Hindawi
Mathematical Problems in Engineering
Volume 2022, Article ID 6540928, 11 pages
https://doi.org/10.1155/2022/6540928

Research Article

Design of Quantitative Trading System Based on Data Mining Method under Software and High-Performance Computing

Shizhou Feng[1,2] **and Jing Du**[3]

[1]*Chongqing College of Mobile Communication, Chongqing 401520, China*
[2]*Chongqing Key Laboratory of Public Big Data Security Technology, Chongqing 401420, China*
[3]*Chongqing College of International Business and Economics, Chongqing 401520, China*

Correspondence should be addressed to Jing Du; 164904121@stu.cuz.edu.cn

Received 9 May 2022; Revised 20 June 2022; Accepted 1 July 2022; Published 10 August 2022

Academic Editor: Xiantao Jiang

The size of funds managed by all hedge funds in the world has exceeded 2.7 trillion US dollars. The funds of various funds and asset management products managed by quantitative investment account for about 30% of the total global trading volume, and in various large stock exchanges around the world, various quantitative investment methods contribute nearly 50% volume of transactions. The construction of a quantitative trading strategy requires first statistical analysis of the information in the securities and futures market and then backtesting the quantitative model with historical data. In view of the practical application of quantitative trading, this study designs a quantitative trading system based on the data mining method. The main development tool used is the numerical computing software MATLAB, and four cores are designed: quantitative stock selection, strategy backtesting, time-series analysis, and portfolio management. The system supports modules for simple trading decisions. It abandons the traditional method of predicting the absolute value of the future price of stock index futures and adopts a new method of predicting the future price trend of stock index futures. This method avoids the huge impact of the accuracy of the absolute value of the prediction on the final investment in the traditional method and also reduces the high dependence of investors on the accuracy of the absolute value. This study also introduces the support-vector machine algorithm in data mining and the quantitative trading system model in data mining. The accuracy of investment transactions in the experiment is also simulated by using the support-vector machine.

1. Introduction

In market trading, general investors will use relevant knowledge in mathematics and statistics, such as quantitative models, to establish corresponding trading strategies. It uses computer technology and financial engineering technology to define and describe various transaction operations in the transaction process, so as to help investors make better decisions in the investment process. Such a processing method can execute the previously established trading strategy according to the established rules and orders during the trading process, so as to avoid human interference and at the same time avoid making irrational decisions. In the early stage of the scale of the financial field, there were many pioneers who dared to explore various feasible investment methods. At present, the development momentum of the global capital market is becoming more and more rapid. The quantitative investment method has been gradually recognized in China and has attracted great attention from investors in related fields such as finance. In different types of financial markets, different types of investors choose and participate in different types of financial market transactions. This is like bonds, futures, stocks, foreign exchange, etc. These investors use the guarantee of their own trading behavior to bear various risks in the trading process and obtain corresponding benefits. Therefore, in financial market transactions, trading methods that can bring considerable income and value will receive great attention from various investors. These trading methods are extremely valuable to investors. In the process of data mining, the most critical point is to continuously train the processed input and output data through the algorithm and obtain the corresponding model. After further validation of the model, it guarantees that the model can express the relationship between input and output in a certain situation. It can also calculate new input data through the model, so as to obtain the expected corresponding new output data. In the entire process of quantifying the model to achieve the final effect, it is necessary to continuously process a large amount of data, which runs through several steps such as the generation of the model, the simulation verification of the model, and the actual verification of the model. In general, data mining techniques are crucial in the entire process of quantitative model implementation.

The traditional forecasting methods are used in various forecasting processes, such as neural network forecasting models. Such traditional forecasting models need to rely on a large amount of data in the process of forecasting. Therefore, if the data volume of the sample is small, such prediction methods cannot perform well, especially for the stock index futures data mentioned in this study. The prediction model constructed according to the relevant theoretical knowledge of the support-vector machine can predict these cases with small sample data well.

Based on the existing transaction data, this study compares and evaluates the prediction models corresponding to the four kernel functions in static and dynamic simulation methods, and selects the optimal kernel function. Through the optimal kernel function, a prediction model with strong practical properties is established based on real transaction data.

2. Related Work

Wei et al. have shown that feature selection (FS) is an important part of data mining and machine learning. Many researchers are working to find more effective, more accurate, and fewer features. Some algorithms have been shown to be effective, such as binary particle swarm optimization (BPSO), genetic algorithm (GA), and support-vector machine (SVM) [1]. Sukawattanavijit et al. proposed the GA-SVM algorithm as a method for classifying RADARSAT-2 (RS2) multifrequency SAR images and multispectral Thaichote images (THEOS). It compares the results of the GA-SVM algorithm with the results of a grid search algorithm [2]. Du et al. argue that support for vector devices (SVMs) is a general error associated with radio-to-boundary bandwidth, while traditional SVMs only estimate the threshold size and ignore the radio frequency sensitivity of related data conversions. Therefore, it can improve the traditional SVM by controlling the radius and threshold [3]. Duan et al. proposed a new OAA-SVM method with the Markov modeling (OAA-SVM-MS). An experimental study of a comparative library confirms that the performance of the OAA-SVM-MS algorithm is significantly better than the classical OAA-SVM algorithm and several other SVM-class algorithms [4]. Zhao et al. face the challenge of designing the best design (MRP) portfolio. This problem plays an important role in accounting arbitrage systems (also known as shoe trading) in financial markets. The goal of the best MRP model is to build documentation of background properties that have satisfactory regression properties and complete conversion properties [5]. Pan and Long document a computational form of expert portfolio analysis for stock investments and trades. It includes two multifactor models and two trading strategies that follow the trends. The smart portfolio technology of investing in stocks goes beyond the classic portfolio economic process. It uses a multifactor model for stock selection and quantitative trading strategies instead of buying and holding [6]. Tsantekidis et al. developed a data mining system that allows traders to be trained in different pairs of forex currencies. This provides a way to help RL representatives throughout the market. It also allows the misuse of repetition of repetitive training models without more serious risk [7]. Euch et al. show that the typical behavior of market participants at high frequency results in inequality and wholesale instability. To this end, it introduces a simple micromodel for asset estimates based on the Hawkes process [8]. Lei et al. believed that the basic idea of the PPDM was to modify the data in such a way that the data mining algorithm could be effective without compromising the security of the suspicious information contained in the data [9]. Chaurasia and Pal analyze breast cancer data from the UCI Machine Learning Wisconsin database. His aim is to develop the standardized models for predicting breast cancer using data mining techniques [10]. Yan and Zheng use the bootstrap method to assess the impact of data mining on basic anomalies. He also found that many basic signals are important predictors of submarket returns, especially after considering data mining. This predictive power is further explained after an increase in emotion and in markets with higher arbitrage restrictions [11]. Slater et al. show that the history of data mining as a way back to research data analysis and that methods for making useful and general decision-making have been established [12]. Huang et al. developed an additional device for calculating rough roughness projections and developing an algorithm based on a compatible matrix. This differs from the static method [13], which calculates a method by updating all relative tables. Mújica-Vargas et al. introduce the kernel fuzzy C-means algorithm, augmented by a Gaussian radio-based kernel based on the M estimator [14]. Srinivas et al. investigated the use of optical decontamination techniques to solve different radial and azimuth patterns of Laguerre-Gaussian (LG) bonds. It also tested the performance of individual and complex LG bonds and compared them with simulations [15]. Although all of these studies have the content investigated in this study, there are some cases of insufficient exposure or insufficient depth.

3. Methods

3.1. Support-Vector Machine Model Based on Data Mining

3.1.1. Linearly Separable Problem. In the support-vector machine theory, if it wants to find the optimal classification line, it needs to find the straight line that makes the distance between the two straight lines L1 and L2 the farthest. The support-vector machines are robust and sparse. The distance between two straight lines is also known as the class interval. In general, as long as the order center is long, there is the general power of the classifier. Similarly, in a multidimensional space, if it wants to find the best level of customization, it only needs to find the maximum level of customization of that unique space. In a certain plane, when the maximum classification interval is reached, the sample points passed by the two straight lines L1 and L2 are the support vectors. It is shown in Figure 1.

To find the optimal classification line, the decision function $g(a)$ formed by the sample set can be defined as follows:

$$(u\cdot a)+s=0. \tag{1}$$

u in this formula is a two-dimensional vector; $(u\cdot a)$ is the inner product of w and a samples; and s is the offset of the classification line. It then normalizes sample set K so that each sample (a_i, b_i) conforms to the following:

$$b_i\left(\left(u\cdot a_i\right)+s\right)\geq 1,\quad i=1,2,\ldots,m. \tag{2}$$

The two lines passing through the support vector can be obtained as follows:

$$L_1\colon\ \left(u\cdot a_i\right)+s=1;\quad L_2\colon\ \left(u\cdot a_i\right)+s=-1. \tag{3}$$

The result $2/\|u\|$ of the classification interval can be calculated. Knowing that the classification interval value located in the optimal plane is the largest, it can be regarded as a restricted optimal problem Min $(\|u\|^2/2)$, and a Lagrange multiplier is introduced as follows:

$$L(u,s,\lambda)=\frac{\|u\|^2}{2}-\sum_{i=1}^{m}\lambda_i\left[b_i\left(ua_i+s\right)-1\right]. \tag{4}$$

λ is the Lagrange multiplier. In order to find the minimum value, the partial derivatives of u and s are calculated here as follows:

$$\begin{aligned}u&=\sum_{i=1}^{m}\lambda_i a_i b_i;\\ \sum_{i=1}^{m}\lambda_i b_i&=0.\end{aligned} \tag{5}$$

Substituting the calculated result into it, we get the following:

$$\min_{\lambda}\left(\frac{1}{2}\sum_{i=1}^{m}\sum_{k=1}^{m}b_i b_k\lambda_i\lambda_k\left(a_i\cdot a_k\right)-\sum_{i=1}^{m}\lambda_i\right),\quad i=1,2\ldots m. \tag{6}$$

Knowing that this problem has constraints, there is only one solution, and the minimum value is as follows:

$$\begin{aligned}u'&=\sum_{i=1}^{m}\lambda_i' a_i b_i;\\ s'&=b_k-u'\cdot a_k.\end{aligned} \tag{7}$$

According to relevant research, when the Lagrange multiplier of the test sample is greater than 0, the sample at this time can play a role in helping classification. These samples that can help the classification are the samples under the intersection of the two straight lines L1 and L2 mentioned in the text—the support vector. In mathematical optimization problems, Lagrange multipliers are a method of finding the extremum of a multivariate function whose variables are constrained by one or more conditions. The Lagrange multipliers of nonsupport vectors are all 0. According to the above, the decision function can be deduced as follows:

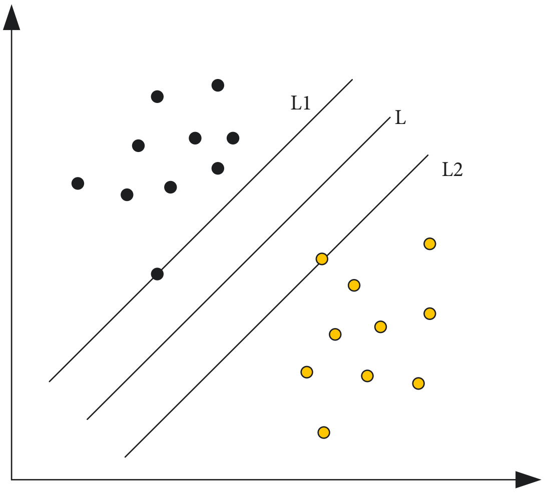

Figure 1: All samples are linearly separable.

$$f(a)=\operatorname{sgn}\left\{\left(u'a+s'\right)\right\}. \tag{8}$$

s' in formula (8) represents the threshold at classification. Its final form can be obtained as follows:

$$f(a)=\operatorname{sgn}\left\{\left(\sum_{i=1}^{m}\lambda_i' b_i\left(a_i\cdot a\right)+s'\right)\right\}. \tag{9}$$

3.1.2. Linear Inseparable Problem. When the situation in Figure 2 is encountered, the samples at this time are in a state of complete linear inseparability. The appropriate solution should be interpreted by the technology of the support-vector machine. At this time, the sample set is linearly separable or tends to be linearly separable. Therefore, when it is necessary to classify and divide the sample set in the high-dimensional space, it can be carried out by determining the generalized optimal classification line.

It maps the sample set of the low-dimensional space to the high-dimensional space, which will change from Figure 2 to Figure 1. $W^m \longrightarrow P$ is to map the data in the low-dimensional space W^m to the high-dimensional space P. This is a nonlinear mapping, and the mapping process of sample set V is as follows:

$$\begin{aligned}V&=\left\{\left(a_i,b_i\right),i=1,2\ldots M,a_i\in W^m,u_i\in\{-1,1\}\right\},\\ \overline{V}&=\left\{\left(\varphi\left(a_i\right),b_i\right),i=1,2\ldots M,a_i\in W^m,\varphi\left(a_i\right)\in P,b_i\in\{-1,1\}\right\}.\end{aligned} \tag{10}$$

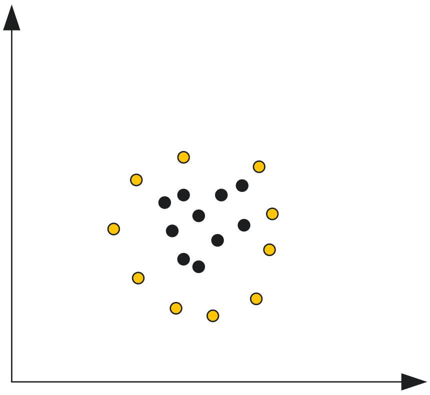

Figure 2: The sample is linearly inseparable.

It can get the following:

$$\begin{aligned}&(u \cdot \varphi(a)) + s = 0,\\ &b_i((u \cdot a_i) + s) \geq 1 - \varepsilon_i, \quad i = 1, 2 \ldots m.\end{aligned} \tag{11}$$

Among them, ε_i is consistent with the definition already mentioned in the text and is a slack variable. In the small-scale field, the equations of the general customization line and its relative constraints are close under the condition of a linear inequality. In order to achieve the best solution to the problem at the highest point, the problem can be solved by means of the Lagrange multiplier and the corresponding levels can be extracted in part at the same time.

$$\min_\lambda \left(\frac{1}{2} \sum_{i=1}^{m} \sum_{k=1}^{m} b_i b_k \lambda_i \lambda_k (\varphi(a_i) \cdot \varphi(a_k)) - \sum_{i=1}^{m} \lambda_i \right), \ i = 1, 2 \ldots m. \tag{12}$$

The decision function in the high-dimensional space can be calculated as follows:

$$f(a) = \operatorname{sgn} \left\{ \left(\sum_{i=1}^{m} \lambda' b_i (\varphi(a_i) \cdot \varphi(a)) + s' \right) \right\}, \quad S \geq \lambda_i \geq 0. \tag{13}$$

According to Mercer's theorem, if a certain kernel function can satisfy all the conditions established in the theorem, then the kernel function must be able to be transcribed into the inner product form in a certain space. Mercer's theorem is that any positive semidefinite function can be used as a kernel function.

$$f(a) = \operatorname{sgn} \left\{ \sum_{i=1}^{m} \lambda \prime b_i H((a_i) \cdot a_k) + s' \right\}, \quad S \geq \lambda_i \geq 0. \tag{14}$$

It classifies samples with linearly inseparable features according to the associated kernel function. While ensuring that the final calculation results are consistent, it can avoid data operations in high-dimensional spaces. In the current research background, there are many kinds of kernel functions commonly used in support-vector machine-related research, such as sigmoid function, linear kernel function, and polynomial kernel function. The use of different kernel functions can create different types of vector machine models. The sigmoid function is a smooth function that facilitates derivation and can compare data. It guarantees that there will be no problem with the data amplitude and is suitable for forward propagation.

Multikernel function, sigmoid kernel function, Gauss kernel function, and line kernel function are shown as follows:

$$H(a, a') = \left(q a^T a' + s \right)^d; \quad q \geq 0, \ s \geq 0, \tag{15}$$

$$\begin{aligned}&H(a, a') = \tan h(q\mu(a, a') + s); \quad q \geq 0, \ \mu(a) > 0, \ s < 0,\\ &H(a, a') = \exp\left\{ -q \,|\, a - a' |^2 \right\}; \quad q \geq 0,\\ &H(a, a') = a^T a'.\end{aligned} \tag{16}$$

Equation (15) is the decision function.

3.2. *Quantitative Trading System Model.*

The real realization of the quantitative program trading system platform can have a huge impact on the trading environment of the market. It designs an efficient and feasible quantitative program trading system platform, which can connect to the trading interfaces of multiple markets at the same time. It manages the transaction data of the market at the same time and can handle related operations required by multiple markets at the same time. The four pillars of trading are psychological quality, innovation ability, capital management, and strategy. The premise of the design of the quantitative program trading system platform is that the design of each module of the platform must conform to the relevant trading rules of the capital market. At the same time, it needs to reserve cross-market and cross-category related settings in advance. Generally speaking, the trading market can be divided into the futures market, the bond market, and the stock market. Various trading markets and related trading categories are embedded in the quantitative program trading system platform, thereby giving birth to a cross-market and cross-category highly interactive trading system platform. Figure 3 shows the overall structure of the quantitative trading system:

In this quantitative program trading system, it is necessary to create the thinking ideas of traders first and then program these thinking ideas. In this case, the idea of trading thinking and its programming are related. It views and optimizes trading ideas through executable programs and uses them as samples for statistical validation. It inputs sample statistical tests and optimizes through relevant strategies, followed by strategy monitoring and maintenance. If it does not pass the test outside of the multisample, it needs to be sent back and reprogrammed; and if it passes, the practical test can be carried out. In the practical test, if it fails, it is necessary to reconstruct the relevant trading thinking. If passed, it can perform policy monitoring and maintenance. Relevant investment analysis methods include company stock selection strategy, excess return attribute analysis, VAR model, etc., all of which incorporate the concept of quantitative investment, as shown in Figure 4.

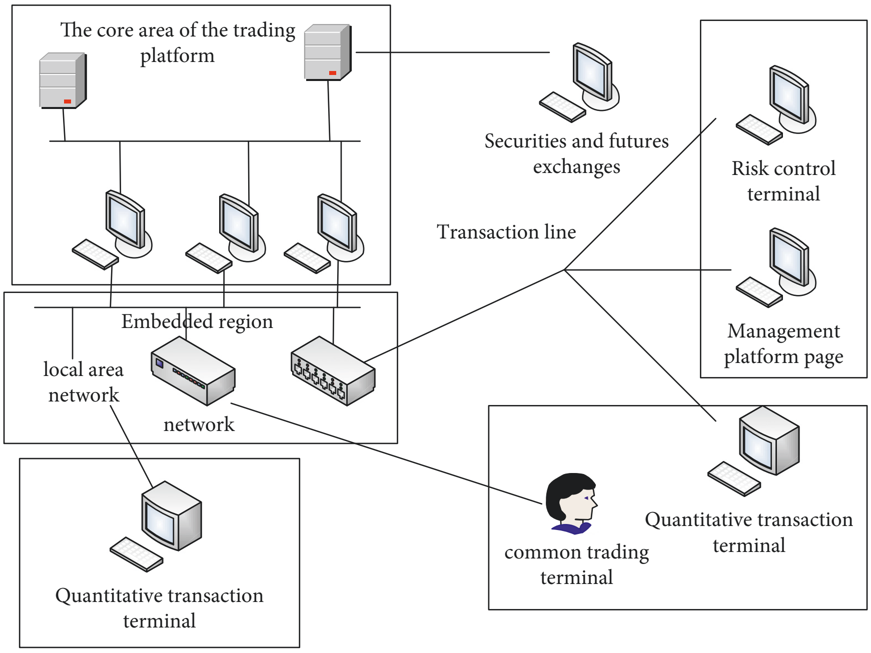
The core area of the trading platform
Securities and futures exchanges
Risk control terminal
Transaction line
Management platform page
Embedded region
local area network
network
Quantitative transaction terminal
common trading terminal
Quantitative transaction terminal

Figure 3: The overall structure of the quantitative trading system.

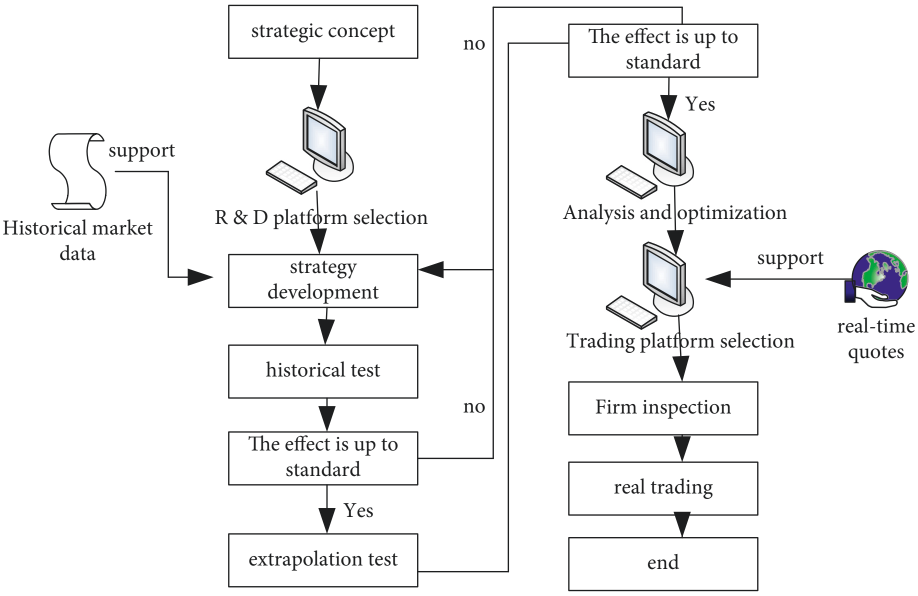

Figure 4: Design scheme of the quantitative trading system.

3.3. Quantitative Transaction Investment Model Based on Data Mining. As shown in Figure 5, the quantitative investment software framework is constructed based on data mining technology. Under the function of this framework, the relevant quantitative investment software is divided into partial databases and four subsystems. These databases mainly include financial product database, historical transaction database, transaction effect database, transaction rule database, and transaction sequence pattern database. Among them, the financial product database contains product information, such as financial statements, announcements, and the announcement. The historical transaction data of various financial products, such as price and quantity, are stored in the historical transaction database. The transaction performance database contains the transaction performance data behind each order. The existence of the transaction rule database can effectively prevent various order problems, such as invalid orders and excess orders. The transaction sequence pattern library holds valid patterns that have been processed by the data mining subsystem. According to the information storage format, the objects used for mining include relational databases, object-oriented databases, data warehouses, text data sources, multimedia databases, spatial databases, temporal databases,

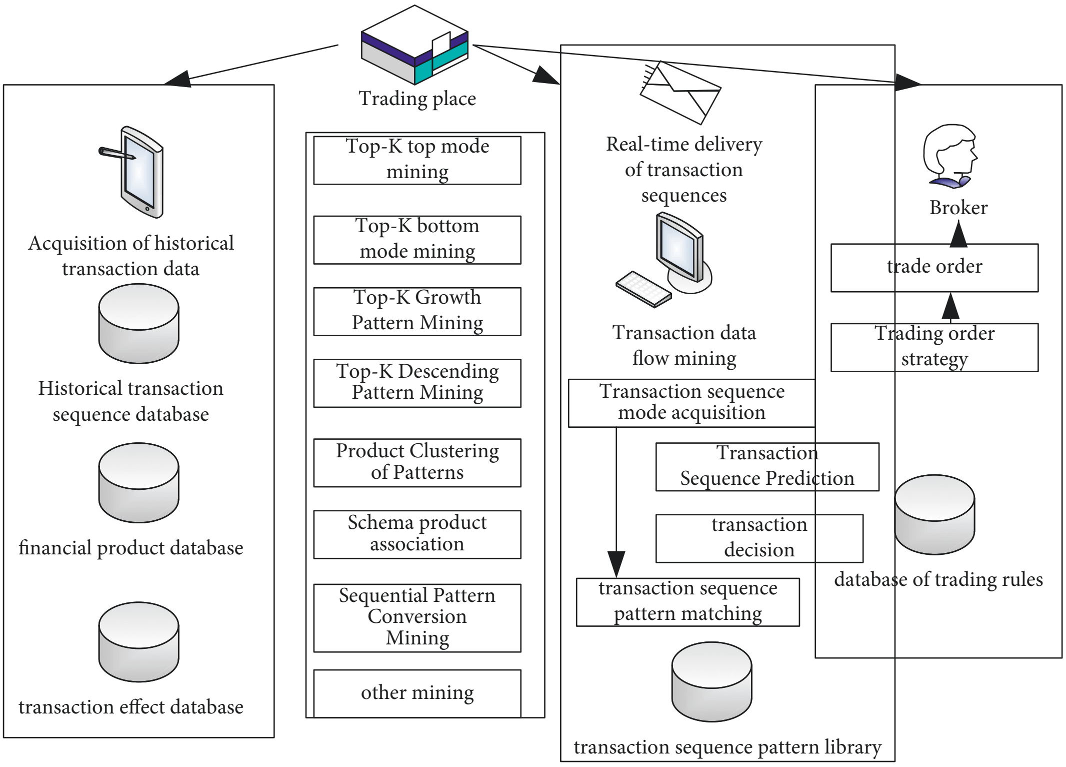
Trading place
Acquisition of historical transaction data
Historical transaction sequence database
financial product database
transaction effect database
Top-K top mode mining
Top-K bottom mode mining
Top-K Growth Pattern Mining
Top-K Descending Pattern Mining
Product Clustering of Patterns
Schema product association
Sequential Pattern Conversion Mining
other mining
Real-time delivery of transaction sequences
Transaction data flow mining
Transaction sequence mode acquisition
Transaction Sequence Prediction
transaction decision
transaction sequence pattern matching
transaction sequence pattern library
Broker
trade order
Trading order strategy
database of trading rules

Figure 5: Quantitative investment system based on the data mining model.

heterogeneous databases, and the internet. From the above description, it can be seen that the functions of each database are independent and interdependent. Each database has a clear division of labor for the data storage of each part. With the support of these databases, the entire framework has a clearer construction route.

In the process of data mining, Bayesian networks are generally used. The Bayesian network transforms the existence of multiple factors into an inference model that can be probabilistically described in the form of a description of nonqualitative causal relationships. In this inference model, there are causal relationships, conditional correlations, etc. among the nodes. Therefore, the relevant characteristics of the Bayesian network can be applied to the formulation of trading strategies. In the stock market, the future price movements of stocks are influenced to some extent by historical prices and other factors. Therefore, there is a causal relationship between the two. It sets the historical data as the upper node in the Bayesian network, through which the future price trend and approximate value of the stock market can be inferred. If the corresponding price of the stock under the maximum conditional probability is found, the future price of the stock market can be predicted. A variety of approximate inference algorithms have been proposed in Bayesian network inference research. It is mainly divided into two categories: simulation-based methods and search-based methods.

As shown in Figure 6, in data mining, there are many aspects that machines need to learn. But its main responsibility is to simulate various learning behaviors of humans through computers. It also uses relevant existing knowledge to reorganize the learned behavior. This in turn enables the acquisition of experience knowledge and through repeated learning to perfect the machine's own flaws. The ultimate goal of machine learning is to estimate the dependency between the input and output of a specific system after continuous learning of the existing training samples. At the same time, it can predict the unknown output according to the estimated result.

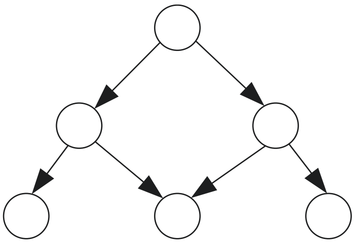

Figure 6: Directed acyclic graph.

4. Quantitative Trading Experiment of Stock Futures

4.1. Experimental Data Preprocessing and Model Static Simulation. In this study, when selecting relevant sample data, after screening and analysis, the relevant data in the main contracts of Shanghai stock index futures are selected, which are basic market data and technical indicator data. In actual situations, the changes and ranges of various data are different. Therefore, these data will also be in different orders of magnitude, and the order of magnitude difference between some data is very large. It is conceivable that when predicting relevant results, the influence of data variables with too large magnitudes on the results is far greater than the influence of data variables with small magnitudes. Therefore, these data variables must be normalized before making predictions. This, thus, ensures that valid information produced even by data variables of small magnitude can be preserved. In the process of prediction, two extreme cases need to be considered, that is, the magnitude of the data variable is extremely large or extremely small. At this time, the support-vector machine model will produce very obvious errors when making predictions. The commonly used methods for normalization are max-min normalization or statistical normalization. The normalization processing method used in this study is the former, and the original samples are normalized by the maximum-minimum normalization method. Then, through the supervisory analysis method, the basic market data and technical indicator data that have been standardized are subjected to dimensionality reduction, respectively. The correlation matrix is calculated by MATLAB software for the sample data after dimension reduction, and the results are shown in Tables 1 and 2:

Table 1: Correlation matrix of futures market input vectors.

	a1	a2	...	a7	a8
a1	1	0.982367	...	0.69447	0.782348
a2	0.985262	1	...	0.785688	0.771344
a3	0.997964	0.980787	...	0.685037	0.72055
...	...	...	...	...	...
a6	0.172649	0.267411	...	0.995843	0.387087
a7	0.710928	0.709601	...	1	0.83417
a8	0.604319	0.785338	...	0.86722	1

Table 2: Correlation matrix of technical indicators.

	a1	a2	...	a11	a12
a1	1	0.985971	...	−0.056733	0.336908
a2	0.984866	1	...	-0.18922	0.314401
a3	0.026125	-0.07812	...	0.681756	0.201492
...	...	...	...	...	...
a10	−0.05673	0.695669	...	0.379624	0.033254
a11	0.270256	0.421153	...	1	0.922858
a12	0.331685	0.214577	...	0.941117	1

It analyzes Tables 1 and 2, where each number is the correlation coefficient between each input variable. According to the data in the table, in the input vector of the basic market and the input vector of technical indicators, there are still many characteristic indicators with high correlation. Therefore, it is also necessary to perform dimensionality reduction processing on the data of the relevant input vectors again through the analysis method. This improves the efficiency and robustness of the model. It uses the princomp function carried by MATLAB to analyze the input vectors in the basic market and technical indicators that have been standardized, respectively. First of all, by analyzing the input vector of the basic market, it can be clearly found that the total contribution of the variance of the first three main factors is very high, which is 94.33%. When performing classification prediction based on the relevant data of the input vector of the basic market, according to the above data, only the first three main factors need to be combined to form the input vector of the prediction model. After analyzing the input vector of technical indicators, it can be seen that the total value of variance contribution of the first five main factors exceeds 94.33% after the vector is processed by the analytical method. Therefore, when classifying and predicting the data of the input vector of the technical indicators, it is only necessary to use the first five factors to form the input vector required by the prediction model. At this point, the original data have been dimensionally reduced from three-dimensional to five-dimensional, thereby improving the operating efficiency of the prediction model.

Under the dual practice of innovation and compatibility analysis, the final input vector of the forecast model can be obtained. In this work, based on the relevant data of the basic product vector and technical indicators, a genetic algorithm is used to find the best positions in the support-vector machine models in different kernel functions. Through relevant input trackers, static simulation is performed for predictive models under four different kernel functions. Based on the previously known data, it predicts how the current closing price will increase or decrease compared to the previous closing price. Through the analytic product entry process, the results of the static simulation of the predictive models under different kernel functions are shown in Table 3.

We can see from the data analysis in Table 3 that the Gaussian radial function in the basic product entry vector is a more obvious function. The index data of this kernel function are also better than the four kernel functions. According to the data in the table, the physical activity is in line with the Gaussian radial function. Comparing the mass kernel function with the Gaussian radial kernel function and linear kernel function, we can see that the multikernel function is more accurate than the other two kernel functions. However, the function of normal, memory, and other indicators is much worse than the other two major functions. In the four kernel functions, the performance of the sigmoid kernel function is the worst and the performance of each rating index is worse than that of the other three kernel functions. Based on the calculations required in this case, the Gaussian radial function is used to construct an appropriate forecast model based on the base product vector. Two-dimensional Gaussian function convolution can be performed in two steps. It first convolves the image with a one-dimensional Gaussian function and then convolves the convolution result with the same one-dimensional Gaussian function in a vertical direction. Here, 36 data were selected to present the results of their simulation, as shown in Figures 7 and 8.

4.2. Dynamic Simulation of Predictive Models. Next, the Gaussian radial nucleus function, linear function, maximum function, and sigmoid nucleus function were used for static simulation using technical index inputs. The results obtained

Table 3: Static simulation data of various kernel functions in the basic market input vector.

Evaluation indicators	Linear kernel function	Polynomial kernel function	Gaussian radial kernel function	Sigmoid kernel function
Accuracy (%)	87.12631	85.81974	87.998	60.98709
Precision (%)	88.69452	95.10187	91.00322	53.71696
Recall (%)	80.21648	67.42716	82.23175	53.11587
F-measure (%)	83.97035	80.10553	86.39344	53.42352

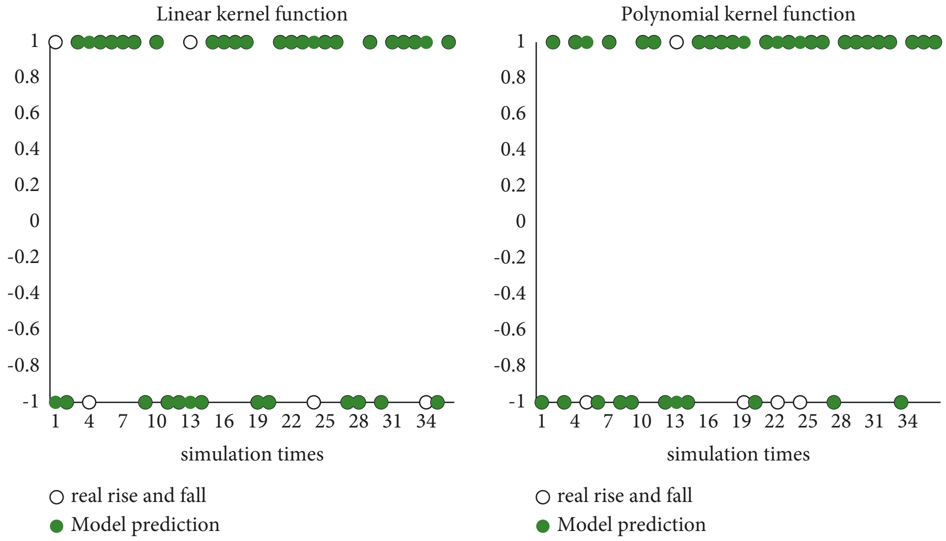

Figure 7: Simulation of linear kernel function and polynomial kernel function under the basic market input vector.

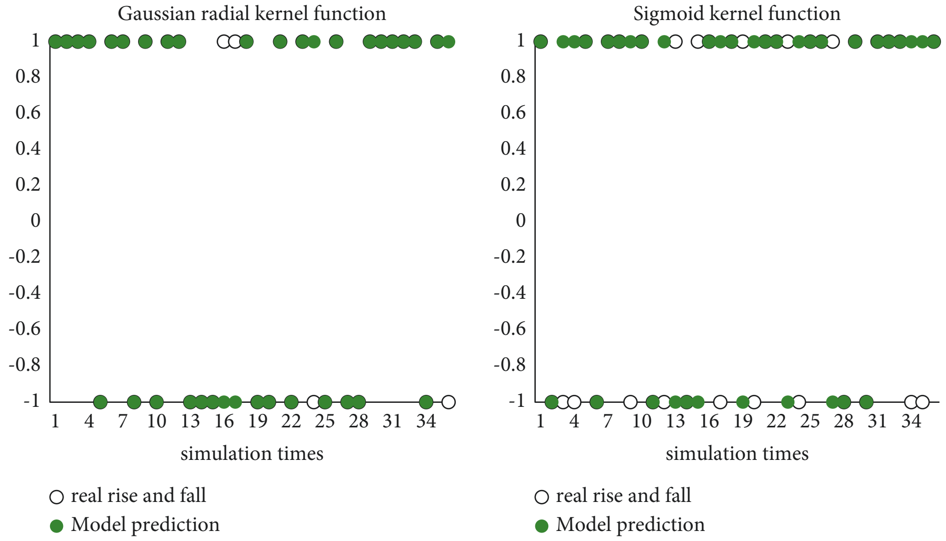

Figure 8: Simulation of Gaussian radial kernel function and sigmoid kernel function under the basic market input vector.

are the same as in the static simulation using the base market input vector. In the case of using the technical indicator input vector, among the four kernel functions, the performance of the Gaussian radial kernel function is still the best, and each evaluation indicator is 1. This result shows that under the condition that the information of a certain day is known, the prediction model of the Gaussian radial kernel function can make an extremely accurate prediction of the rise and fall of the closing price of the day. In the subsequent process of building a dynamic prediction model with technical indicators as the core, the Gaussian radial kernel function will still be used.

According to the simulation processing in the above static and dynamic situations, it can be seen that under the condition of static simulation, the obtained prediction model has a high recognition rate. However, when dealing with the prediction of the closing price under the known trading data of the day, the static prediction model cannot accurately predict the rise and fall of the closing price of the day. Therefore, the static forecasting models have no

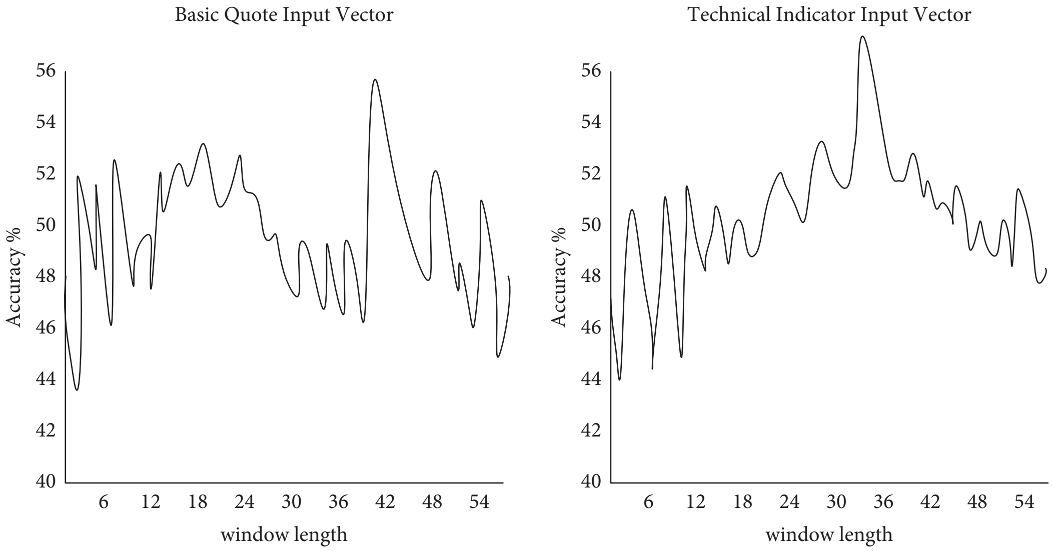

Figure 9: The optimal window length of the basic market input vector and the technical indicator input vector.

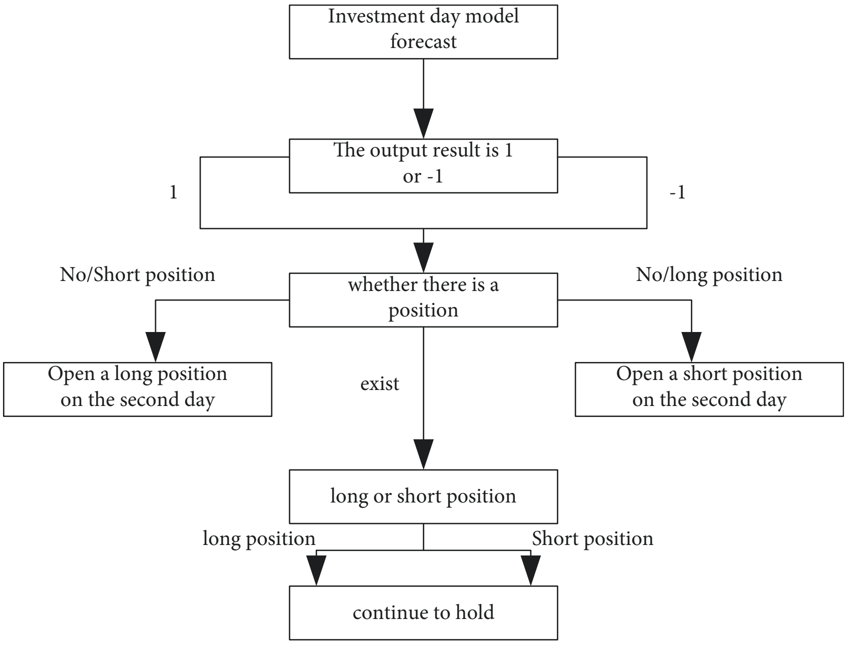
Investment day model forecast
The output result is 1 or -1
1
-1
No/Short position
whether there is a position
No/long position
Open a long position on the second day
exist
Open a short position on the second day
long or short position
long position
Short position
continue to hold

Figure 10: Prediction strategy flow.

practical value and cannot provide meaningful assistance in the actual investment process [16]. Accordingly, this study will create a new dynamic prediction model through the optimal kernel function selected by static simulation, so as to improve the practical application value of the classification prediction model. Through the principal component analysis method, the input vectors corresponding to the basic market and technical indicators are obtained, respectively. It performs dynamic simulation processing on these two vectors, respectively, and finally obtains the window length under the optimal rolling time corresponding to the two. The result is shown in Figure 9. According to the data in the figure, the optimal rolling window length corresponding to the input vector of the basic market is 40. In this case, the prediction accuracy rate is 55.68%. The prediction accuracy corresponding to the input vector of technical indicators is 57.28%. The optimal rolling window length at this time is 33.

4.3. Construction of the Quantitative Trading Strategy Model. In this study, based on the dynamic classification prediction model established with the Gaussian radial kernel function as the core obtained in the previous study, the preliminary trading strategies based on the two systems of the basic market input vector and the technical indicator input vector will be established, respectively. Then, the performance of the trading strategies under the two input vector systems is rigorously compared, and the optimal input vector system is selected. This system is used to predict the rise and fall of the main contract prices of Shanghai stock index futures. In the process of the actual processing of the trading strategy, the dynamic simulation processing of the trading measurement is carried out in this study to find the optimal rolling window length. It also establishes the input vector and builds the prediction model according to the respective data corresponding to the basic market after the closing of the trading day and the technical indicators. Then, it predicts the ups and downs of the closing price within one day after the trading day, and at the same time, converts the obtained prediction results into real trading signals. The process is shown in Figure 10.

5. Discussion

When investors speculate on stock index futures, they usually implement corresponding measures according to their predictions about the future price changes of stock index futures. If an investor predicts that the future price of stock index futures will fall, they will sell the contract and take a short position. If investors predict that the future price of stock index futures will rise, they will buy the contract and establish a long position at the same time. In the speculative process of stock index futures, the absolute value of the future price of stock index futures is generally predicted, and the future price trend of stock index futures is predicted based on this value. Taking these predictions as the basis for speculation on stock index futures, if the absolute value of the predicted future price of the stock index futures is too large and differs greatly from the actual value, it will bring a large loss to the transaction. The support-vector machines have excellent performance in classification problems. Among the existing research studies on the regression forecasting of the stock market based on the support-vector machines, most of the research studies use a certain kind of kernel function instead of a variety of kernel functions to establish the forecasting model. It does not compare and analyze the performance of various kernel functions.

6. Conclusions

Based on the research and analysis of similar quantitative investment literature, combined with the needs and characteristics of each link of investment decision-making, and considering the existing technical level, a quantitative trading system based on the data mining method is designed. The system mainly includes 4 core modules of quantitative stock selection, strategy backtesting, expansion function and portfolio management, and 1 auxiliary function of data preprocessing. Through the processing of multidimensional and multilevel historical data, in the form of actual cases, the specific application of each module is discussed, and the results with reference significance for actual transaction decision-making are given. Overall, the system design is relatively complete, but there is still a lot of room for improvement. The quantitative stock selection module not only introduces the most classic and widely used multifactor stock selection model, but also discusses the stock selection model based on the classification algorithm and clustering algorithm. The investment portfolio constructed by using the stock selection results of the multifactor strategy can significantly outperform the performance benchmark over a long period of time, with a steadily rising excess return. For the stock selection model based on the classification algorithm and the clustering algorithm, the former uses the SVM algorithm to construct a stock rise and fall classifier, and the latter superimposes the hierarchical clustering and K-means clustering to construct a two-stage clustering model, and finally selects, the most profitable stock class is selected as the target of quantitative investment. The first model uses stock price technical data, and the second model uses company financial data, each with its own emphasis and complementary to each other. The development of quantitative investment in China is in full swing. The content studied in this study is only the tip of the iceberg in its broad field. Compared with the complexity of actual transaction decision-making, the research in this study is not deep enough. Some functional modules are not up to standard due to the technical level. Not reducing or even giving up, for example, intelligent optimization plays a key role in the improvement of trading strategies, but it has been stranded due to the complexity of algorithm implementation. The existing functions are relatively single, and the bonding between the modules is not tight enough. With the advancement of science and technology, the improvement of the system, and the development of the market, the application of quantitative investment in China's securities and futures market will continue to deepen, and emerging fields such as high-frequency trading and artificial intelligence financial advisors will be developed.

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 no conflicts of interest.

Acknowledgments

This research study was sponsored by Project of Association of Fundamental Computing Education in Chinese Universities, Ideological and Political Reform and Practice of "Trinity" Course under the Background of New Engineering—Taking Python Language Programming as an Example (2022-AFCEC-635), and Project of Chongqing College of Mobile Communication, Exploration and Practice of Ideological and Political Construction in the "Trinity" Course of New Engineering (22JG318). The authors thank these projects for supporting this article.

References

[1] J. Wei, R. Zhang, Z. Yu et al., "A BPSO-SVM algorithm based on memory renewal and enhanced mutation mechanisms for

feature selection," *Applied Soft Computing*, vol. 58, no. 1, pp. 176–192, 2017.

[2] C. Sukawattanavijit, J. Chen, and H. Zhang, "GA-SVM algorithm for improving land-cover classification using SAR and optical remote sensing data," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 3, pp. 284–288, 2017.

[3] J. Z. Du, W. G. Lu, X. H Wu, J. Y. Dong, and W. M. Zuo, "L-SVM a radius-margin-based SVM algorithm with LogDet regularization," *Expert Systems with Applications*, vol. 102, no. JUL, pp. 113–125, 2018.

[4] Y. Duan, B. Zou, J. Xu, F. Chen, J. Wei, and Y. Y. Tang, "OAA-SVM-MS a fast and efficient multi-class classification algorithm," *Neurocomputing*, vol. 454, no. 2, pp. 448–460, 2021.

[5] Z. Zhao, R. Zhou, and D. P. Palomar, "Optimal mean-reverting portfolio with leverage constraint for statistical arbitrage in finance," *IEEE Transactions on Signal Processing*, vol. 67, no. 7, pp. 1681–1695, 2019.

[6] H. Pan and M. Long, "Intelligent portfolio theory and application in stock investment with multi-factor models and trend following trading strategies," *Procedia Computer Science*, vol. 187, no. 1, pp. 414–419, 2021.

[7] A. Tsantekidis, N. Passalis, and A. S. Toufa, "Price trailing for financial trading using deep reinforcement learning," *Transactions on Neural Networks and Learning Systems*, IEEE, vol. 01, no. 99, , pp. 1–10, 2020.

[8] O. El Euch, M. Fukasawa, and M. Rosenbaum, "The microstructural foundations of leverage effect and rough volatility," *Finance and Stochastics*, vol. 22, no. 2, pp. 241–280, 2018.

[9] L. X. Lei, C. J. Chunxiao, J. W. Jian, Y. Jian, and R. Yong, "Information security in big data: privacy and data mining," *IEEE Access*, vol. 2, no. 2, pp. 1149–1176, 2014.

[10] V. Chaurasia and S. Pal, "A novel approach for breast cancer detection using data mining techniques," *Social Science Electronic Publishing*, vol. 3297, no. 1, pp. 2320–9801, 2017.

[11] X. S. Yan and L. Zheng, "Fundamental analysis and the cross-section of stock returns: a data-mining approach," *Review of Financial Studies*, vol. 30, no. 4, pp. 1382–1423, 2017.

[12] S. Slater, S. Joksimović, V. Kovanovic, R. S. Baker, and D. Gasevic, "Tools for educational data mining," *Journal of Educational and Behavioral Statistics*, vol. 42, no. 1, pp. 85–106, 2017.

[13] Y. Huang, T. Li, C. Luo, H. Fujita, and S. Horng, "Matrix-based dynamic updating rough fuzzy approximations for data mining," *Knowledge-Based Systems*, vol. 119, no. MAR, pp. 273–283, 2017.

[14] D. Mújica-Vargas, B. E. Carvajal-Gámez, G. Ochoa, and J. Rubio, "Robust Gaussian-base radial kernel fuzzy clustering algorithm for image segmentation," *Electronics Letters*, vol. 55, no. 15, pp. 835–837, 2019.

[15] S. Pachava, A. Dixit, and B. Srinivasan, "Modal decomposition of Laguerre Gaussian beams with different radial orders using optical correlation technique," *Optics Express*, vol. 27, no. 9, Article ID 13182, 2019.

[16] J. Zhang, N. Wu, J. Li, and F. Zhou, "A novel differential fault analysis using two-byte fault model on AES Key schedule," *IET Circuits, Devices and Systems*, vol. 13, no. 5, pp. 661–666, 2019.