - \pi_{m,2}(\bar{w}_{2,2}^{f*}, \bar{p}_{e,2,2}^{f*}) > 0$;
if $\alpha^+ < \alpha < 1$, $\pi_{m,1}(w_1^{f*}, p_{e,1}^{f*}) - \pi_{m,2}(\bar{w}_{2,2}^{f*}, \bar{p}_{e,2,2}^{f*}) < 0$.

We make $\alpha_2 = \alpha^+$. We need to compare the size of α_2 and

α_1 .

(2)

$$\max(\alpha_2, \alpha_1) \begin{cases} \alpha_2 & \text{if } 0 < \theta < \bar{\theta}_2 \\ \alpha_1 & \text{if } \theta = \bar{\theta}_6 \end{cases} \quad (\text{D.7})$$

D.3. Comparison of the Equilibrium Solutions of R2 and R1. By comparing the solutions of R2 and R3, the more profitable solutions are the final equilibrium solutions for the manufacturer and the retailer. We use (w^{f**}, p_e^{f**}) to denote them. Based on Tables 10 and 11, the results are as follows.

Analysis 1. If $\{0 < \theta \leq \bar{\theta}_6\} \cap \{0 < \alpha \leq 1/2\} \cup \{\bar{\theta}_2 < \theta \leq \bar{\theta}_6\} \cap \{1/2 < \alpha < 1\}$, compare $(w_1^{f*}, p_{e,1}^{f*})$ with $(\bar{w}_{2,2}^{f*}, \bar{p}_{e,2,2}^{f*})$.

$$\pi_{m,1}(w_1^{f*}, p_{e,1}^{f*}) - \pi_{m,2}(\bar{w}_{2,2}^{f*}, \bar{p}_{e,2,2}^{f*}) = \frac{(\theta c + a - c)^2 \alpha (Y7\alpha^2 + Y8\alpha + Y9)}{(17\theta^2\alpha^2 + 24\theta^2\alpha + 8\theta^2 - 16\alpha^2 - 24\alpha - 8)(7\theta^2 - 9)}$$

where

$$Y7 = 3\theta^4 + 19\theta^3 + \frac{89}{4}\theta^2 - \theta - 7;$$

$$Y8 = 3\theta^4 + 19\theta^3 + 25\theta^2 + 3\theta - 6; \quad (\text{D.5})$$

$$Y9 = \theta^4 + 4\theta^3 + 6\theta^2 + 4\theta + 1.$$

For all $0 < \theta \leq \bar{\theta}_6$, $17\theta^2\alpha^2 + 24\theta^2\alpha + 8\theta^2 - 16\alpha^2 - 24\alpha - 8 > 0$, $Y1 < 0$.

By making $Y7\alpha^2 + Y8\alpha + Y9 = 0$, we can obtain

$(\alpha_2 - \alpha_1)$ is a monotone increasing function of θ . There exists a threshold $\bar{\theta}_7$ ($\bar{\theta}_7 \in (\bar{\theta}_2, \bar{\theta}_6)$). If $\theta \in (0, \bar{\theta}_7)$, $\alpha_2 < \alpha_1$; if $\theta \in [\bar{\theta}_7, \bar{\theta}_6)$, $\alpha_1 \leq \alpha_2 < 1$.

Solution 1 of Analysis 1: If $\{0 < \theta \leq \bar{\theta}_6\} \cap \{0 < \alpha \leq \alpha_1\}$. Simultaneously, $0 < \theta \leq \bar{\theta}_7$, $0 < \alpha \leq \alpha_2$, and $\pi_{m,1}(w_1^{f*}, p_{e,1}^{f*}) - \pi_{m,2}(\bar{w}_{2,2}^{f*}, \bar{p}_{e,2,2}^{f*}) \geq 0$; the equilibrium solution is

$$(w^{f**}, p_e^{f**}) = (w_1^{f*}, p_{e,1}^{f*}) \quad (\text{E1})$$

Simultaneously, $0 < \theta \leq \bar{\theta}_7$, $\alpha_2 < \alpha \leq \alpha_1$, and $\pi_{m.1}(w_1^{f*}, p_{e1}^{f*}) - \pi_{m.2}(\bar{w}_{2,2}^{f*}, \bar{p}_{e,2,2}^{f*}) \leq 0$; the equilibrium solution is

$$(w^{f**}, p_e^{f**}) = (\bar{w}_{2,2}^{f*}, \bar{p}_{e,2,2}^{f*}) \quad (E2)$$

Simultaneously, $\bar{\theta}_7 < \theta \leq \bar{\theta}_6$, $0 < \alpha \leq \alpha_2$, and $\pi_{m.1}(w_1^{f*}, p_{e1}^{f*}) - \pi_{m.2}(\bar{w}_{2,2}^{f*}, \bar{p}_{e,2,2}^{f*}) \geq 0$; the equilibrium solution is

$$(w^{f**}, p_e^{f**}) = (w_1^{f*}, p_{e,1}^{f*}) \quad (E3)$$

Solution 2 of Analysis 1: If $\{\bar{\theta}_2 < \theta \leq \bar{\theta}_6\} \cap \{\alpha_1 < \alpha < 1\}$. Simultaneously, $\bar{\theta}_2 < \theta \leq \bar{\theta}_7$, $\alpha_1 < \alpha < 1$, and $\pi_{m.1}(w_1^{f*}, p_{e1}^{f*}) - \pi_{m.2}(\bar{w}_{2,2}^{f*}, \bar{p}_{e,2,2}^{f*}) < 0$; the equilibrium solution is

$$(w^{f**}, p_e^{f**}) = (\bar{w}_{2,2}^{f*}, \bar{p}_{e,2,2}^{f*}) \quad (E4)$$

Simultaneously, $\bar{\theta}_7 < \theta \leq \bar{\theta}_6$, $\alpha_1 < \alpha \leq \alpha_2$; $\pi_{m.1}(w_1^{f*}, p_{e1}^{f*}) - \pi_{m.2}(\bar{w}_{2,2}^{f*}, \bar{p}_{e,2,2}^{f*}) \geq 0$; the equilibrium solution is

$$(w^{f**}, p_e^{f**}) = (w_1^{f*}, p_{e,1}^{f*}) \quad (E5)$$

Simultaneously, $\bar{\theta}_7 < \theta \leq \bar{\theta}_6$, $\alpha_2 < \alpha < 1$, and $\pi_{m.1}(w_1^{f*}, p_{e1}^{f*}) - \pi_{m.2}(\bar{w}_{2,2}^{f*}, \bar{p}_{e,2,2}^{f*}) < 0$; the equilibrium solution is

$$(w^{f**}, p_e^{f**}) = (\bar{w}_{2,2}^{f*}, \bar{p}_{e,2,2}^{f*}) \quad (E6)$$

Analysis 2. If $\{0 < \theta \leq \bar{\theta}_2\} \cap \{\alpha_1 < \alpha < 1\}$, then compare $(\bar{w}_{1,1}^{f*}, \bar{p}_{e,1,1}^{f*})$ with $(\bar{w}_{2,2}^{f*}, \bar{p}_{e,2,2}^{f*})$.

Solution of Analysis 2. It is easy to prove that $\pi_{m.1}(\bar{w}_{1,1}^{f*}, \bar{p}_{e,1,1}^{f*}) - \pi_{m.2}(\bar{w}_{2,2}^{f*}, \bar{p}_{e,2,2}^{f*}) < 0$; the equilibrium solution is

$$(w^{f**}, p_e^{f**}) = (\bar{w}_{2,2}^{f*}, \bar{p}_{e,2,2}^{f*}) \quad (E7)$$

Analysis 3. If $\{\bar{\theta}_6 < \theta \leq \bar{\theta}_5\} \cap \{0 < \alpha < 1\}$, compare $(w_1^{f*}, p_{e,1}^{f*})$ with $(w_2^{f*}, p_{e,2}^{f*})$.

$$\begin{aligned} & \pi_{m.1}(w_1^{f*}, p_{e,1}^{f*}) - \pi_{m.2}(w_2^{f*}, p_{e,2}^{f*}) \\ &= \frac{(\theta c + a - c)^2 (\theta + 1) (Y10\alpha^2 + Y11\alpha + Y12)}{(17\theta^2\alpha^2 + 24\theta^2\alpha + 8\theta^2 - 16\alpha^2 - 24\alpha - 8)(8 - 9\theta^2)} \end{aligned} \quad (D.8)$$

where

$$\begin{aligned} Y10 &= 18\theta^3 - 6\theta^2 - 16\theta + 8; \\ Y11 &= 27\theta^3 - 24\theta + 8; \\ Y12 &= 9\theta^3 + 3\theta^2 - 8\theta. \end{aligned} \quad (D.9)$$

It is easy to prove that $Y10 > 0$ under this condition. By making $Y10\alpha^2 + Y11\alpha + Y12$ equal to zero, we can obtain

$$\begin{aligned} \alpha^+ &= \frac{-9\theta^2 - 6\theta + 4 + \sqrt{9\theta^4 + 12\theta^3 + 4\theta^2 + 16\theta + 16}}{(\theta + 1)(3\theta - 2)}, \\ \alpha^- &= \frac{-9\theta^2 - 6\theta + 4 - \sqrt{9\theta^4 + 12\theta^3 + 4\theta^2 + 16\theta + 16}}{(\theta + 1)(3\theta - 2)}. \end{aligned} \quad (D.10)$$

Make $\alpha_3 = \alpha^+$; for all $\bar{\theta}_6 < \theta \leq \bar{\theta}_5$, $\alpha_3 > 0$.

(1) If $0 < \alpha \leq \alpha_3$, then $\pi_{m.1}(w_1^{f*}, p_{e,1}^{f*}) - \pi_{m.2}(w_2^{f*}, p_{e,2}^{f*}) \leq 0$.

If $\alpha_3 < \alpha$, then $\pi_{m.1}(w_1^{f*}, p_{e,1}^{f*}) - \pi_{m.2}(w_2^{f*}, p_{e,2}^{f*}) > 0$.

(2) There exists a threshold $\bar{\theta}_8 \in (\bar{\theta}_6, \bar{\theta}_5)$; if $\bar{\theta}_6 < \theta \leq \bar{\theta}_8$, $\alpha_3 < 1$; if $\bar{\theta}_8 < \theta \leq \bar{\theta}_5$, $\alpha_3 > 1$.

Solution of Analysis 3. $\bar{\theta}_6 < \theta \leq \bar{\theta}_5$; $0 < \alpha < 1$. Simultaneously, $\bar{\theta}_6 < \theta \leq \bar{\theta}_8$, $0 < \alpha \leq \alpha_3$, and $\pi_{m.1}(w_1^{f*}, p_{e,1}^{f*}) - \pi_{m.2}(w_2^{f*}, p_{e,2}^{f*}) \geq 0$; the equilibrium solution is

$$(w^{f**}, p_e^{f**}) = (w_1^{f*}, p_{e,1}^{f*}) \quad (E8)$$

Simultaneously, $\bar{\theta}_6 < \theta \leq \bar{\theta}_8$, $\alpha_3 < \alpha < 1$, and $\pi_{m.1}(w_1^{f*}, p_{e,1}^{f*}) - \pi_{m.2}(w_2^{f*}, p_{e,2}^{f*}) < 0$; the equilibrium solution is

$$(w^{f**}, p_e^{f**}) = (w_2^{f*}, p_{e,2}^{f*}) \quad (E9)$$

Simultaneously, $\bar{\theta}_8 < \theta \leq \bar{\theta}_5$, $0 < \alpha < 1$, and $\pi_{m.1}(w_1^{f*}, p_{e,1}^{f*}) - \pi_{m.2}(w_2^{f*}, p_{e,2}^{f*}) > 0$; the equilibrium solution is

$$(w^{f**}, p_e^{f**}) = (w_1^{f*}, p_{e,1}^{f*}) \quad (E10)$$

Analysis 4. If $\{\bar{\theta}_5 < \theta \leq \bar{\theta}_3\} \cap \{0 < \alpha < 1\}$, then compare $(w_1^{f*}, p_{e,1}^{f*})$ with $(w_{2,1}^{f*}, p_{e,2,1}^{f*})$.

Solution of Analysis 4: $\bar{\theta}_5 < \theta \leq \bar{\theta}_3$; $0 < \alpha < 1$. $\pi_{m.1}(w_1^{f*}, p_{e,1}^{f*}) > \pi_{m.2}(\bar{w}_{2,1}^{f*}, \bar{p}_{e,2,1}^{f*})$; the equilibrium solution is

$$(w^{f**}, p_e^{f**}) = (w_1^{f*}, p_{e,1}^{f*}) \quad (E11)$$

Analysis 5. If $\{\bar{\theta}_3 < \theta < 1\} \cap \{0 < \alpha < 1\}$, then compare $(\bar{w}_{1,4}^{f*}, \bar{p}_{e,1,4}^{f*})$ with $(\bar{w}_{2,1}^{f*}, \bar{p}_{e,2,1}^{f*})$.

Solution of Analysis 5. $\pi_{m.1}(\bar{w}_{1,4}^{f*}, \bar{p}_{e,1,4}^{f*}) - \pi_{m.2}(\bar{w}_{2,1}^{f*}, \bar{p}_{e,2,1}^{f*}) > 0$; the equilibrium solution is

$$(w^{f**}, p_e^{f**}) = (\bar{w}_{1,4}^{f*}, \bar{p}_{e,1,4}^{f*}) \quad (E12)$$

We have summarized (E1)-(E12), and the complete equilibrium solutions are shown in Table 4.

Data Availability

The numerical experiment in this paper is mainly based on the inferred conclusion in this paper, and the corresponding parameter assignment is explained at the beginning of Section 5. Parameter setting mainly refers to the parameter setting of other similar studies, and references have been cited in this paper.

Disclosure

We declare that the work presented in this manuscript represents our original research that has not been published previously and is not under consideration for publication elsewhere, in whole or in part.

Conflicts of Interest

No conflicts of interest exist in the submission of this manuscript, which is approved by all authors for publication.

References

- [1] K. Takahashi, T. Aoi, D. Hirotsu, and K. Morikawa, "Inventory control in a two-echelon dual-channel supply chain with setup of production and delivery," *International Journal of Production Economics*, vol. 133, no. 2, pp. 403–415, 2011.
- [2] "Accenture-Consumers-In-The-New-Full-Report-Chinese," 2018, <https://www.accenture.com/cn-zh/insight-consumers-in-the-new>.
- [3] <http://www.camia.cn/content/775.html>.
- [4] M. Rabin, "Incorporating fairness into game theory and economics," *The American Economic Review*, vol. 83, no. 5, pp. 1281–1302, 1993.
- [5] E. Fehr and K. M. Schmidt, "A theory of fairness, competition, and co-operation," *The Quarterly Journal of Economics*, vol. 114, no. 3, pp. 817–868, 1999.
- [6] D. Kahneman, J. L. Knetsch, and R. H. Thaler, "Fairness and the assumptions of economics," *Journal of Business*, vol. 59, no. 4, pp. S285–S300, 1986.
- [7] N. Kumar and L. K. Scheer, "The effects of supplier fairness on vulnerable resellers," *Journal of Marketing Research*, vol. 32, no. 1, pp. 54–65, 1995.
- [8] D. Zhao, "Game analysis on vertical monopoly based on fixed resale minimum price—Take the case of Moutai and Wuliangy," *Times Finance*, no. 3, pp. 88–89, 2016.
- [9] C. Xing, "Direct stores: game chips between Moutai and dealers," *Wine World*, no. 4, pp. 26–27, 2012.
- [10] X. Zhong, "The delicate strategic partnership in supply chain management - inspiration of Gome and Gree marketing war," *China Collective Economy*, no. 05, p. 50, 2007.
- [11] W. K. Chiang, D. Chhajed, and J. D. Hess, "Direct marketing, indirect profits: a strategic analysis of dual channel design," *Management Science*, vol. 49, no. 1, pp. 1–20, 2003.
- [12] G. Ya-jun and Z. Li-qiang, "The conflict and coordination in dual channel based on e-market," *System Engineering Theory and Practice*, vol. 28, no. 9, pp. 59–66, 2008.
- [13] K. Cattani, W. Gilland, H. S. Heese, and J. Swaminathan, "Boiling frogs: pricing strategies for a manufacturer adding a direct channel that competes with the traditional channel," *Production Engineering Research and Development*, vol. 15, no. 1, pp. 40–56, 2006.
- [14] Q. Fang, L. Ren, and Y. Wang, "Pricing strategy of retailer dual-channel supply chain considering predominant power," *Journal of Wuhan University of Science and Technology*, vol. 40, no. 4, pp. 302–306, 2017.
- [15] R. Yan and Z. Pei, "Retail services and firm profit in a dual-channel market," *Journal of Retailing and Consumer Services*, vol. 16, no. 4, pp. 306–314, 2009.
- [16] J. F. Tian, T. J. Fan, and J. L. Hu, "Pricing policies in a dual-channel supply chain with manufacture services," in *Proceedings of the International Conference on Computer Information Systems & Industrial Applications*, 2015.
- [17] Q. Xu, Z. Liu, and B. Shen, "The impact of price comparison service on pricing strategy in a dual-channel supply chain," *Mathematical Problems in Engineering*, vol. 2013, Article ID 613528, 13 pages, 2013.
- [18] C. Shen, Z. Xiong, and W. Yan, "Research on dual channel pricing and coordination strategy under the network price comparison," *Chinese Journal of Management Science*, vol. 22, no. 1, pp. 84–93, 2014.
- [19] B. Li, M. Zhu, Y. Jiang, and Z. Li, "Pricing policies of a competitive dual-channel green supply chain," *Journal of Cleaner Production*, vol. 112, Part 3, pp. 2029–2042, 2016.
- [20] F. Zhang and J. Ma, "Research on the complex features about a dual-channel supply chain with a fair caring retailer," *Communications in Nonlinear Science and Numerical Simulation*, vol. 30, no. 1-3, pp. 151–167, 2016.
- [21] F. Zhang and C. Wang, "Dynamic pricing strategy and coordination in a dual-channel supply chain considering service value," *Applied Mathematical Modelling: Simulation and Computation for Engineering and Environmental Systems*, vol. 54, pp. 722–742, 2018.
- [22] B. Dan, G. Y. Xu, and C. Liu, "Pricing policies in a dual-channel supply chain with retail services," *International Journal of Production Economics*, vol. 139, no. 1, pp. 312–320, 2012.
- [23] Z. Ding, *Research on The Pricing Strategy and Coordination Contracts of Dual Channel Supply Chain*, Hefei University of Technology, 2015.
- [24] J. Zhao, X. Hou, Y. Guo, and J. Wei, "Pricing policies for complementary products in a dual-channel supply chain," *Applied Mathematical Modelling: Simulation and Computation for Engineering and Environmental Systems*, vol. 49, pp. 437–451, 2017.
- [25] J. Lin and J. Wang, "Research of manufacturers' channel strategy under dual-channel supply chain based on differentiated product," *Chinese Journal of Management Science*, vol. 26, no. 6, pp. 72–84, 2018.
- [26] Y. Qu, Z. Guan, R. Qu, and T. Ye, "Impact of members' fairness preference and loss-averse on order strategy in hybrid dual channel supply chain," *Chinese Journal of Management*, vol. 14, no. 01, pp. 129–138, 2017.
- [27] X. Wei, Q. Lin, and Y. Qin, "Optimal pricing strategies of dual-channel supply chain under risk aversion and fairness preference," *Journal of Chongqing University of Technology (Social Science)*, vol. 30, no. 12, pp. 50–58, 2016.
- [28] T. Nie and S. Du, "Dual-fairness supply chain with quantity discount contracts," *European Journal of Operational Research*, vol. 258, no. 2, pp. 491–500, 2017.
- [29] Q.-H. Li and B. Li, "Dual-channel supply chain equilibrium problems regarding retail services and fairness concerns," *Applied Mathematical Modelling*, 2016.

- [30] F. Xu, R. Chuge, and W. Fan, "Impact of horizontal fairness and vertical fairness on strategies in dual-channel supply chain," *Journal of Engineering System*, vol. 29, no. 04, pp. 527–536, 2014.
- [31] L. Wang, K. Cheng, and W. Shiwei, "Study on pricing strategies of dual-channel supply chain under fairness preference," *Chinese Journal of Management Science*, vol. 20, no. S2, pp. 563–568, 2012.
- [32] B. Li, Y. Li, L. Hou, and P. Hou, "Impact of fair-minded retailer on decision of supply chain in dual-channel," *Control and Decision*, vol. 30, no. 05, pp. 955–960, 2015.
- [33] X. Yue and J. Liu, "Demand forecast sharing in a dual-channel supply chain," *European Journal of Operational Research*, vol. 174, no. 1, pp. 646–667, 2006.
- [34] S. Huang, C. Yang, and H. Liu, "Pricing and production decisions in a dual-channel supply chain when production costs are disrupted," *Economic Modelling*, vol. 30, no. 1, pp. 521–538, 2013.
- [35] G. Y. Xu, B. Dan, X. M. Zhang, and C. Liu, "Coordinating a dual-channel supply chain with risk-averse under a two-way revenue sharing contract," *International Journal of Production Economics*, vol. 147, no. 1, pp. 171–179, 2014.
- [36] W. Wang, L. Bo, N. Liao, and L. Xu, "Redefining customer experience in the new retail era- McKinsey China digital consumer research," *Science and Technology of China*, vol. 2017, no. 09, pp. 24–28, 2017.
- [37] Niuli, *Trends in Cosmetics Sales: Online and Offline Occupy Equal Shares*, 2017, <http://www.China.china.Chinairn.com/hyzz/20170109/140751848.shtml>.
- [38] M. Liu, E. Cao, C. K. Salifou et al., "Pricing strategies of a dual-channel supply chain with risk aversion," *Transportation Research Part E: Logistics and Transportation Review*, vol. 90, pp. 108–120, 2016.
- [39] C. Yuan, L. Yan, and G. Chai, "A dual-channel cournot game model with remarks on the policy of equal prices on the two channels of suning," *Forecasting*, vol. 33, no. 05, pp. 65–70, 2014.
- [40] G. Charness and M. Rabin, "Understanding social preferences with simple tests," *The Quarterly Journal of Economics*, vol. 117, no. 3, pp. 817–869, 2002.
- [41] G. F. Loewenstein, L. Thompson, and M. H. Bazerman, "Social utility and decision making in interpersonal contexts," *Journal of Personality and Social Psychology*, vol. 57, no. 3, pp. 426–441, 1989.
- [42] T.-H. Ho and X. Su, "Peer-induced fairness in games," *American Economic Review*, vol. 99, no. 5, pp. 2022–2049, 2009.
- [43] V. Pavlov and E. Katok, *Fairness and Coordination Failures in Supply Chain Contracts. Working Paper*, University of Texas at Dallas, 2009, http://www.utdallas.edu/~ekatok/fair_theory.pdf.
- [44] X. Kong, "KUKA: Online and offline integration," *China Chain Store*, no. 10, pp. 52–53, 2015.
- [45] J. Guo, "OPPO and Vivo: Mobile phone marketing strategy analysis," *Modern SOE Research*, no. 02, p. 96, 2018.
- [46] http://www.sohu.com/a/128912758_121344.

Research Article

An Improved Demand Forecasting Model Using Deep Learning Approach and Proposed Decision Integration Strategy for Supply Chain

Zeynep Hilal Kilimci ¹, A. Okay Akyuz ^{1,2}, Mitat Uysal,¹ Selim Akyokus ³,
M. Ozan Uysal,¹ Berna Atak Bulbul ² and Mehmet Ali Ekmis²

¹Department of Computer Engineering, Dogus University, Istanbul, Turkey

²OBASE Research & Development Center, Istanbul, Turkey

³Department of Computer Engineering, Istanbul Medipol University, Istanbul, Turkey

Correspondence should be addressed to Berna Atak Bulbul; berna.bulbul@obase.com

Received 1 February 2019; Accepted 5 March 2019; Published 26 March 2019

Guest Editor: Thiago C. Silva

Copyright © 2019 Zeynep Hilal Kilimci et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Demand forecasting is one of the main issues of supply chains. It aimed to optimize stocks, reduce costs, and increase sales, profit, and customer loyalty. For this purpose, historical data can be analyzed to improve demand forecasting by using various methods like machine learning techniques, time series analysis, and deep learning models. In this work, an intelligent demand forecasting system is developed. This improved model is based on the analysis and interpretation of the historical data by using different forecasting methods which include time series analysis techniques, support vector regression algorithm, and deep learning models. To the best of our knowledge, this is the first study to blend the deep learning methodology, support vector regression algorithm, and different time series analysis models by a novel decision integration strategy for demand forecasting approach. The other novelty of this work is the adaptation of boosting ensemble strategy to demand forecasting system by implementing a novel decision integration model. The developed system is applied and tested on real life data obtained from SOK Market in Turkey which operates as a fast-growing company with 6700 stores, 1500 products, and 23 distribution centers. A wide range of comparative and extensive experiments demonstrate that the proposed demand forecasting system exhibits noteworthy results compared to the state-of-art studies. Unlike the state-of-art studies, inclusion of support vector regression, deep learning model, and a novel integration strategy to the proposed forecasting system ensures significant accuracy improvement.

1. Introduction

Since competition is increasing day by day among retailers at the market, companies are focusing more predictive analytics techniques in order to decrease their costs and increase their productivity and profit. Excessive stocks (overstock) and out-of-stock (stockouts) are very serious problems for retailers. Excessive stock levels can cause revenue loss because of company capital bound to stock surplus. Excess inventory can also lead to increased storage, labor, and insurance costs, and quality reduction and degradation depending on the type of the product. Out-of-stock products can result in

lost for sales and reduced customer satisfaction and store loyalty. If customers cannot find products at the shelves that they are looking for, they might shift to another competitor or buy substitute items. Especially at middle and low level segments, customer's loyalty is quite difficult for retailers [1]. Sales and customer loss is a critical problem for retailers. Considering competition and financial constraints in retail industry, it is very crucial to have an accurate demand forecasting and inventory control system for management of effective operations. Supply chain operations are cost oriented and retailers need to optimize their stocks to carry less financial risks. It seems that retail industry will face more

competition in future. Therefore, the usage of technological tools and predictive methods is becoming more popular and necessary for retailers [2]. Retailers in different sectors are looking for automated demand forecasting and replenishment solutions that use big data and predictive analytics technologies [3]. There has been extensive set of methods and research performed in the area of demand forecasting. Traditional forecasting methods are based on time-series forecasting approaches. These forecasting approaches predict future demand based on historical time series data which is a sequence of data points measured at successive intervals in time. These methods use a limited number of historical time-series data related with demand. In the last two decades, data mining and machine learning models have drawn more attention and have been successfully applied to time series forecasting. Machine learning forecasting methods can use a large amount of data and features related with demand and predict future demand and patterns using different learning algorithms. Among many machine learning methods, deep learning (DL) methods have become very popular and have been recently applied to many fields such as image and speech recognition, natural language processing, and machine translation. DL methods have produced better predictions and results, when compared with traditional machine learning algorithms in many researches. Ensemble learning (EL) is also another methodology to boost the system performance. An ensemble method is composed of two parts: ensemble generation and ensemble integration [4]. In ensemble generation part, a diverse set of base prediction models is generated by using different methods or samples. In integration part, the predictions of all models are combined by using an integration strategy.

In this study, we propose a novel model to improve demand forecasting process which is one of the main issues of supply chains. For this purpose, nine different time series methods, support vector regression algorithm (SVR), and DL approach based demand forecasting model are constructed. To get the final decision of these models for the proposed system, nine different time series methods, SVR algorithm, and DL model are blended by a new integration strategy which is reminiscent of boosting ensemble strategy. The other novelty of this work is the adaptation of boosting strategy to demand forecasting model. In this way, the final decision of the proposed system is based on the best algorithms of the week by gaining more weight. This makes our forecasts more reliable with respect to the trend changes and seasonality behaviors. The proposed system is implemented and tested on SOK Market real-life data. Experiment results indicate that the proposed system presents noticeable accuracy enhancements for demand forecasting process when compared to single prediction models. In Table 3, the enhancements obtained by the novel integration method compared to single best predictor model are presented. To the best of our knowledge, this is the first study to consolidate the deep learning methodology, different time series analysis models, and a novel integration strategy for demand forecasting process. The rest of this paper is organized as follows: Section 2 gives a summary of related work about demand forecasting, time series methods, DL approach, and ensemble learning

methodologies. Section 3 describes the proposed framework. Sections 4, 5, and 6 present experiment setup, experimental results, and conclusions, respectively.

2. Related Work

This section gives a summary of some of researches about demand forecasting, EL, and DL. There are many application areas of automatic demand forecasting methodologies at the literature. Energy load demand, transportation, tourism, stock market, and retail forecasting are some of important application areas of automatic demand forecasting. Traditional approaches for demand forecasting use time series methods. Time series methods include Naïve method, average method, exponential smoothing, Holt's linear trend method, exponential trend method, damped trend methods, Holt-Winters seasonal method, moving averages, ARMA (Autoregressive Moving Average), and ARIMA (Autoregressive Integrated Moving Average) models [5]. Exponential smoothing methods can have different forms depending on the usage of trend and seasonal components and additive, multiplicative, and damped calculations. Pegels presented different possible exponential smoothing methods in graphical form [6]. The types of exponential smoothing methods are further extended by Gardder [7] to include additive and multiplicative damped trend methods. ARMA and ARIMA (also called the Box-Jenkins method named after the statisticians George Box and Gwilym Jenkins) are most common methods that are applied to find the best fit of a model to historical values of a time series [8].

Intermittent Demand Forecasting methods try to detect intermittent demand patterns that are characterized with zero or varied demands at different periods. Intermittent demand patterns occur in areas like fashion retail, automotive spare parts, and manufacturing. Modelling of intermittent demand is a challenging task because of different variations. One of the influential methods about intermittent demand forecasting is proposed by Croston [9]. Croston's method uses a decomposition approach that uses separate exponentially smoothed estimates of the demand size and the interval between demand incidences. Its superior performance over the single exponential smoothing (SES) method has been demonstrated by Willemain [10]. To address some limitations on Croston's method, some additional studies were performed by Syntetos-Boylan [11, 12] and Teunter, Syntetos, and Babai [13]. Some applications use multiple time series that can be organized hierarchically and can be combined using bottom up and top down approaches at different levels in groups based on product types, geography, or other features. A hierarchical forecasting framework is proposed by Hydman et al. [14] that provides better forecasts produced by either a top-down or a bottom-up approach.

On supply chain context, since there are a high number of time series methods, automatic model selection becomes very crucial [15]. Aggregate selection is a single source of forecasts and is chosen for all the time series. Also all combinations of trend and cyclical effects in additive and multiplicative form should be considered. Petropoulos et al. in 2014 analyzed via regression analysis the main determinants of forecasting

accuracy. Li et al. introduced a revised mean absolute scaled error (RMASE) in their study as a new accuracy measure for intermittent demand forecasting which is a relative error measure and scale-independent [16]. In addition to time series methods, artificial intelligence approaches are becoming popular with the growth of big data technologies. An initial attempt was made in the study of Garcia [17].

In recent years, EL is also popular and used by researchers in many research areas. In the study of Song and Dai [18], they proposed a novel double deep Extreme Learning Machine (ELM) ensemble system focusing on the problem of time series forecasting. In the study by Araque et al., DL based sentiment classifier is developed [19]. This classifier serves as a baseline when compared to subsequent results proposing two ensemble techniques, which aggregate baseline classifier with other surface classifiers in sentiment analysis. Tong et al. proposed a new software defect predicting approach including two phases: the deep learning phase and two-stage ensemble (TSE) phase [20]. Quia presented an ensemble method [21] composed of empirical mode decomposition (EMD) algorithm and DL approach both together in his work. He focused on electric load demand forecasting problem comparing different algorithms with their ensemble strategy. Qi et al. present the combination of Ex-Adaboost learning strategy and the DL research based on support vector machine (SVM) and then propose a new Deep Support Vector Machine (DeepSVM) [22].

In classification problems, the performance of learning algorithms mostly depends on the nature of data representation [23]. DL was firstly proposed in 2006 by the study of Geoff Hinton who reported a significant invention in the feature extraction [24]. After then, DL researchers generated many new application areas in different fields [25]. Deep Belief Networks (DBN) based on Restricted Boltzmann Machines (RBMs) is another representative algorithm of DL [24] where there are connections between the layers but none of them among units within each layer [24]. At first, a DBN is implemented to train data in an unsupervised learning way. DBN learns the common features of inputs as much as possible it can. Then, DBN can be optimized in a supervised way. With DBN, the corresponding model can be constructed for classification or other pattern recognition tasks. Convolutional Neural Networks (CNN) is another instance of DL [22] and multiple layers neural networks. In CNN, each layer contains several two-dimensional planes, and those are composed of many neurons. The main principal advantage of CNN is that the weight in each convolutional layer is shared among each of them. In other words, the neurons use the same filter in each two-dimensional plane. As a result, feature tunable parameters reduce computation complexity [22, 25]. Auto Encoder (AE) is also being thought to train on deep architectures in an acquisitive layer-wise manner [22]. In neural network (NN) systems, it is supposed that the output itself can be thought as the input data. It is possible to obtain the different data representations for the original data by adjusting the weight of each layer. Input data is composed of encoder and decoder. AE is a NN for reconstructing the input data. Other types of DL methods can be found in [24, 26, 27].

Our work differs from the above mentioned literature studies in that this is the very first attempt of employing different time series methods, SVR algorithm, and DL approach for demand forecasting process. Unlike the literature studies, a novel final decision integration strategy is proposed to boost the performance of demand forecasting system. The details of the proposed study can be found in Section 3.

3. Proposed Framework

This section gives a summary of base forecasting techniques such as time series and regression methods, support vector regression model, feature reduction approach, deep learning methodology, and a new final decision integration strategy.

3.1. Time Series and Regression Methods. In our proposed system, nine different time series algorithms including moving average (MA), exponential smoothing, Holt-Winters, ARIMA [28] methods, and three different Regression Models are employed. In time series forecasting models, the classical approach is to collect historical data, analyze these data underlying feature, and utilize the model to predict the future [28]. Table 1 shows algorithm definitions and parameters which are used in the proposed system. These algorithms are commonly used forecasting algorithms in time series demand forecasting domain.

3.2. Support Vector Regression. Support Vector Machines (SVM) is a powerful classification technique based on a supervised learning theory developed and presented by Vladimir Vapnik [29]. The background works for SVM depend on early studies of Vapnik's and Alexei Chervonenkis's on statistical learning theory, about 1960s. Although the training time of even the fastest SVM can be quite slow, their main properties are highly accurate and their ability to model complex and nonlinear decision boundaries is really powerful. They show much less proneness to overfitting than other methods. The support vectors can also provide a very compact description of the learned model.

We also use SVR algorithm in our proposed system, which is regression implementation of SVM for continuous variable classification problems. SVR algorithm is being used for continuous variable prediction problems as a regression method that preserves all the main properties (maximal margin) as well as classification problems. The main idea of SVR is the computation of a linear regression function in a high dimensional feature space. The input data are mapped by means of a nonlinear function in high dimensional space. SVR has been applied in different variety of areas, especially on time series and financial prediction problems; handwritten digit recognition, speaker identification, object recognition, convex quadratic programming, and choices of loss functions are some of them [4]. SVR is a continuous variable prediction method like regression. In this study, SVR is used to forecast sales demands by using the input variables explained in Table 1.

TABLE I: Time series algorithms used in demand forecasting.

Models	Formulation	Definition of Variables
Regression Model 1	$y_t = \beta_0 + \beta_1 X_{1,t} + \beta_2 X_{2,t} + \beta_3 X_{3,t} + \beta_4 X_{4,t} + \epsilon_t$	Variability in weekly customer demand; special days and holidays (X1), discount dates (X2), days when the product is closed for sale (X3), and sales (X4) that cannot explain these three activities but are exceptional.
Regression Model 2	$y_t = \beta_0 + \beta_1 X_{1,t} + \beta_2 X_{2,t} + \beta_3 X_{3,t} + \sum_{j=1}^{52} \gamma_j W_{j,t} + \epsilon_t$	R1 is modeled by adding weekly partial-regressive terms ($W_{j,t}; j = 1, \dots, 52$).
Regression Model 3	$y_t = \beta_0 + \beta_1 X_{1,t} + \beta_2 X_{2,t} + \beta_3 X_{3,t} + \sum_{j=1}^{12} \alpha_j M_{j,t} + \sum_{j=1}^4 \gamma_j \bar{W}_{j,t} + \epsilon_t$	The R3 model adds the shadow regression terms to the R1 model for each month of the year $M_{j,t}, j = 1, \dots, 12$, and for the months of the week $\bar{W}_{j,t}, j = 1, \dots, 4$ end result.
SVR (Support Vector Regression)	$y_t = \text{SVR}(\text{input variables})$	The SVR prediction is used for forecasting y_t by using inputs X_1, X_2, X_3 mentioned regression model 1, each month of the year $M_{j,t}, j = 1, \dots, 12$, and the months of the week $\bar{W}_{j,t}, j = 1, \dots, 4$.
Exponential Smoothing Model	$y_t = \alpha \hat{y}_{t-1} + (1 - \alpha) y_{t-1} + \epsilon_t$	Exponential smoothing is generally used for products where there is no trend or seasonality in the model.
Holt-Trend Methods	$\begin{aligned} \hat{y}_t &= L_t + T_t \\ L_t &= \alpha y_t + (1 - \alpha)(L_t - 1) \\ T_t &= \beta(L_t - L_{t-1}) + (1 - \beta)T_{t-1} \end{aligned}$	Additive Holt-Trend model has been evaluated for products with level (L) and trend (T).
Holt-Winters Seasonal Models	$\begin{aligned} \hat{y}_t &= L_t + T_t + S_{t+p+1} \\ L_t &= \alpha y_t + (1 - \alpha)(L_t + T_t) \\ T_t &= \beta(L_t - L_{t-1}) + (1 - \beta)T_{t-1} \\ S_{t+p+1} &= \gamma \left(\frac{D_{t+1}}{L_{t+1}} \right) + (1 - \gamma)S_{t+1} \end{aligned}$	Additive Holt-Winters Seasonal models have been evaluated for products with level (L), seasonality (S) and trends (T)
Two-Level Model 1	$\begin{aligned} y1_t &= \alpha \hat{y}_{t-1} + (1 - \alpha) y_{t-1} + \epsilon_t \\ y2_t &= \beta_0 + \beta_1 X_{1,t} + \beta_2 X_{2,t} + \beta_3 X_{3,t} + \beta_4 X_{4,t} + \epsilon_t \\ y_t &= y1_t - y2_t + \epsilon_t \end{aligned}$	In this method, R1 model was first applied to the time series. Later, the demand values estimated by this model are subtracted from the customer demand data and the residual values are fitted with exponential correction and Holt-Winters Exponential Smoothing Model. Estimates of these two metrics were collected and the product demand estimate was calculated and evaluated.
Two-Level Model 2	$\begin{aligned} y1_t &= L_t + T_t \\ L_t &= \alpha y1_t + (1 - \alpha)(L_t - 1) \\ T_t &= \beta(L_t - L_{t-1}) + (1 - \beta)T_{t-1} \\ y2_t &= \beta_0 + \beta_1 X_{1,t} + \beta_2 X_{2,t} + \beta_3 X_{3,t} + \beta_4 X_{4,t} + \epsilon_t \\ y_t &= y1_t - y2_t + \epsilon_t \end{aligned}$	In this method, R1 model was first applied to the time series. Later, the demand values estimated by this model are subtracted from the customer demand data and the residual values are fitted with exponential correction and Holt-Winters Trend Model. Estimates of these two metrics were collected and the product demand estimate was calculated and evaluated.
Two-Level Model 3	$\begin{aligned} y1_t &= L_t + T_t + S_{t+p+1} \\ L_t &= \alpha y_t + (1 - \alpha)(L_t + T_t) \\ T_t &= \beta(L_t - L_{t-1}) + (1 - \beta)T_{t-1} \\ S_{t+p+1} &= \gamma \left(\frac{D_{t+1}}{L_{t+1}} \right) + (1 - \gamma)S_{t+1} \\ y2_t &= \beta_0 + \beta_1 X_{1,t} + \beta_2 X_{2,t} + \beta_3 X_{3,t} + \beta_4 X_{4,t} + \epsilon_t \\ y_t &= y1_t - y2_t + \epsilon_t \end{aligned}$	In this method, R1 model was first applied to the time series. Later, the demand values estimated by this model are subtracted from the customer demand data and the residual values are fitted with exponential correction and Holt-Winters Seasonal Model. Estimates of these two metrics were collected and the product demand estimate was calculated and evaluated.

3.3. *Deep Learning.* Machine learning approach can analyze features, relationships, and complex interactions among features of a problem from samples of a dataset and learn a model, which can be used for demand forecasting. Deep learning (DL) is a machine learning technique that applies deep neural network architectures to solve various complex problems. DL has become a very popular research topic among researchers and has been shown to provide impressive results in image processing, computer vision, natural language processing, bioinformatics, and many other fields [25, 26].

In principle, DL is an implementation of artificial neural networks, which mimic natural human brain [30]. However, a deep neural network is more powerful and capable of analyzing and composing more complex features and relationships than a traditional neural network. DL requires high computing power and large amounts of data for training. The recent improvements in GPU (Graphical Processing Unit) and parallel architectures enabled the necessary computing power required in deep neural networks. DL uses successive layers of neurons, where each layer extracts more complex and abstracts features from the output of previous layers.

Thus, a DL can automatically perform feature extraction in itself without any preprocessing step. Visual object recognition, speech recognition, and genomics are some of the fields where DL is applied successfully [26].

A multilayer feedforward artificial neural network (MLFANN) is employed as a deep learning algorithm in this study. In a feedforward neural network, information moves from the input nodes to the hidden nodes and at the end to the output nodes in successive layers of the network without any feedback. MLFANN is trained with stochastic gradient descent using backpropagation. It uses gradient descent algorithm to update the weights purposing to minimize the squared error among the network output values and the target output values. Using gradient descent, each weight is adjusted according to its contribution value to the error. For deep learning part, H2O library [31] is used as an open source big data artificial intelligence platform.

3.4. Feature Extraction. In this work, there are some issues because of the huge amount of data. When we try to use all of the features of all products in each store with 155 features and 875 million records, deep learning algorithm took days to finish its job due to limited computational power. For the purpose, reducing the number of features and the computational time is required for each algorithm. In order to overcome this problem and to accelerate clustering and modelling phases, PCA (Principal Component Analysis) is used as a feature extraction algorithm. PCA is a commonly used dimension reduction algorithm that presents the most significant features. PCA transforms each instance of the given dataset from d dimensional space to a k dimensional subspace in which the new generated set of k dimensions are called the Principal Components (PC) [32]. Each principal component is directed to a maximum variance excluding the variance, accounted for in all its preceding components. As a result, the first component covers the maximum variance and so on. In brief, Principal Components are represented as

$$PC_1 = a_1X_1 + a_2X_2+\dots \quad (1)$$

where PC_i is i^{th} principal component, X_j is j^{th} original feature, and a_j is the numerical coefficient for feature X_j .

3.5. Proposed Integration Strategy. Decision integration approach is being used by the way of combining the strengths of different algorithms into a single collaborated method philosophy. By the help of this approach, it is aimed to improve the success of the proposed forecasting system by combining different algorithms. Since each algorithm can be more sensitive or can have some weaknesses under different conditions, collecting decision of each model provides more effective and powerful results for decision making processes.

There are two main EL approaches in the literature, called homogeneous and heterogeneous EL. If different types of classifiers are used as base algorithms, then such a system is called heterogeneous ensemble, otherwise, homogenous ensemble. In this study, we concentrate on heterogeneous ensembles. An ensemble system is composed of two parts:

ensemble generation and ensemble integration [33–36]. In ensemble generation part, a diverse set of models are generated using different base classifiers. Nine different time series and regression methods, support vector regression model, and deep learning algorithm are employed as base classifiers in this study. There are many integration methods that combine decisions of base classifiers to obtain a final decision [37–40]. For integration step, a new decision integration strategy is proposed for demand forecasting model in this work. We draw inspiration from boosting ensemble model [4, 41] to construct our proposed decision integration strategy. The essential concept behind boosting methodology is that, in prediction or classification problems, final decisions can be calculated as weighted combination of the results.

In order to integrate the prediction of each forecasting algorithm, we used two integration strategies for the proposed approach. The first integration strategy selects the best performing forecasting method among others and uses that method to predict the demand of a product in a store for the next week. The second integration strategy chooses the best performing forecasting methods of current week and calculates the prediction by combining weighted predictions of winners. In our decision integration strategy, final decision is determined by regarding contribution of all algorithms like democratic systems. Our approach considers decisions of forecasting algorithms; those are performing better by considering the previous 4 weeks of this year and last year transformation of current week and previous week. While algorithms are getting better at forecasts by the time, their contribution weights are increasing accordingly, or vice versa. In first decision integration strategy, we do not consider contribution of all algorithms in fully democratic manner; we only look at contributions of the best algorithms of related week. In other words, final decision is maintained by integrating the decisions of only the best ones (selected according to their historical decisions) for each store and product. On the other hand, the best algorithms of a week can change according to each store, product couple, since every algorithm has different behavior for different products and locations.

The second decision integration strategy is based on the weighted average of mean absolute percentage error (MAPE) and mean absolute deviation (MAD) [42]. These are two popular evaluation metrics to assess the performance of forecasting models. The equations for calculation of these forecast accuracy measures are as follows:

MAPE (Mean Absolute Percentage Error)

$$MAPE = \frac{1}{n} \sum_{t=1}^n \frac{|F_t - A_t|}{A_t} \quad (2)$$

$$MAD = \frac{1}{n} \sum_{t=1}^n |F_t - A_t| \quad (3)$$

MAD (Mean Absolute Deviation)

where

- (i) F_t is the expected or estimated value for period t
- (ii) A_t is the actual value for period t
- (iii) n is also taking the value of the number of periods

Accuracy of a model can be simply computed as follows: Accuracy = 1-MAPE. If the values of the above two measures are small, it means that forecasting model performs better [43].

Replenishment for demand forecasting in retail industry needs some modifications in general EL strategies, regarding retail specific trends and dynamics. For this purpose, the proposed second integration strategy takes calculation of weighted average of MAPE for each week of previous 4 weeks and additionally MAPE of the previous year's same week and previous year following week's trend in Equation (4). This makes our forecasts more reliable with respect to trend changes and seasonality behaviors. MAPE of each week is being calculated by the formula defined in Equation (2) above. The proposed model automatically changes weekly weights of each member algorithm by the way of their average MAPE for each week at store and product level. Based on the results of MAPE weights, the best algorithms of the week gain more weight on final decision.

Average MAPE of each algorithm for a product at each store is calculated with Equation (4). Suppose that MAPEs of an algorithm are 0.2, 0.3, 0.1, and 0.1 for the previous 4 weeks of the current week and 0.1 and 0.2 for the previous year's same week and prior week. Assume that weekly coefficients are 25%, 20%, 10%, 5%, 30%, and 10%, respectively. Then, the average MAPE of an algorithm for the current week will be $0.3 * 0.25 + 0.3 * 0.2 + 0.1 * 0.1 + 0.1 * 0.05 + 0.1 * 0.3 + 0.2 * 0.1 = 0.2$ according to Equation (4).

$$M_{\text{avg}} = \sum_{k=1}^n C_k M_k \quad (4)$$

In Equation (4), the sum of coefficients should be 1; that is, $\sum_{k=1}^n C_k = 1$. M_k means MAPE of the related weeks, where

- (i) M_1 is the MAPE of previous week
- (ii) M_2 is the MAPE of 2 weeks before related week
- (iii) M_3 is the MAPE of 3 weeks before related week
- (iv) M_4 is the MAPE of 4 weeks before related week
- (v) M_5 is the MAPE of the previous year's same week
- (vi) M_6 is the MAPE of the previous year's previous week

The proposed decision integration system computes a MAPE for each algorithm for the coming forecasting week per store and product level. In addition, the effects of special days are taken into consideration. Christmas, Valentine's day, mother's day, start of Ramadan, and other religious days can be thought as some examples of special days. System users, considering the current year's calendar, can define special days manually. Thus, trends of special days are computed by using previous year's trends automatically. This enables the consideration of seasonality and special events that results in evaluation

of more accurate forecasting values. Meanwhile, seasonality and other effects can be taken under consideration, as well. Following MAPE calculation of the algorithms, new weights (W_i where $i=L..n$) are being assigned to each of them according to their weighted average for the current week.

The next step is the definition of the best algorithms of the week for each store and product couple. We only take 30% of the best performing methods into consideration as the best algorithms. After making many empirical observations, this ratio is giving ultimate results according to the dataset. It is obvious that these parameters are very specific to dataset characteristics and also depend on which algorithms are included in the integration strategy. Scaling is the last preprocess for the calculation of final decision forecast. Suppose that we have n algorithms (A_1, A_2, \dots, A_n) and k of them are the best models of the week for a store and product couple, in which $k \leq n$. The weight of each winner is being scaled according to its weighted average in Equation (5).

$$W'_t = \frac{W_t}{\sum_{j=1}^k W_j}, \quad t \text{ in } (1, \dots, k) \quad (5)$$

Assume that there are 3 best algorithms and their weights are $W_1 = 30\%$, $W_2 = 18\%$, and $W_3 = 12\%$, respectively. Their scaled new weights are going to be

$$\begin{aligned} W'_1 &= \frac{30}{60} = 50\%, \\ W'_2 &= \frac{18}{60} = 30\%, \\ W'_3 &= \frac{12}{60} = 20\%, \end{aligned} \quad (6)$$

respectively, according to Equation (5).

Scaling makes our calculation more comprehensible. After scaling the weight of each algorithm, the system gives ultimate decision according to new weights by considering the performance of each algorithm's with Equation (5).

$$F_{\text{avg}} = \sum_{j=1}^k F_j W'_j \quad (7)$$

In Equation (7), the main constraint is $\sum_{j=1}^k W'_j = 1$, k is the number of champion algorithms, and F_1 is the forecast of the related algorithm. Suppose that the best performing algorithms are A_1, A_2 , and A_3 and algorithm A_1 forecasts sales quantity as 20 and A_2 says it will be 10 for the next week; A_3 forecast is 5. Let us assume that their scaled weights are 50%, 30%, and 20%, respectively. Then the weighted forecast is as follows according to Equation (7):

$$F_{\text{avg}} = 20 * 0.50 + 10 * 0.30 + 5 * 0.20 = 14 \quad (8)$$

Every algorithm has vote right according to its weight. Finally, if a member algorithm does not appear in the list of the best algorithms for a product and store couple for a specific period, it is automatically being put into a black list, so that it will not be used anymore for a product, store level

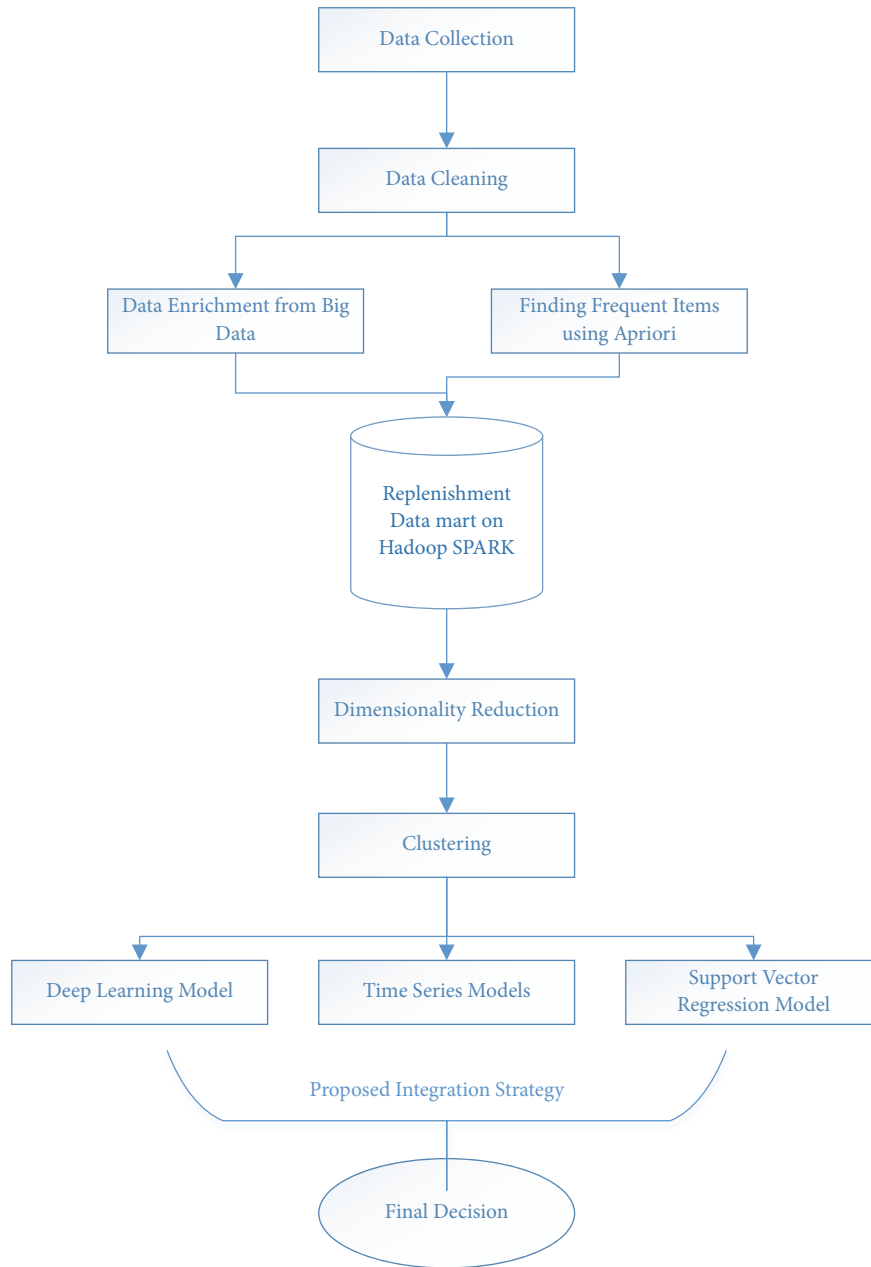


FIGURE 1: Flowchart of the proposed system.

as it is marked. This enables faster computation time in which the system disregards poor performing algorithms. The algorithm of the proposed demand forecasting system and the flowchart of the proposed system are given in Algorithm 1 and Figure 1, respectively.

4. Experiment Setup

We use an enhanced dataset which includes SOK Market's real life sales and stock data with enriched features. The dataset consists of 106 weeks of sales data that includes 7888 distinct products for different stores of SOK Market. On each store, around 1500 products are actively sold while the rest

rarely become active. The forecasting is being performed for each week to predict demand of the following week. Three-fold cross validation methodology is applied for testing data and then decides with the best accurate one of them for the next coming week forecast. This is a common approach making demand forecasting in retail [44]. Furthermore, outside weather information, such as daily temperature and other weather conditions, is joined to the model as new variables. Since it is known that there is a certain correlation between shopping trends and weather conditions at retail in most circumstances [33], we enriched the dataset with the weather information. The dataset also includes mostly sales related features filled from relational legacy data sources.

Given: n is the number of stores, m is the number of products, t is the number of algorithms in the system and A_t is an algorithm with index t . $s_{m,n}$ is the matrix which includes the number of best performing algorithms for each store, and product, $B_{m,n}$ is the matrix which contains the set of blacklist of algorithms for each store, and product, and $F_{i,j}$ is the matrix which stores final decision of each forecast where i is the number of stores and j is the number of products.

```

for i=1:n
  for j=1:m
    for k=1:t
      if  $A_k$  is in list  $B_{i,j}$  then continue
      else run  $A_k$ 
        Calculate algorithm weight  $W_k$ 
      end if
    end for
    for k=1:t
      Choice best performing algorithms and locate in  $\{s_{i,j}\}$ 
    end for
    for z=1:  $s_{i,j}$ 
      do scaling for  $A_z$ 
    end for
    Calculate proposed integration strategy, and store in  $F_{i,j}$ 
  end for
end for
return all  $F_{n,m}$ 

```

ALGORITHM 1: The algorithm of the proposed demand forecasting system.

Deep learning approach is used to estimate customer demand for each product at each store of SOK Market retail chain. There is also very apparent similarity between products, which are mostly in the same basket in terms of sales trends. For this purpose, Apriori algorithm [4] is utilized to find correlated products of each product. Apriori algorithm is a fast way to find out most associated products which are at the same basket. The most associated product's sales related features are added to dataset, as well. This takes the advantage of relationships among products with similar sales trends. Addition of associated product of a product and its features enables DL algorithm to learn better within different hidden layers. In summary, it is observed in extensive experiments that enhancements of training dataset with weather data and most associated products sales data made our estimates more powerful with the usage of DL algorithm.

The dataset includes weekly based data of each product at each store for the latest 8 weeks and sales quantities of the last 4 weeks in the same season. Historically, 2 years of data are included for each 5500 stores and 1500 products. Data size is about 875 million records within 155 features. These features include sales quantity, customer return quantity, received product quantities from distribution centers, sales amount, discount amount, receipt counts of a product, customer counts who bought a specific product, sales quantity for each day of week (Monday, Tuesday, etc.), maximum and minimum stock level of each day, average stock level for each week, sales quantity for each hour of the days, and also sales quantities of last 4 weeks of the same season. Since there is a relationship between products which are generally in the same basket, we prepared the same features defined above for

the most associated products, as well. Detailed explanation of the dataset is presented in Table 2.

In Table 2 fields given with square brackets mean that it is an array for each week (Ex. Sales_quantity_week_0 means sales quantity of current week and sales_quantity_week_1 is for sales quantity for one week before and so on). Regarding experimental results, taking weekly data for the latest 8 weeks and 4 weeks of the same season at previous year gives more acceptable trends of changes in seasonality.

For DL implementation part, H2O library [31] is used as an open source big data artificial intelligence platform. H2O is a powerful machine learning library and gives us opportunity to implement DL algorithms on Hadoop Spark big data environments. It puts together the power of highly advanced machine learning algorithms and provides to get benefit of truly scalable in-memory processing for Spark big data environment on one or many nodes via its version of Sparkling Water. By the help of in-memory processing capability of Spark technologies, it provides faster parallel platform to utilize big data for business requirements to get maximum benefit. In this study, we use H2O version 3.14.0.2 on 64-bit, 16-core, and 32GB memory machine to observe the power of parallelism during DL modelling while increasing the number of neurons.

A multilayer feedforward artificial neural network (MLFANN) is employed as a deep learning algorithm. MLFANN is trained with stochastic gradient descent using backpropagation. Using gradient descent, each weight is adjusted according to its contribution value to the error. H2O has also advanced features like momentum training, adaptive learning rate, L1 or L2 regularization, rate annealing,

TABLE 2: Dataset explanation.

Name	Description	Example
Yearweek	Related yearweek, weeks are starting from Monday to Sunday	201801
Store	Store number	1234
Product	Product Identification Number	1
Product_Adjactive	Associated product with the product according to apriori algorithm. Most frequent product at the same basket with a specific product.	2
Stock_In_Quantity_Week[0-8]	Stock increase quantity of the product in related week, ex. stock transfer quantity from distribution center to store.	50
Return_Quantity_Week[0-8]	Stock return quantity from customers at a specific week and store	20
Sales_Quantity_Week[0-8]	Weekly sales quantity of related product at a specific store	120
Sales_Amount_Week[0-8]	Total sales amount of the product at the customer receipt	2500
Discount_Amount_Week[0-8]	Discount amount of the product if there is any	500
Customer_Count_Week[0-8]	How many customers bought this product at a specific week	30
Receipt_Count_Week[0-8]	Distinct receipt count for related product	20
Sales_Quantity_Time[9-22]	Hourly sales quantity of related product from 9 am to 22 pm.	5
Last4weeks_Day[1-7]	Total sales of each weekday of last 4 weeks. Total sales of Mondays, Tuesdays... etc.	10
Last8weeks_Day[1-7]	Total sales of each weekday of last 8 weeks. Total sales of Mondays, Tuesdays... etc.	10
Max_Stock_Week[0-8]	Maximum stock quantity of related week.	12
Min_Stock_Week[0-8]	Minimum stock quantity of related week	2
Avg_Stock_Week[0-8]	Average stock quantity of related week	5
Sales_Quantity_Adj_Week[0-8]	Sales quantity of most associated product	14
Temperature_Weekday[1-7]	Daily temperature of weekdays. Monday, Tuesday... etc.	22
Weekly_Avg_Temperature[0-8]	Average weather temperature of related week.	23
Weather_Condition_Weekday[1-7]	Nominal variable; rainy, snowy, sunny, cloudy etc.	Rainy
Sales_Quantity_Next_Week	Target variable of our classification problem	25

dropout, and grid search. Gaussian distribution is applied because of the usage of continuous variable as response variable. H2O performs very well in our environment when 3 level hidden layers, 10 nodes each of them, and totally 300 epochs are set as parameters.

After implementing dimension reduction step, K-means is employed as a clustering algorithm. K-means is a greedy and fast clustering algorithm [45] that tries to partition samples into k clusters in which each sample is near to the cluster center. K is selected as 20 because, after several trials and empirical observations, the most even distribution on dataset is reached and divided our dataset into 20 different clusters for each store. After that part, deep learning algorithm is applied and obtained 20 different DL models for each store. Then, demand forecasting for each product is performed by using its cluster's model on a store basis. The main reason of using clustering is time and computing lack for DL algorithm. Instead of making modelling for each store, 20 models are generated for each store in product level. Our trials with sample dataset bias difference are less than 2%, so this speed-up is very beneficial and a good option for researchers. The forecasting result of the DL model is transferred into our forecasting system that makes its final decision by considering decisions of 11 forecasting methods including DL model.

Forecasting solutions should be scalable, process large amounts of data, and extract a model from data. Big data

environments using Spark technologies give us opportunity for implementing algorithms with scalability that shares tasks of machine learning algorithms among nodes of parallel architectures. Considering huge amounts of samples and large number of features, even the computational power of parallel architecture is not enough in some of cases. For this reason, dimension reduction is needed for big data applications. In this study, PCA is employed as a feature extraction step.

5. Experimental Results

Forecasting solutions should be scalable, process large amounts of data, and extract a model from data. Big data environments using Spark technologies give us opportunity for implementing algorithms with scalability that shares tasks of machine learning algorithms among nodes of parallel architectures. Considering huge amounts of samples and large number of features, even the computational power of parallel architecture is not enough in some of cases. For this reason, dimension reduction is needed for big data applications. In this study, PCA is employed as a feature extraction step.

SOK Market sells 21 different groups of items. We assess the performance of the proposed forecasting system on group basis as shown in Table 3. Table 3 shows MAPE (Mean Absolute Percentage Error) of integration strategies (S_1) and (S_2)

TABLE 3: Demand forecasting improvements per product group.

Product Groups	Method I without Proposed Integration Strategy		Method II with Proposed Integration Strategy		Proposed Integration Strategy with Deep Learning		Percentage Success Rate $P_1 = (S_1 - S_2)/S_2$	Percentage Success Rate $P_2 = (S_1 - S_D)/S_D$	Improvement $D = P_2 - P_1$
	(S_1) MAPE	(S_2) MAPE	(S_2) MAPE	(S_D) MAPE	(S_2) MAPE	(S_D) MAPE			
Baby Products	0.5157	0.3081	0.3081	0.2927	0.2927	0.2927	43.25%	43.25%	2.99%
Bakery Products	0.3482	0.2059	0.2059	0.1966	0.1966	0.1966	40.85%	43.52%	2.67%
Beverage	0.3714	0.2316	0.2316	0.2207	0.2207	0.2207	37.64%	40.58%	2.94%
Biscuit-Chocolate	0.3358	0.2077	0.2077	0.1977	0.1977	0.1977	38.14%	41.13%	2.99%
Breakfast Products	0.4443	0.2770	0.2770	0.2661	0.2661	0.2661	37.65%	40.11%	2.45%
Canned-Paste-Sauces	0.3836	0.2309	0.2309	0.2198	0.2198	0.2198	39.80%	42.69%	2.89%
Cheese	0.3953	0.2457	0.2457	0.2345	0.2345	0.2345	37.84%	40.68%	2.84%
Cleaning Products	0.4560	0.2791	0.2791	0.2650	0.2650	0.2650	38.79%	41.89%	3.10%
Cosmetics Products	0.5397	0.3266	0.3266	0.3148	0.3148	0.3148	39.49%	41.67%	2.19%
Deli Meats	0.4242	0.2602	0.2602	0.2488	0.2488	0.2488	38.65%	41.36%	2.70%
Edible Oils	0.4060	0.2299	0.2299	0.2215	0.2215	0.2215	43.36%	45.45%	2.09%
Household Goods	0.5713	0.3656	0.3656	0.3535	0.3535	0.3535	36.01%	38.13%	2.12%
Ice Cream-Frozen	0.5012	0.3255	0.3255	0.3106	0.3106	0.3106	35.05%	38.03%	2.98%
Legumes-Pasta-Soup	0.3850	0.2397	0.2397	0.2269	0.2269	0.2269	37.74%	41.07%	3.33%
Nuts-Chips	0.3316	0.2049	0.2049	0.1966	0.1966	0.1966	38.20%	40.71%	2.51%
Poultry Eggs	0.4219	0.2527	0.2527	0.2403	0.2403	0.2403	40.11%	43.04%	2.94%
Ready Meals	0.4613	0.2610	0.2610	0.2520	0.2520	0.2520	43.42%	45.36%	1.94%
Red Meat	0.2514	0.1616	0.1616	0.1532	0.1532	0.1532	35.71%	39.06%	3.35%
Tea-Coffee Products	0.4347	0.2650	0.2650	0.2535	0.2535	0.2535	39.04%	41.68%	2.64%
Textile Products	0.5418	0.3048	0.3048	0.2907	0.2907	0.2907	43.74%	46.34%	2.60%
Tobacco Products	0.3791	0.2378	0.2378	0.2290	0.2290	0.2290	37.29%	39.61%	2.32%
Average	0.4238	0.2582	0.2582	0.2469	0.2469	0.2469	38.99%	41.68%	2.69%

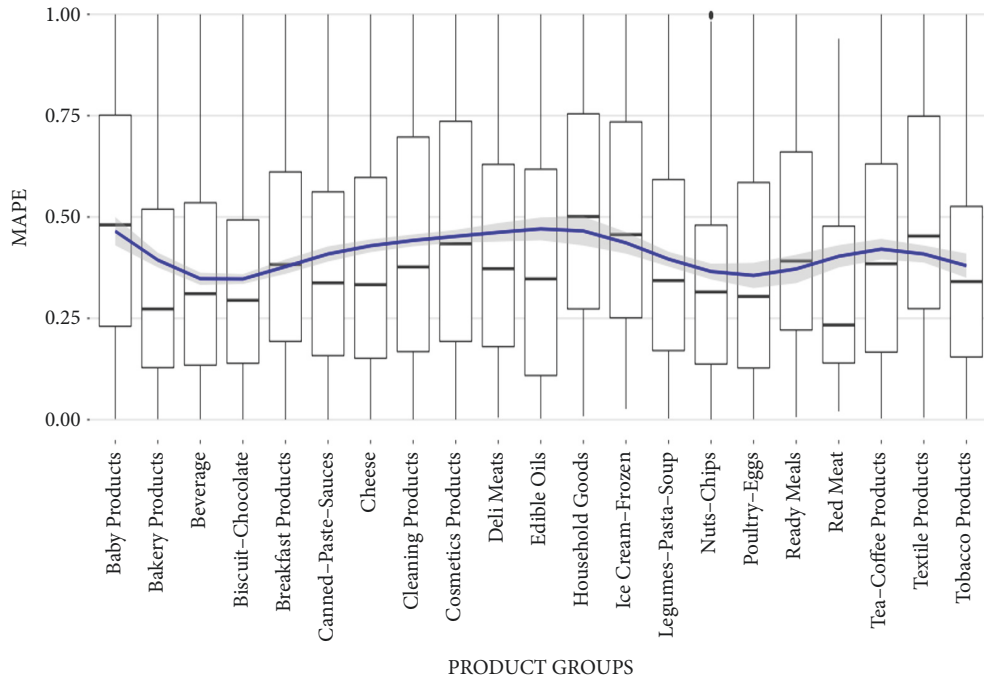


FIGURE 2: MAPE distribution according to product groups with S_1 integration strategy.

in columns 2 and 3. The average percentage forecasting errors of strategies (S_1) and (S_2) are 0.42 and 0.26, respectively. As it can be seen from Table 3, the percentage success rate of (S_2) in accordance with (S_1) is represented in column 5. The current integration strategy (S_2) provides 38.99% improvement over the first integration strategy (S_1) average. In some product groups like Textile Products, the error rate is reduced from 0.54 to 0.31 with 43.74% enhancement in MAPE.

With this work, a further improvement is obtained by utilizing DL model compared to the usage of just 10 algorithms which are based on forecasting strategy. Column 4 of Table 3 demonstrates MAPEs of 21 different groups of products after the addition of DL model. After the inclusion of DL model into the proposed system, we observed around 2% to 3.4% increase in demand forecast accuracies in some of product groups. The error rates of some product groups like “baby products,” “cosmetic products,” and “household goods” are usually higher than others, since these product groups do not have continuous demand by consumers. For example, the error rate for “household goods” group is 0.35 while the error rate for “red meat” group is 0.15. The error rates in food products are usually lower, since customers consider their daily consumption regularly to buy these products. However, the error rates are higher in baby, textile, and household goods. Customers buy products from these groups depending on whether they like the product or not selectively. In addition, the prices of goods in these groups and promotion effects are usually high relative to the prices of goods in other groups.

Moreover, it is observed that the top benefitted product groups for Method I accomplishes over 40% success rate on Baby Products, Bakery Products, Edible Oils, Poultry Eggs, Ready Meals, and Textile Products when the column

of percentage success rate is analyzed. The common point of these groups is that each one of them is being consumed by a specific customer segment, regularly. For instance, baby products group is being chosen by families who have kids; Edible Oils, Poultry Eggs, and Bakery Products are being preferred by frequently and continuously shopping customer segments; and Ready Meals are being preferred by mostly bachelor, single, and working consumers, etc. Furthermore, the inclusion of DL into the forecasting system indicates that some other consuming groups (for example, Cleaning Products, Red Meats, and Legumes Pasta Soup) exhibit better performance than the others with additional over 3%.

Figure 2 shows the box plots of mean percentage errors (MAPE) for each group of products after application of (S_1) integration strategy. The box plots enable us to analyze distributional characteristics of forecasting errors for product groups. As it can be seen from the figure, there are different medians for each product group. Thus, the forecasting errors are usually dissimilar for different products groups. The interquartile range box represents the middle 50% of scores in data. The lengths of interquartile range boxes are usually very tall. This means that there are quite number of different prediction errors within products of a given group.

Figure 3 presents the box plots of MAPEs of the proposed forecasting system that apply proposed integration approach (S_2) which combines the predictions of 10 different forecasting algorithms.

Figure 4 shows after adding DL to our system results. The lengths of interquartile range boxes are narrower when compared to the ones in Figures 2 and 3. This means that the distribution of forecasting errors is less than Figures 2 and 3. Finally, the integration of DL into the proposed forecasting system generates results that are more accurate.

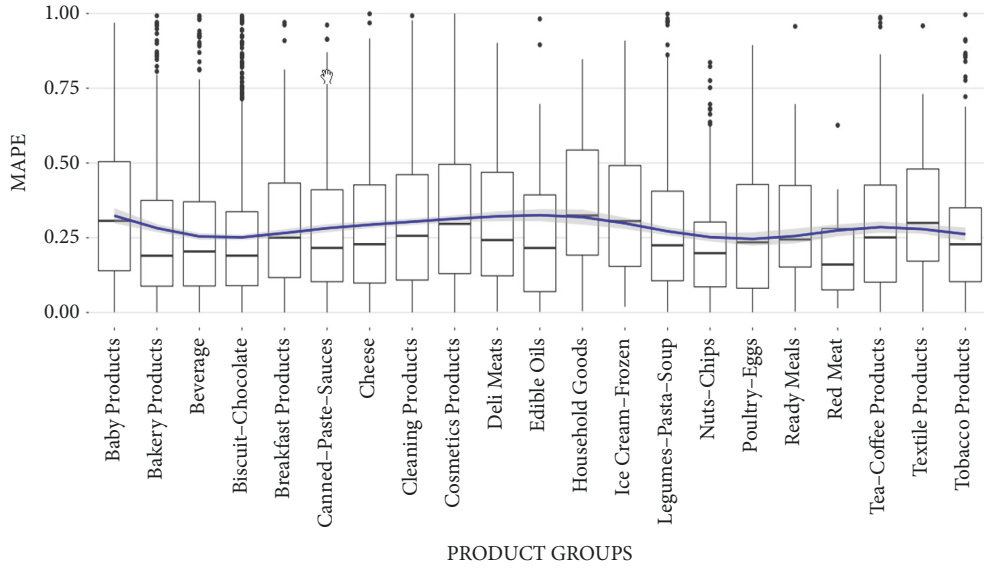


FIGURE 3: MAPE distribution according to product groups with S_2 integration strategy.

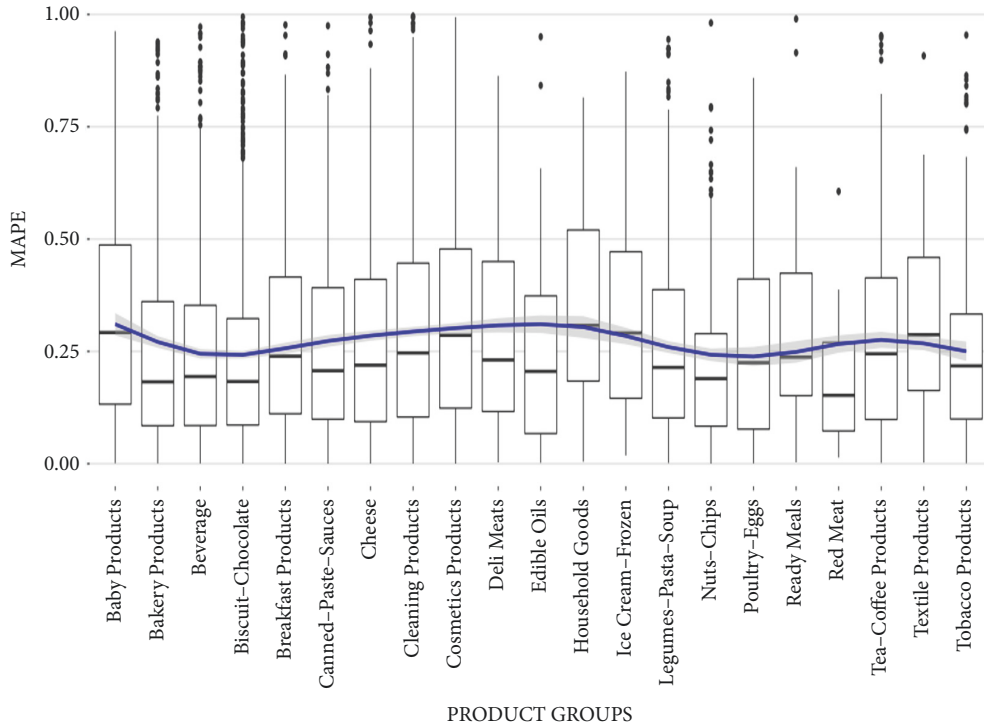


FIGURE 4: MAPE distribution according to product groups after deep learning algorithms.

For example, for the (S_1) integration strategy in Baby Products product group MAPE distribution for the 1st quartile of data 5 is observed between 0% and 23%, for 2nd quartile between 23% and 48%, and for 3rd and 4th between 48% and 75% in Figure 2. After applying integration strategy for Baby Products product group, MAPE distribution becomes 0%–16% for the 1st quartile, 16%–30% for the 2nd quartile of the data, and 30%–50% for the rest. This gives around 8% improvement for the first quartile, from 8% to 18% enhancement for the second quartile, and from 18% to 25% advancement for the rest of the data according to boundaries

differences. Median of MAPE distribution is analyzed as 48%. After applying the proposed integration strategy, it is observed as 27%, which means 21% improvement. Approximately 1%–3% enhancement is observed for each quartile of the data with the inclusion of deep learning strategy in Figure 4.

For Cleaning Products, similar improvements are observed with the others; for integration strategy MAPE distribution of the 1st quartile of data is observed between 0% and 19%, for the 2nd quartile it is between 19% and 40%, and for the rest of data it is observed between 40% and

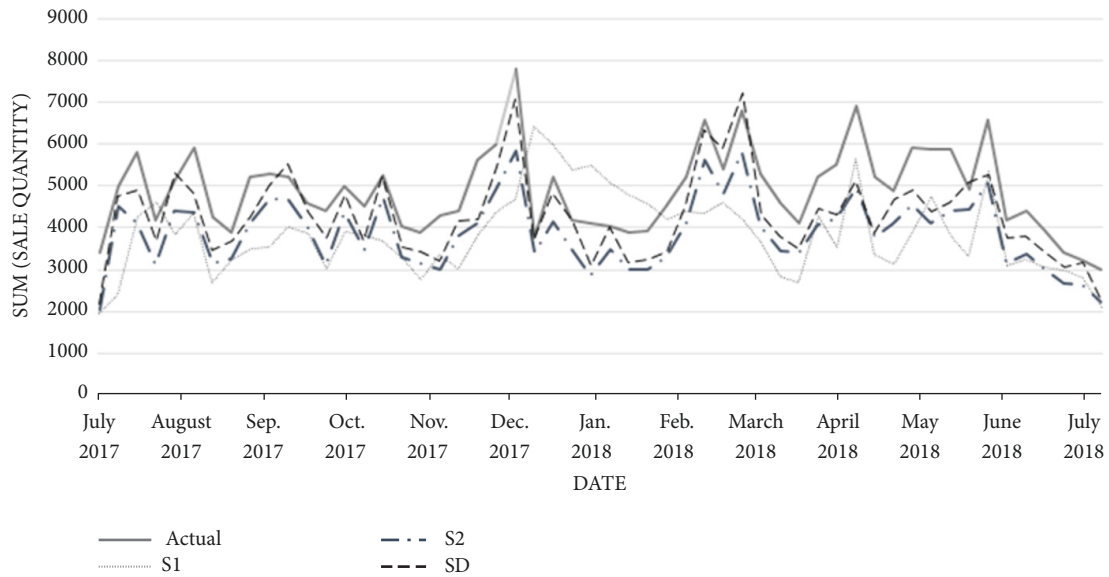


FIGURE 5: Accuracy comparison among integration strategies for one sample product group.

73% in Figure 2. After performing the proposed integration strategy for Cleaning Products, MAPE distribution becomes 0%–10% for the 1st quartile of data, 10%–17% for the 2nd quartile, and 27%–43% for the 3rd and 4th quartiles together. Improvements are observed as follows: 9% for the 1st quartile, from 9% to 13% for the 2nd quartile, and from 13% to 20% for the rest of the data at boundaries. After implementation of the integration strategy for the 1st and 2nd quartiles, the system advancement is observed as 2%, and for the 3rd and 4th quartiles it is %3, respectively. Briefly, integration strategy improvement is remarkable for each product group and consolidation of DL to the proposed integration strategy advances results between almost 2% and 3%, additionally.

In Figure 5, the accuracy comparison among integration strategies of one of the best performing sample groups is presented during 1-year period. Particularly, Christmas and other seasonality effects can be seen very clearly at Figure 5 and integration strategy with DL (S_D) is performing with the best accuracy.

It is hard to compare the performance of our proposed system with the other studies because of the lack of works with similar combinations of deep learning approach, the proposed decision integration strategies, and different learning algorithms for demand forecasting process. Another difficulty is the deficiency of real-life dataset similar to SOK dataset in order to compare the state-of-art studies. Although the results of proposed system are given in this study, we also report the results of a number of research works here on demand forecasting domain. A greedy aggregation-decomposition method has solved a real-world intermittent demand forecasting problem for a fashion retailer in Singapore [46]. They report 5.9% MAPE success rate with a small dataset. Authors compare different approaches such as statistical model, winter model, and radius basis function neural network (RBFNN) with SVM for demand forecasting process in [47]. As a result, they conclude the study by the

fact that the success of SVM algorithm surpasses others with around 7.7% enhancement at average MAPE results level. A novel demand forecasting model called SHeNSVM (Selective and Heterogeneous Ensemble of Support Vector Machines) is proposed in [48]. The proposed model presents that the individual SVMs are trained by different samples generated by bootstrap algorithm. After that, genetic algorithm is employed for retrieving the best individual combination schema. They report 10% advancement with SVM algorithm and 64% MAPE improvement. They only employ beer data with 3 different brands in their experiments. Tugba Efindigil et al. propose a novel forecasting mechanism which is modeled by artificial intelligence approaches in [49]. They compared both artificial neural networks and adaptive network-based fuzzy inference system techniques to manage the fuzzy demand with incomplete information. They reached around 18% MAPE rates for some products during their experiments.

6. Conclusion

In retail industry, demand forecasting is one of the main problems of supply chains to optimize stocks, reduce costs, and increase sales, profit, and customer loyalty. To overcome this issue, there are several methods such as time series analysis and machine learning approaches to analyze and learn complex interactions and patterns from historical data.

In this study, there is a novel attempt to integrate the 11 different forecasting models that include time series algorithms, support vector regression model, and deep learning method for demand forecasting process. Moreover, a novel decision integration strategy is developed by drawing inspiration from the ensemble methodology, namely, boosting. The proposed approach considers the performance of each model in time and combines the weighted predictions of the best performing models for demand forecasting process. The

proposed forecasting system is tested and carried out real life data obtained from SOK Market retail chain. It is observed that the inclusion of different learning algorithms except time series models and a novel integration strategy advanced the performance of demand forecasting system. To the best of our knowledge, this is the very first attempt to consolidate deep learning methodology, SVR algorithm, and different time series methods for demand forecasting systems. Furthermore, the other novelty of this work is the adaptation of boosting ensemble strategy to the demand forecasting model. In this way, the final decision of the proposed system is based on the best algorithms of the week by gaining more weight. This makes our forecasts more reliable with respect to the trend changes and seasonality behaviors. Moreover, the proposed approach performs very well integrating with deep learning algorithm on Spark big data environment. Dimension reduction process and clustering methods help to decrease time consuming with less computing power during deep learning modelling phase. Although a review of some of similar studies is presented in Section 5, as it is expected, it is very difficult to compare the results of other studies with ours because of the use of different datasets and methods. In this study, we compare results of three models where model one selects the best performing forecasting method depending on its success on previous period with 42.4% MAPE on average. The second model with the novel integration strategy results in 25.8% MAPE on average. The last model with the novel integration strategy enhanced with deep learning approach provides 24.7% MAPE on average. As a result, the inclusion of deep learning approach into the novel integration strategy reduces average prediction error for demand forecasting process in supply chain.

As a future work, we plan to enrich the set of features by gathering data from other sources like economic studies, shopping trends, social media, social events, and location based demographic data of stores. New variety of data sources contributions to deep learning can be observed. One more study can be done to determine the hyperparameters for deep learning algorithm. In addition, we also plan to use other deep learning techniques such as convolutional neural networks, recurrent neural networks, and deep neural networks as learning algorithms. Furthermore, our another objective is to use heuristic methods MBO (Migrating Birds Optimization) and other related algorithms [50] to optimize some of coefficients/weights which were determined empirically by trial-and-error like taking 30% percent of the best performing methods in our current system.

Data Availability

This dataset is private customer data, so the agreement with the organization SOK does not allow sharing data.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this article.

Acknowledgments

This work is supported by OBASE Research & Development Center.

References

- [1] P. J. McGoldrick and E. Andre, "Consumer misbehaviour: promiscuity or loyalty in grocery shopping," *Journal of Retailing and Consumer Services*, vol. 4, no. 2, pp. 73–81, 1997.
- [2] D. Grewal, A. L. Roggeveen, and J. Nordfält, "The Future of Retailing," *Journal of Retailing*, vol. 93, no. 1, pp. 1–6, 2017.
- [3] E. T. Bradlow, M. Gangwar, P. Kopalle, and S. Voleti, "the role of big data and predictive analytics in retailing," *Journal of Retailing*, vol. 93, no. 1, pp. 79–95, 2017.
- [4] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers, USA, 2013.
- [5] R. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*, OTexts, Melbourne, Australia, 2018, <http://otexts.org/fpp2/>.
- [6] C. C. Pegels, "Exponential forecasting: some new variations," *Management Science*, vol. 12, pp. 311–315, 1969.
- [7] E. S. Gardner, "Exponential smoothing: the state of the art," *Journal of Forecasting*, vol. 4, no. 1, pp. 1–28, 1985.
- [8] D. Gujarati, *Basic Econometrics*, Mcgraw-Hill, New York, NY, USA, 2003.
- [9] J. D. Croston, "Forecasting and stock control for intermittent demands," *Operational Research Quarterly*, vol. 23, no. 3, pp. 289–303, 1972.
- [10] T. R. Willemain, C. N. Smart, J. H. Shockor, and P. A. DeSautels, "Forecasting intermittent demand in manufacturing: a comparative evaluation of Croston's method," *International Journal of Forecasting*, vol. 10, no. 4, pp. 529–538, 1994.
- [11] A. Syntetos, *Forecasting of intermittent demand*, Brunel University, 2001.
- [12] A. A. Syntetos and J. E. Boylan, "The accuracy of intermittent demand estimates," *International Journal of Forecasting*, vol. 21, no. 2, pp. 303–314, 2005.
- [13] R. H. Teunter, A. A. Syntetos, and M. Z. Babai, "Intermittent demand: linking forecasting to inventory obsolescence," *European Journal of Operational Research*, vol. 214, no. 3, pp. 606–615, 2011.
- [14] R. J. Hyndman, R. A. Ahmed, G. Athanasopoulos, and H. L. Shang, "Optimal combination forecasts for hierarchical time series," *Computational Statistics & Data Analysis*, vol. 55, no. 9, pp. 2579–2589, 2011.
- [15] R. Fildes and F. Petropoulos, "Simple versus complex selection rules for forecasting many time series," *Journal of Business Research*, vol. 68, no. 8, pp. 1692–1701, 2015.
- [16] C. Li and A. Lim, "A greedy aggregation–decomposition method for intermittent demand forecasting in fashion retailing," *European Journal of Operational Research*, vol. 269, no. 3, pp. 860–869, 2018.
- [17] F. Turrado García, L. J. García Villalba, and J. Portela, "Intelligent system for time series classification using support vector machines applied to supply-chain," *Expert Systems with Applications*, vol. 39, no. 12, pp. 10590–10599, 2012.
- [18] G. Song and Q. Dai, "A novel double deep ELMs ensemble system for time series forecasting," *Knowledge-Based Systems*, vol. 134, pp. 31–49, 2017.

- [19] O. Araque, I. Corcuera-Platas, J. F. Sánchez-Rada, and C. A. Iglesias, "Enhancing deep learning sentiment analysis with ensemble techniques in social applications," *Expert Systems with Applications*, vol. 77, pp. 236–246, 2017.
- [20] H. Tong, B. Liu, and S. Wang, "Software defect prediction using stacked denoising autoencoders and twostage ensemble learning," *Information and Software Technology*, vol. 96, pp. 94–111, 2018.
- [21] X. Qiu, Y. Ren, P. N. Suganthan, and G. A. J. Amaratunga, "Empirical mode decomposition based ensemble deep learning for load demand time series forecasting," *Applied Soft Computing*, vol. 54, pp. 246–255, 2017.
- [22] Z. Qi, B. Wang, Y. Tian, and P. Zhang, "When ensemble learning meets deep learning: a new deep support vector machine for classification," *Knowledge-Based Systems*, vol. 107, pp. 54–60, 2016.
- [23] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: a review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.
- [24] G. E. Hinton, S. Osindero, and Y. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, 2006.
- [25] Y. Bengio, "Practical recommendations for gradient-based training of deep architectures," in *Neural Networks: Tricks of the Trade*, vol. 7700 of *Lecture Notes in Computer Science*, pp. 437–478, Springer, Berlin, Germany, 2nd edition, 2012.
- [26] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning Review," *International Journal of Business and Social Science*, vol. 3, 2015.
- [27] S. Kim, Z. Yu, R. M. Kil, and M. Lee, "Deep learning of support vector machines with class probability output networks," *Neural Networks*, vol. 64, pp. 19–28, 2015.
- [28] P. J. Brockwell and R. A. Davis, *Time Series: Theory And Methods*, Springer Series in Statistics, New York, NY, USA, 1989.
- [29] V. N. Vapnik, *Statistical Learning Theory*, Wiley- Interscience, New York, NY, USA, 1998.
- [30] S. Haykin, *Neural Networks and Learning Machines*, Macmillan Publishers Limited, New Jersey, NJ, USA, 2009.
- [31] A. Candel, V. Parmar, E. LeDell, and A. Arora, *Deep Learning with H2O*, United States of America, 2018.
- [32] K. Keerthi Vasani and B. Surendiran, "Dimensionality reduction using principal component analysis for network intrusion detection," *Perspectives in Science*, vol. 8, pp. 510–512, 2016.
- [33] L. Rokach, "Ensemble-based classifiers," *Artificial Intelligence Review*, vol. 33, no. 1-2, pp. 1–39, 2010.
- [34] R. Polikar, "Ensemble based systems in decision making," *IEEE Circuits and Systems Magazine*, vol. 6, no. 3, pp. 21–45, 2006.
- [35] Reid. Sam, *A review of heterogeneous ensemble methods*, Department of Computer Science, University of Colorado at Boulder, 2007.
- [36] D. Gopika and B. Azhagusundari, "An analysis on ensemble methods in classification tasks," *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 3, no. 7, pp. 7423–7427, 2014.
- [37] Y. Ren, L. Zhang, and P. N. Suganthan, "Ensemble classification and regression-recent developments, applications and future directions," *IEEE Computational Intelligence Magazine*, vol. 11, no. 1, pp. 41–53, 2016.
- [38] U. G. Mangai, S. Samanta, S. Das, and P. R. Chowdhury, "A survey of decision fusion and feature fusion strategies for pattern classification," *IETE Technical Review*, vol. 27, no. 4, pp. 293–307, 2010.
- [39] M. Woźniak, M. Graña, and E. Corchado, "A survey of multiple classifier systems as hybrid systems," *Information Fusion*, vol. 16, no. 1, pp. 3–17, 2014.
- [40] G. Tsoumakas, L. Angelis, and I. Vlahavas, "Selective fusion of heterogeneous classifiers," *Intelligent Data Analysis*, vol. 9, no. 6, pp. 511–525, 2005.
- [41] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, part 2, pp. 119–139, 1997.
- [42] S. Chopra and P. Meindl, *Supply Chain Management: Strategy, Planning, and Operation*, Prentice Hall, 2004.
- [43] G. Kushwaha, "Operational performance through supply chain management practices," *International Journal of Business and Social Science*, vol. 217, pp. 65–77, 2012.
- [44] G. E. Box, G. M. Jenkins, G. Reinsel, and G. M. Ljung, *Time Series Analysis: Forecasting and Control*, Wiley Series in Probability and Statistics, John Wiley & Sons, Hoboken, NJ, USA, 5th edition, 2015.
- [45] W.-L. Zhao, C.-H. Deng, and C.-W. Ngo, "k-means: a revisit," *Neurocomputing*, vol. 291, pp. 195–206, 2018.
- [46] C. Li and A. Lim, "A greedy aggregation-decomposition method for intermittent demand forecasting in fashion retailing," *European Journal of Operational Research*, vol. 269, no. 3, pp. 860–869, 2018.
- [47] L. Yue, Y. Yafeng, G. Junjun, and T. Chongli, "Demand forecasting by using support vector machine," in *Proceedings of the Third International Conference on Natural Computation (ICNC 2007)*, pp. 272–276, Haikou, China, August 2007.
- [48] L. Yue, L. Zhenjiang, Y. Yafeng, T. Zaixia, G. Junjun, and Z. Bofeng, "Selective and heterogeneous SVM ensemble for demand forecasting," in *Proceedings of the 2010 IEEE 10th International Conference on Computer and Information Technology (CIT)*, pp. 1519–1524, Bradford, UK, June 2010.
- [49] T. Efeendigil, S. Önüt, and C. Kahraman, "A decision support system for demand forecasting with artificial neural networks and neuro-fuzzy models: a comparative analysis," *Expert Systems with Applications*, vol. 36, no. 3, pp. 6697–6707, 2009.
- [50] E. Duman, M. Uysal, and A. F. Alkaya, "Migrating birds optimization: a new metaheuristic approach and its performance on quadratic assignment problem," *Information Sciences*, vol. 217, pp. 65–77, 2012.

Research Article

Is Deep Learning for Image Recognition Applicable to Stock Market Prediction?

Hyun Sik Sim ¹, Hae In Kim,² and Jae Joon Ahn ²

¹Department of Industrial & Management Engineering, Kyonggi University, Suwon 16227, Republic of Korea

²Department of Information & Statistics, Yonsei University, Wonju 03722, Republic of Korea

Correspondence should be addressed to Jae Joon Ahn; ahn2615@yonsei.ac.kr

Received 6 December 2018; Accepted 10 February 2019; Published 19 February 2019

Guest Editor: Thiago C. Silva

Copyright © 2019 Hyun Sik Sim et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Stock market prediction is a challenging issue for investors. In this paper, we propose a stock price prediction model based on convolutional neural network (CNN) to validate the applicability of new learning methods in stock markets. When applying CNN, 9 technical indicators were chosen as predictors of the forecasting model, and the technical indicators were converted to images of the time series graph. For verifying the usefulness of deep learning for image recognition in stock markets, the predictive accuracies of the proposed model were compared to typical artificial neural network (ANN) model and support vector machine (SVM) model. From the experimental results, we can see that CNN can be a desirable choice for building stock prediction models. To examine the performance of the proposed method, an empirical study was performed using the S&P 500 index. This study addresses two critical issues regarding the use of CNN for stock price prediction: how to use CNN and how to optimize them.

1. Introduction

Stock markets have random walk characteristics. Random walk characteristics in stock markets mean that the stock price moves independently at every point in time. Due to the random walk characteristic, stock market prediction using past information is very challenging [1]. In addition, Carpenter et al. [2] insisted that the stock market can be influenced by complex factors, such as business and economic conditions and political and personal issues. There is a high degree of uncertainty in the stock market, which makes it difficult to predict stock price movements [3].

With the globalization and development of information and communication technology (ICT), however, many people are looking toward stock markets for earning excess returns under a convenient investment environment. Therefore, the study of stock market prediction has been a very important issue for investors.

Stock market prediction methods can be categorized into fundamental analysis and technical analysis [4]. Fundamental analysis is a method of analyzing all elements that affect the intrinsic value of a company, and technical analysis is a way of predicting future stock price through graph analysis.

When fundamental analysis is applied, some problems may occur. For example, forecasting timeliness can be reduced, subjectivity can be intervened, and the difference between stock price and intrinsic value can be maintained for a long time [5]. Due to the limitation of fundamental analysis, many studies related to stock market prediction using technical analysis have been conducted.

In recent years, many researchers have suggested that artificial neural networks (ANNs) provide an opportunity to achieve profits exceeding the market average by using technical indicators as predictors in stock markets [6–9]. Shin et al. [10] proposed a stock price prediction model based on deep learning techniques using open-high-low-close (OHLC) price and volume and derived technical indicators in the Korean stock market.

However, since many financial market variables are intertwined with each other directly or indirectly, it is difficult to predict future stock price movements by using technical indicators alone, even when applying a typical deep learning model.

In this study, a stock price prediction model based on convolutional neural network (CNN) and technical analysis is proposed to validate the applicability of new learning

methods in stock markets. Unlike typical neural network structures, the CNN, which is most commonly applied to analyze visual imagery, can improve learning performance by convolution and pooling processes [11]. For applying the CNN, various technical indicators, which are used for technical analysis, have been generated as predictors (input variables) of the prediction model, and these technical indicators were converted to images of the time series graph. This study compared the forecasting accuracies of the proposed model and the typical ANN model as well as support vector machine (SVM) model to verify the usefulness of deep learning for image recognition in the stock market.

The remainder of this paper is organized as follows. Section 2 describes the theoretical background for typical ANN, SVM, and CNN. Section 3 introduces the proposed model for stock market prediction in this study. Section 4 demonstrates the empirical results and analysis. Finally, we draw conclusions in Section 5.

2. Background

2.1. Typical ANN. A typical ANN model is a data processing system consisting of layers, connection strengths (weights), a transfer function, and a learning algorithm. The ANN has a structure in which relations between input and output values are learned through iterative weight adjustments. The neural network structure consists of a fully connected layer, in which all neurons are combined with adjacent layers.

The ANN consists of a perceptron, called a neuron, and the overall structure of the general ANN is given in Figure 1(a). The general ANN consists of three layers: the input layer, the hidden layer, and the output layer. In the input layer, the neurons correspond to each input variable. The neurons in the hidden layer and output layer perform the function of calculating the summation of input values and weights in the previous layer.

Figure 1(b) represents the relationship between input and output values in each layer. In Figure 1(b), x_1 , x_2 , and x_3 represent input signals and have weights of w_1 , w_2 and w_3 , respectively. The net input function combines the input signal and weight linearly and converts the value through the activation function to output the signal y .

The fully connected layer structure may cause a problem, in which spatial information is lost by ignoring the shape of the data [12]. To increase the representation ability of the data in the ANN model, the number of hidden neurons is increased, or hidden layers are added. However, a vanishing gradient problem occurs when a backpropagation algorithm carries error information from the output layer toward the input layer [13].

2.2. SVM. SVM, developed by Vapnik [14], is an artificial intelligence learning method. It is a machine learning technique based on statistical learning theory and structural risk minimization. The purpose is to identify the optimal separating hyperplane to divide two or more classes of data with the learning mechanism by training the input data. SVM is a type of supervised learning to predict and classify items

and it is well known as useful machine learning algorithm for classification [15].

Assume that there are n number of data points existing in the eigenspace, $\{(\bar{x}_1, c_1), (\bar{x}_2, c_2), \dots, (\bar{x}_n, c_n)\}$, the symbol $C_1 \in \{+1, -1\}$ indicates the classification for data point \bar{x}_1 . These data points serve as the training data for the identification of the optimal separating hyperplane as

$$\bar{w} \cdot \bar{x} - \alpha = 0 \quad (1)$$

The symbol \bar{w} denotes the separating margin and α is a constant. There could be multiple solutions to \bar{w} , but the optimal \bar{w} is the one with the maximum margin. Equation (2) is the solution to the optimization problem:

$$\begin{aligned} & \text{minimize} \quad \frac{1}{2} \|\bar{w}\|^2 \\ & \text{subject to} \quad c_i (\bar{w} \cdot \bar{x} - \alpha) \geq 1, \quad 1 \leq i \leq n \end{aligned} \quad (2)$$

After the network learning obtains the w with the maximum margin, it is then possible to establish the classification \hat{C} by using (3) on the test data that has yet to be classified.

$$\hat{C} = \begin{cases} -1, & \text{if } \bar{w} \cdot \bar{x} - \alpha \leq -1 \\ +1, & \text{if } \bar{w} \cdot \bar{x} - \alpha \geq +1 \end{cases} \quad (3)$$

2.3. CNN. The CNN, as a deep learning technique, is a model that imitates the visual processing of living organisms that recognize patterns or images. The CNN has a structure in which one or more convolutional layers and pooling layers are added to a fully connected layer, which results in an ANN structure.

Figure 2 shows the structure of LeNet-5, which is the most famous CNN algorithm. According to Figure 2, a five-layer CNN was established. LeNet-5 is composed of two convolutional layers for the first two layers and three fully connected layers for the remaining three layers. First, the image of the input layer is filtered through the convolutional layer to extract appropriate features [16].

The convolutional layer is the first layer to extract features from an input image. Convolution preserves the relationship between pixels by learning image features using small squares of input data. Convolution is a mathematical operation that requires two inputs, such as an image matrix and a filter or kernel.

A convolution operation is an elementwise matrix multiplication operation, where one of the matrices is the image and the other is the filter or kernel that turns the image into something else. The output of this is the final convoluted image. If the image is larger than the size of the filter, the filter is moved to various parts of the image to perform the convolution operation. If the convolution operation is performed each time, a new pixel is generated in the output image.

In image processing, there are few sets of filters that are used to perform several tasks. The convolution of an image with different filters (kernels) can perform operations, such as edge detection, blurring, and sharpening, by applying filters.

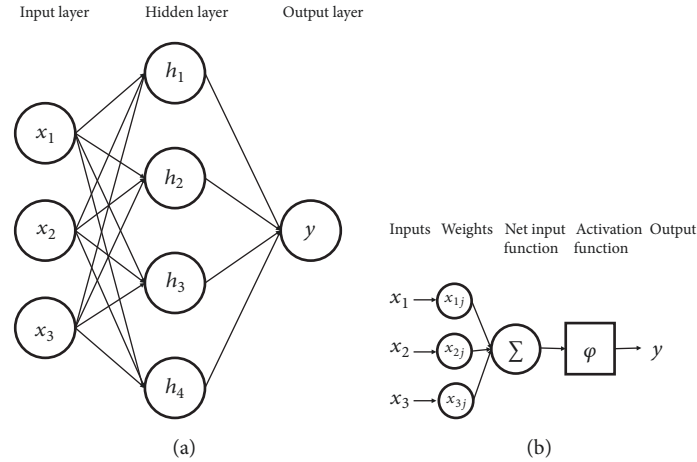


FIGURE 1: Typical ANN structure. (a) The overall structure of the general ANN. (b) The relationship between the input and output values in each layer.

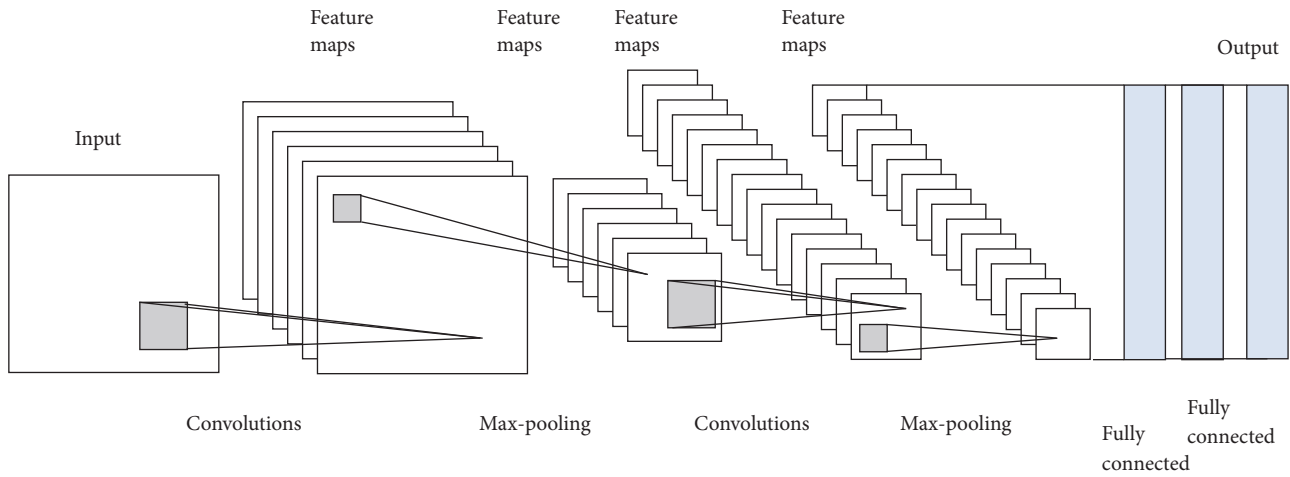


FIGURE 2: The LeNet-5 structure.

In CNNs, filters are not defined. The value of each filter is learned during the training process [17]. Every filter is spatially small (in terms of width and height) but extends through the full depth of the input volume. During the forward pass, each filter is moved across the width and height of the input volume, and dot products are computed between the entries of the filter and the input at any position. As the filter is moved over the width and height of the input volume, a 2-dimensional feature map that gives the responses of that filter is produced at every spatial position [18].

Intuitively, the network learns filters that activate when they see some type of visual feature, such as an edge of some orientation of the first layer or eventually the entire honeycomb or wheel-like patterns within the higher layers of the network. An entire set of filters is generated in each convolutional layer, and each one produces a separate 2-dimensional feature map.

Figure 3 shows the process of generating a feature map for a convolutional layer. The original image is the one on the left, and the matrix of numbers in the middle is the convolutional

matrix or filter. Consider a 4 x 4 matrix, whose image pixel values are 0, 1, 2, and 3, and a 3 x 3 filter matrix, as shown in Figure 3. Then, the convolution of the 4 x 4 image matrix multiplies with the 3 x 3 filter matrix, which results in the feature map, as shown in Figure 3.

The activation functions of every convolutional layer and the first two fully connected layers are shown in (4) (i.e., ReLU (Rectified Linear Unit)). The ReLU function is used to solve the vanishing gradient, which does not reflect the output error of the neural network as it moves away from the output layer in the process of the neural network [19].

$$\max(0, x)$$

$$f(x) = \begin{cases} 0 & \text{for } x < 0 \\ x & \text{for } x \geq 0 \end{cases} \quad (4)$$

Generally, the pooling layer is located after the convolutional layer. The pooling layer was introduced for two main reasons [20]. The first was to perform down sampling (i.e., to reduce

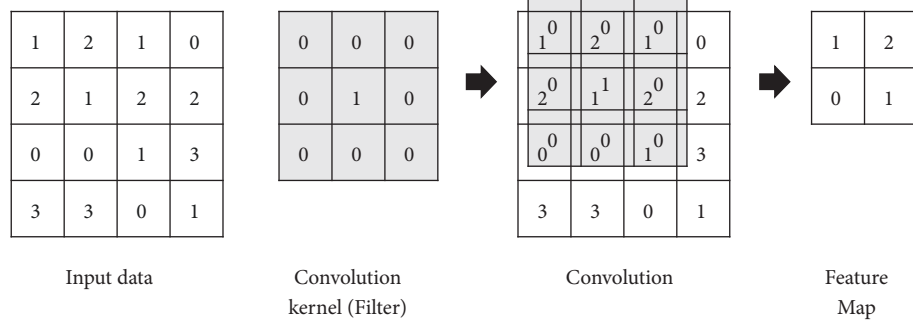


FIGURE 3: The process of generating the feature map of the convolutional layer.

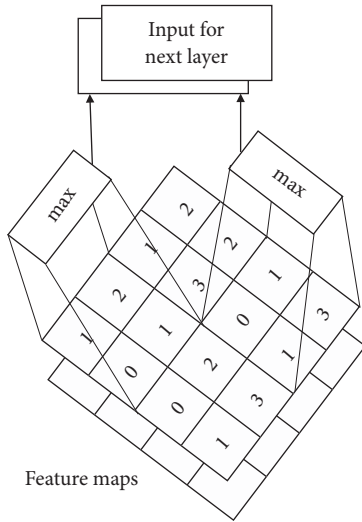


FIGURE 4: The process in the max pooling layer.

the amount of computation that needs to be done), and the second was to only send the important data to the next layers in the CNN max pooling layers by taking the largest element from the rectified feature map, as shown in Figure 4. The most common form is a pooling layer, with filters of size 2×2 , which are applied with a stride of 2 down samples every depth slice in the input by 2 along both the width and height, discarding 75% of the activation. These values are then linked to a fully connected layer, such as an ANN structure, to output the label-specific prediction probabilities

3. CNN Architecture for Building a Stock Price Prediction Model

3.1. Input Image Generation. In this study, historical S&P 500 minute data are used, and these time series data are divided into 30 minute increments for stock price prediction. When learning a prediction model, the closing price and technical indicators are considered as input variables, and target variables are set to values expressed as 1 or 0. If the target has a value of 0, the closing price at time $t - 1$ is higher than the closing price at time t , as shown in (5). In other words, the stock price prediction model proposed in this study learns the moving pattern of the independent variables

for 30 minutes and forecasts the increase or decrease in the stock price after one minute.

$$\text{target} = \begin{cases} 0 & \text{for } \text{close price}_t < \text{close price}_{t-1} \\ 1 & \text{for } \text{close price}_t \geq \text{close price}_{t-1} \end{cases} \quad (5)$$

Table 1 shows the technical indicators used in this study. Nine technical indicators are selected for the prediction model (refer to [21]): simple moving average (SMA), exponential moving average (EMA), rate of change (ROC), moving average convergence divergence (MACD), fast %K, slow %D, upper band, lower band, and %B. Finally, the technical indicators calculated by Table 1 are standardized to have a value between 0 and 1 for converting to images of time series graph.

Now, the technical indicators are converted to the images of a time series graph using the input image of the CNN. Finally, 1100 input images in the training period and 275 input images in the test period are generated. Figure 5 shows the example of the input images in the test period when applying only 3 input variables. In Figure 5, the red line, green line, and blue line indicate the closing prices of the S&P 500 index, SMA 20, and EMA 20, respectively.

3.2. CNN Parameter Settings for the Best Prediction Model Architecture. In this study, the LeNet-5 algorithm is used for stock price prediction. The CNN structure of this study is shown in Figure 6. The $64 \times 64 \times 3$ input image is filtered in the first convolutional layer by $3 \times 3 \times 3$ kernels, with a stride of 1 pixel. Then, max pooling is used in the pooling layer. The main purpose of the pooling operation is to reduce the size of the image as much as possible, taking a 2×2 matrix to minimize pixel loss and obtain the correct characteristic region [22].

The second convolutional layer filters the output of the first convolutional layer using $3 \times 3 \times 3$ kernels, with a stride of 1 pixel. After the pooling process is performed once again, flattening, which is a process of converting a two-dimensional array into one long continuous linear vector, is performed. That is, the process of converting a pooled image pixel into a one-dimensional single vector is performed.

In the fully connected layer, the entire connection of 512 neural networks is performed. The number of neurons in both of the first two fully connected layers is 512. Then,

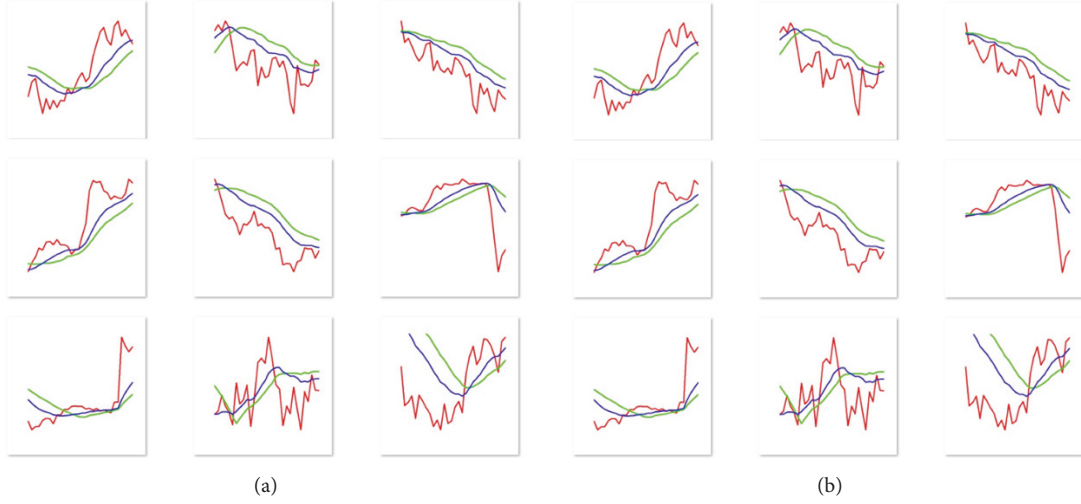


FIGURE 5: Example of the input image. (a) Generated input image when the closing price increases after 1 minute. (b) Generated input image when the closing price decreases after 1 minute.

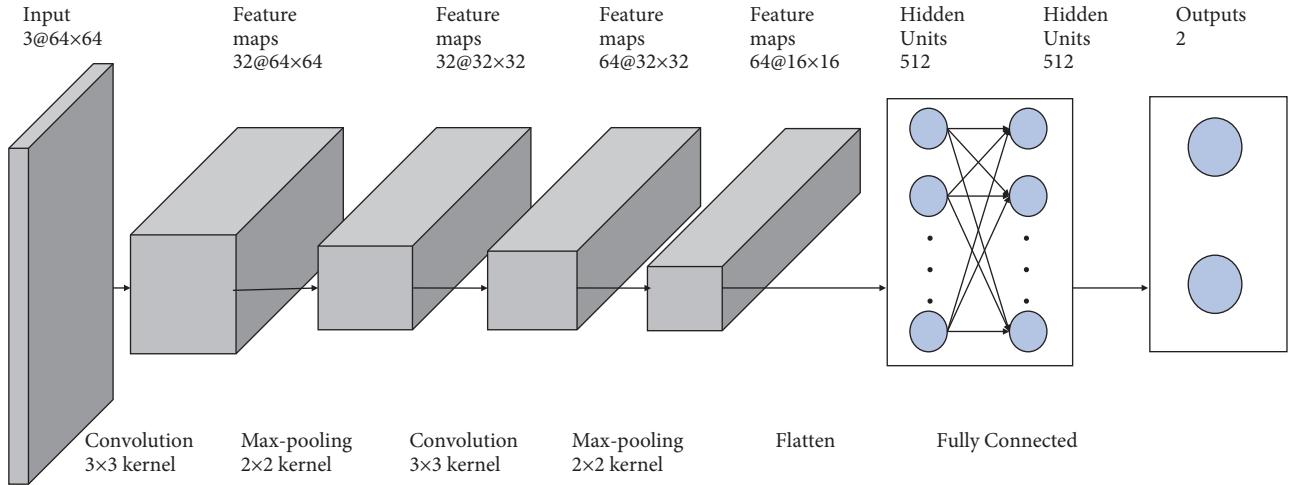


FIGURE 6: The architecture of the CNN for the prediction model.

because the process is a binary classification, the connection goes through an output layer that contains only one node. The last layer uses the sigmoid activation function.

Adaptive Optimization Methods. Stochastic gradient descent (SGD) has been widely used when training CNN models. Despite its simplicity, SGD performs well empirically across a variety of applications but also has strong theoretical foundations [23].

Training neural networks is equivalent to solving the nonconvex optimization problem in

$$\min_{w \in \mathbb{R}^n} f(w) \quad (6)$$

where f represents a loss function. The iterations of SGD can be described in

$$w_k = w_{k-1} - \alpha_{k-1} \widehat{\nabla} f(w_{k-1}) \quad (7)$$

where w_k denotes the k^{th} iteration, α_k represents a (tuned) step size sequence (also called the learning rate), and $\widehat{\nabla} f(w_k)$ denotes the stochastic gradient computed at w_k .

The Adam optimization algorithm is an algorithm that can be used instead of the classical SGD procedure to update network weights iteratively based on training data. The Adam algorithm is popular in the field of deep learning because it achieves good results quickly [24]. The updated Adam equation can be represented in

$$w_k = w_{k-1} - \alpha_{k-1} \frac{\sqrt{1 - \beta_2^k}}{1 - \beta_1^k} \frac{m_{k-1}}{\sqrt{v_{k-1} + \epsilon}} \quad (8)$$

where

$$\begin{aligned} m_{k-1} &= \beta_1 m_{k-2} + (1 - \beta_1) \widehat{\nabla} f(w_{k-1}) \\ v_{k-1} &= \beta_2 v_{k-2} + (1 - \beta_2) \widehat{\nabla} f(w_{k-1})^2 \end{aligned} \quad (9)$$

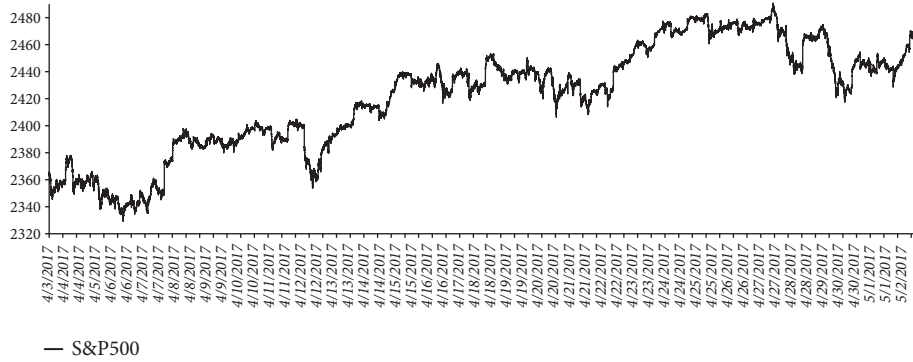


FIGURE 7: Minute closing prices of the S&P 500 index.

$\beta \in [0, 1)$ represents a momentum parameter, and v_0 is initialized to 0.

Dropout. The dropout method introduced by Hinton et al. [25] is known as a very effective way to reduce overfitting when applying neural networks with many hidden layers. This method consists of setting the output of each hidden neuron in the chosen layer to zero with some probability (usually 50%). In this paper, the dropout method was applied after the pooling operations.

Loss Function. The ANN uses the loss function as an indicator to determine the optimal weight parameter through learning [26]. In this study, the mean square error (MSE) and cross entropy error (CEE) were adopted to comprise the objective function (loss function). Equations (10) and (11) show the MSE measure and CEE measure, respectively. y_k represents the output of the neural networks, and t_k represents the target value in (10) and (11).

$$MSE = \frac{1}{n} \sum_k^n (y_k - t_k)^2 \quad (10)$$

When calculating the MSE, the neurons in all output layers are entered. This loss function is most commonly used because it is simple to calculate. Basically, the difference between the output of the model and the target distance is used as an error. The advantage of squaring the distance difference is that the difference between data with small distance differences and the large data error becomes larger, which has the advantage of being able to know exactly where the error is located.

$$CEE = -\sum_k^n t_k \log y_k \quad (11)$$

The CEE only counts the neuron corresponding to the target, which results in a larger penalty as it moves farther from the target.

Epoch and Batch Sizes. An epoch consists of one full training cycle for the data. An epoch is an iteration over the entire training data and target data provided. The epochs are equal to 2500 in this study. The batch size is a term used in machine

learning and refers to the number of training examples utilized in one iteration. The batch size is 1 [27] in this study.

Steps per epoch indicate the number of batch iterations before a training epoch is considered finished. These steps represent the total number of steps (i.e., batches of samples) before declaring one epoch finished and starting the next epoch.

4. Empirical Studies

4.1. Experimental Settings. In this study, the empirical analysis covers a 1-month period. The dataset consists of minute data of the S&P 500 index from 10:30 pm on April 3, 2017, to 2:15 pm on May 2, 2017. The entire dataset covers 41,250 minutes. Figure 7 shows a time series graph of the S&P 500 closing price during the analysis period.

Among the entire dataset, 33,000 minutes are allocated for the training data (80% of the entire data), and 8,250 minutes are allocated for the testing data (20% of the entire data). When the time series data are converted into an image every 30 minutes, the training data consist of 1,100 input images, and the testing data consist of 275 input images.

For experimenting with the CNN algorithm, the technical indicators used for forecasting the stock price in [21] are employed as input variables here.

To evaluate the forecasting accuracy, the following three measurements are employed: hit ratio, sensitivity, and specificity (see (12)–(14)).

$$\text{hit ratio} = \frac{n_{0,0} + n_{1,1}}{n} \quad (12)$$

$$\text{sensitivity} = \frac{n_{0,0}}{n_{0,0} + n_{0,1}} \quad (13)$$

$$\text{specificity} = \frac{n_{1,1}}{n_{1,0} + n_{1,1}} \quad (14)$$

In (12)–(14), $n_{0,0}$ and $n_{0,1}$ represent the number of predicted values of 0 and the number of predicted values of 1 when the actual value is 0, respectively. Additionally, $n_{1,0}$ and $n_{1,1}$ represent the number of predicted values of 0 and the number of predicted values of 1 when the actual value is 1, respectively. The hit ratio is a metric or measure of the prediction model performance when the target variable is binary. While

TABLE 1: Technical indicators used for the proposed prediction model.

Technical indicators	Formula (n=20)	Explanation
SMA	$\frac{\sum(Price, n)}{n}$	n = Time Period
EMA	$Close(i) \cdot P + (EMA(i-1)) \cdot (1-P)$	$Close(i)$ = The closing price at time i $EMA(i-1)$ = Exponentially moving average of the closing price at time $i-1$ P = the percentage using the price value
MACD	$FastMA - Slow MA$	Fast MA is the moving average (5) Slow MA is the moving average (20)
ROC	$100 \cdot \frac{Close - Close\ n\ ago}{Close\ n\ ago}$	
Fast %K	$100 \cdot \frac{Close - Low}{High - Low}$	
Slow %D	$SMA(Slow\ K\%,\ Dma)$	$Slow\ K\%$ = SMA (Fast %K, KMA) KMA = Period of moving average used to smooth the slow %K values
Upper Band	$Middle\ Band + (y \cdot n - standard\ deviation)$	$Middle\ Band$ = n-period moving average
Lower Band	$Middle\ Band - (y \cdot n - standard\ deviation)$	y = factor applied to the standard deviation value
%B	$\frac{Price - Lower\ Band}{Upper\ Band - Lower\ Band}$	

TABLE 2: Input variables for each CNN model.

	Input variables
CNN1	Closing price
CNN2	Closing price, SMA, EMA
CNN3	Closing price, SMA, EMA, ROC, MACD
CNN4	Closing Price, SMA, EMA, ROC, MACD, Fast %K, Slow %D, Upper Band, Lower Band

the hit ratio is simply a measure of discrepancies between the predicted value and actual observations, sensitivity and specificity measure the conditional discrepancies depending on actual observations.

4.2. Experimental Results. In this study, TensorFlow was used for the experiment. TensorFlow is a famous deep learning development framework in which grammar is developed in the form of a Python library. To verify the usefulness of the technical indicators as an input variable, four CNN models are constructed with different technical indicators. The CNN models are created by applying 0, 2, 4, and 9 technical indicators. In this study, these models are called CNN1, CNN2, CNN3, and CNN4, respectively. Table 2 presents the input variables applied to these four models.

Table 3 shows the accuracies of the four models. To determine the adaptive optimization method, all CNN parameters (except for the adaptive optimization method) are applied equally to each model. Here, the dropout probability, batch size, and epoch are fixed at 0.5, 1, and 2500, respectively. Additionally, the steps per epoch in the training and testing data were set to 250 and 50, respectively, and the loss function was the CEE.

As shown in Table 3, when the SGD optimizer is used for the adaptive optimization method, CNNs achieve a high level of predictive performance. CNN1, which is

the prediction model without technical indicators, has the highest hit ratio among the four models. Therefore, technical indicators cannot affect the positive impact of the CNN on stock price forecasting. However, a large difference between the sensitivity and specificity of CNN1 indicates that an overfitting problem occurs due to considering only one input variable.

Table 4 shows the accuracies of the four models with SGD optimizers using different loss functions. From Table 4, we know that the use of the MSE as a loss function increases the predictability rather than the use of the CEE.

The accuracies of the four models with the SGD optimizer and MSE loss function using different dropout probabilities are given in Table 5. CNN1 has the highest hit ratio (0.85) when the dropout probability is 0. The results in Table 5 show that an increase in the dropout probability does not contribute to the predictive performance of the CNN, which is interesting because dropout options are widely known to play an important role in deep learning architecture construction. In the case of this experiment, however, since the learning image of CNN models is simpler than the character recognition or text recognition generally applied to CNNs, it is considered that the dropout option has a negative effect.

Table 6 shows the accuracies of the four CNN models with different steps per epoch when applying the SGD optimizer,

TABLE 3: Accuracy comparison for CNNs with different optimizers during the test period.

	Optimizer	Hit ratio	Specificity	Sensitivity
CNN1	Adam	0.63	0.9545	0.3587
	SGD	0.65	0.9596	0.3810
CNN2	Adam	0.52	0.5822	0.4800
	SGD	0.56	0.5248	0.6144
CNN3	Adam	0.58	0.6460	0.4408
	SGD	0.60	0.6429	0.4840
CNN4	Adam	0.56	0.54	0.5810
	SGD	0.58	0.5939	0.5868

TABLE 4: Accuracy comparison for CNNs with different loss functions during the test period.

	Loss function	Hit ratio	Specificity	Sensitivity
CNN1	MSE	0.66	0.6611	0.6508
	CEE	0.65	0.9596	0.3810
CNN2	MSE	0.67	0.6825	0.6209
	CEE	0.56	0.5248	0.6144
CNN3	MSE	0.62	0.6151	0.6302
	CEE	0.60	0.6429	0.4840
CNN4	MSE	0.62	0.6114	0.6393
	CEE	0.58	0.5939	0.5868

TABLE 5: Accuracy comparison for CNNs with different dropout probabilities during the test period.

	Dropout probability	Hit ratio	Specificity	Sensitivity
CNN1	0	0.85	0.9593	0.6971
	0.25	0.67	0.6904	0.6507
	0.5	0.66	0.6611	0.6508
CNN2	0	0.62	0.6679	0.5878
	0.25	0.68	0.6992	0.6744
	0.5	0.67	0.6825	0.6209
CNN3	0	0.64	0.9559	0.2487
	0.25	0.64	0.9091	0.3012
	0.5	0.62	0.6151	0.6302
CNN4	0	0.66	0.6548	0.6872
	0.25	0.62	0.6040	0.6375
	0.5	0.62	0.6114	0.6393

TABLE 6: Accuracy comparison for CNNs with different steps per epoch during the test period.

	Steps per epoch (train / test)	Hit ratio	Specificity	Sensitivity
CNN1	400 / 100	0.68	0.7221	0.6413
	800 / 200	0.54	0.5324	0.5434
CNN2	400 / 100	0.53	0.6155	0.46
	800 / 200	0.52	0.6257	0.4118
CNN3	400 / 100	0.61	0.9304	0.31
	800 / 200	0.52	0.4824	0.5238
CNN4	400 / 100	0.54	0.6220	0.4734
	800 / 200	0.54	0.7112	0.3450

TABLE 7: Predictive accuracies of ANNs and SVMs.

	Hit ratio	Specificity	Sensitivity
ANN1	0.4872	0.6866	0.2979
ANN2	0.5602	0.5674	0.5522
ANN3	0.5653	0.6561	0.4801
ANN4	0.5573	0.6269	0.4626
SVM1	0.48	0.8881	0.0922
SVM2	0.4655	0.8582	0.0922
SVM3	0.5018	0.4851	0.5177
SVM4	0.5455	0.5149	0.5745

TABLE 8: Optimized parameters for CNNs.

Parameter	Considered value (option)	Selected Value (option)
Adaptive optimization method	Adam, SGD	SGD
Dropout probability	0 ~ 0.5	0
Loss function	MSE, CEE	MSE
Steps per epoch (train / test)	200 / 50 ~ 800 / 200	200 / 50

MSE loss function and dropout probability of 0. Based on the results of Table 6, we can realize that an increase in steps per epoch causes an overfitting problem and results in a decrease in accuracy. As a result, it is not effective in increasing the number of steps for stock price prediction based on a CNN using technical indicators.

To verify the performance of CNN models, ANN and SVM models are generated and their accuracies are evaluated. The same input variables for CNNs in Table 2 are applied to ANNs and SVMs. Before exploring the ANN and SVM for stock price prediction, small preliminary experiments were performed to obtain proper parameter settings for the successful implementation of the ANN and SVM. As a result, the number of hidden layers, the number of hidden units, and the activation function of ANN are set to be 1, 3, and sigmoid, respectively. And SVM uses polynomial kernel to make a nonlinear classification interface.

Based on the results show in Table 7, when the ANN and SVM are applied, technical indicators are shown to be input variables positively affecting the stock price prediction, as opposed to when the CNN is applied. Nevertheless, the predictive performances of the ANN and SVM are lower than that of the CNN (refer to Table 5 when dropout probability is 0). Therefore, CNNs using input images can be a useful method for stock price prediction. In practice, CNN models are good at detecting patterns in images such as lines. CNNs can detect relationships among images that humans cannot find easily; the structure of neural networks can help detect complicated relationships among features. For example, in CNN, color images are composed of RGB channels, and the features of input for each channel can be extracted. This allows CNN to extract features better than when it uses a vectorized input such as ANN [28].

5. Concluding Remarks

In this study, we attempted to check the applicability of the CNN for stock market prediction. Previously, many

researchers have suggested that ANNs offer a chance to achieve profits in financial markets. Therefore, this study determined the predictive performances of the CNN and ANN to validate the usefulness of the CNN. In addition, SVM, well known for useful classification algorithm, was employed to verify the usefulness of the CNN.

To design the CNN architecture, this study focused on two points. First, the CNN parameters were optimized. For this, the experiments were performed over the parameter range given in Table 8, and the best experiments were obtained. Second, technical indicators, which are well known as efficient input variables in stock price forecasting, were verified to play a role as a suitable input image for CNNs when technical indicators are converted into images.

Our empirical experiments demonstrate the potential usefulness of the CNN by showing that it could improve the predictive performance more than the ANN. In this sense, the CNN appears to be a desirable choice for building stock prediction models. In addition, technical indicators were input variables that did not positively affect the stock price prediction when the CNN was implemented for the prediction model. This result is because technical indicators cannot be good input variables, as they are similar to the moving pattern of the closing price. Therefore, building a stock price prediction model with better performance can be expected if other factors that move opposite the stock price, such as gold price and interest rate, are considered as input variables for the CNN. As a result of this study, it is difficult to predict technical indicators of stock market by general data mining classification technique. Therefore, CNN, which is a deep learning method that analyzes time series data into graphs, can be a useful for stock price prediction.

Data Availability

The data used in this study can be accessed via <https://www.kesci.com/home/dataset/5bbdc2513631bc00109c29a4/files>.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

- [1] E. F. Fama, "Random walks in stock market prices," *Financial Analysts Journal*, vol. 51, no. 1, pp. 75–80, 1995.
- [2] G. A. Carpenter, S. Grossberg, N. Markuzon, J. H. Reynolds, and D. B. Rosen, "Artmap: a neural network architecture for incremental learning supervised learning of analog multidimensional maps," *IEEE Transactions in Neural Networks*, vol. 3, no. 5, pp. 698–713, 1992.
- [3] Y. Kara, M. Acar Boyacioglu, and Ö. K. Baykan, "Predicting direction of stock price index movement using artificial neural networks and support vector machines: the sample of the Istanbul stock exchange," *Expert Systems with Applications*, vol. 38, no. 5, pp. 5311–5319, 2011.
- [4] B. G. Malkiel, *A Random Walk Down Wall Street*, W. W. Norton & Company, New York, NY, USA, 1999.
- [5] J. L. Bettman, S. J. Sault, and E. L. Schult, "Fundamental and technical analysis: substitutes or complements?" *Accounting & Finance*, vol. 49, no. 1, pp. 21–36, 2009.
- [6] T. Fischer and C. Krauss, "Deep learning with long short-term memory networks for financial market prediction," *European Journal of Operational Research*, vol. 270, no. 2, pp. 654–669, 2018.
- [7] J. Lee, "A stock trading system based on supervised learning of highly volatile stock price patterns," *Journal of Korean Institute of Information Scientists and Engineers*, vol. 19, no. 1, pp. 23–29, 2013.
- [8] C.-M. Hsu, "A hybrid procedure for stock price prediction by integrating self-organizing map and genetic programming," *Expert Systems with Applications*, vol. 38, no. 11, pp. 14026–14036, 2011.
- [9] Y. K. Kwon, S. S. Choi, and B. R. Moon, "Stock prediction based on financial correlation," in *Proceedings of the 7th annual conference on Genetic and evolutionary computation*, pp. 2061–2066, ACM, Wash, D.C., USA, June 2005.
- [10] D. Shin, K. Choi, and C. Kim, "Deep learning model for prediction rate improvement of stock price using RNN and LSTM," *The Journal of Korean Institute of Information Technology*, vol. 15, no. 10, pp. 9–16, 2017.
- [11] Y. J. Song and J. W. Lee, "A design and implementation of deep learning model for stock prediction using tensorflow," *KIISE Transactions on Computing Practices*, vol. 23, no. 11, pp. 799–801, 2017.
- [12] R. J. Schalkoff, *Artificial Neural Networks*, vol. 1, McGraw-Hill, New York, NY, USA, 1997.
- [13] Y. Bengio, P. Simard, and P. Frasconi, "Backpropagation applied to handwritten zip code recognition," *Neural Computation*, vol. 1, no. 4, pp. 541–551, 1989.
- [14] V. N. Vapnik, *The Nature of Statistical Learning Theory*, Springer, New York, NY, USA, 1995.
- [15] C. L. Jan, "An effective financial statements fraud detection model for the sustainable development of financial markets: evidence from Taiwan," *Sustainability*, vol. 10, no. 2, p. 513, 2018.
- [16] Y. LeCun, B. Boser, J. S. Denker et al., "Gradient-based learning applied to document recognition," in *Shape, Contour and Grouping in Computer Vision*, pp. 319–345, Springer, Berlin, Germany, 1999.
- [17] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Object recognition with gradient-based learning," in *Proceedings of the IEEE*, vol. 86, pp. 2278–2324, 1998.
- [18] S. Sarraf and G. Tofghi, "Classification of alzheimer's disease using fmri data and deep learning convolutional neural networks," *Computer Vision and Pattern Recognition*, Article ID 1603.08631, 2016, <https://arxiv.org/abs/1603.08631>.
- [19] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, vol. 15, pp. 315–323, 2011.
- [20] D. Cires, U. Meier, J. Masci, L. M. Gambardella, and J. Schmidhuber, "Flexible, high performance convolutional neural networks for image classification," in *Proceedings of The Twenty-Second International Joint Conference on Artificial Intelligence*, pp. 1237–1242, 2011.
- [21] S. B. Achelis, *Technical Analysis from A to Z*, McGraw Hill, New York, NY, USA, 2001.
- [22] M. Abadi, A. Chu, and I. Goodfellow, "Deep learning with differential privacy," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, ACM, 2016.
- [23] H. Robbins and S. Monro, "A stochastic approximation method," *Annals of Mathematical Statistics*, pp. 102–109, 1985.
- [24] D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," *Machine Learning*, Article ID 1412.6980, 2014, <https://arxiv.org/abs/1412.6980>.
- [25] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proceedings of the 26th Annual Conference on Neural Information Processing Systems (NIPS '12)*, pp. 1097–1105, Lake Tahoe, Nev, USA, December 2012.
- [26] J. B. Hampshire and A. H. Waibel, "Novel objective function for improved phoneme recognition using time-delay neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 1, no. 2, pp. 216–228, 1990.
- [27] D. R. Wilson and T. R. Martinez, "The general inefficiency of batch training for gradient descent learning," *Neural Networks*, vol. 16, no. 10, pp. 1429–1451, 2003.
- [28] Y. Tsai, J. Chen, and J. Wang, "Predict forex trend via convolutional neural networks," *Journal of Intelligent Systems*, Article ID 1801.03018, 2018.