Research Article

Application and Legal Governance of Computational Technologies in Credit Investigation: An Interdisciplinary Perspective

Ying Zhang, Zhen Xie, and Xiuwei Qu

1Post-Doc at the Institute of Studies for AI & Law, Tsinghua University, Beijing 100084, China
2Assistant Researcher at the Institute for Digital Technology and Law, China University of Political Science and Law, Beijing 100088, China
3Credit Department, Postal Savings Bank of China, Beijing 100032, China
4School of Date Science, Henan Mechanical and Electric Vocational College, Zhengzhou, Henan 451191, China

Correspondence should be addressed to Ying Zhang; zhangying.365@hotmail.com

Received 25 January 2022; Revised 4 March 2022; Accepted 8 March 2022; Published 4 July 2022

Copyright © 2022 Ying Zhang 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 purpose of this interdisciplinary study is to investigate the applications of modern computational technologies in the field of credit investigation and discuss the related legal governance measurements in order to unblock the bottlenecks in the future development of credit investigation market. By analyzing the computational technologies and algorithms most commonly used in credit data collection and data storage, data transmission and data access, data analysis and processing, data calculation, result output and effect evaluation, this paper summarizes and proposes a unified general process of modern credit investigation, pointing out that in this general process, low data quality, privacy violation, algorithmic bias are the main challenges in the big data era, and countermeasures like data quality control, privacy protection, and algorithm governance need to be taken seriously into account in order to further explore the great potential of the credit investigation under the legal framework.

1. Introduction

As a fundamental business in financial sector, credit investigation has a significant impact on economic-social development. Take China as an example, the modernization of Chinese credit investigation system had started in 2010s, according to estimates [1] in 2012 the Chinese credit investigation market had reached 0.3 billion dollars and had a positive contribution to the growth of consumption and investment which accounted for 4.28% of the Chinese GDP growth. Currently, the Chinese credit investigation market has already exceeded 13 billion dollars in 2021, and still has much room for the improvement compared to the U.S with the mature credit investigation market of $130 billion in 2021 [2].

Today, in the general context of the widespread use of computational technologies, several overall trends in the credit investigation industry can be observed: firstly, the credit data resource has expanded from traditional credit data to alternative data (Figure 1). Secondly, credit investigation businesses has expanded its scope from credit reporting, information inquiry and credit scoring to other modernized businesses such as credit antifraud, illegal intrusion/access detection, user profiling, account tracking and risk monitoring. Thirdly, the increased private investment on big data and computing power has led to a shift in credit service providers from traditional credit bureaus to big data platforms. And fourthly, modern credit investigation is so efficient that it puts forward more urgent requirements for legal compliance in different stages of credit data collection and storage, data transmission and access, data analysis and processing, data calculation, result output, and evaluation.

In recent years, many researches have pointed out the key issues in the evolutionary progress of modern credit
investigation. For instance, Stiglitz and Weiss [3] pointed out that the information asymmetry between banks and enterprises is one of the main reasons of SMEs’ financing difficulty, and credit rationing built on models can greatly reduce information asymmetry in the credit market. Abdou and Pointon [4] carried out a survey on the statistical models such as stepwise logistic regression, linear programming, or neural network that could transform relevant data into numerical measures to assist the credit decision-making. Marqués et al. [5] illustrated different resampling algorithms to resolve the core issue of class imbalance, and discussed the application of logistic regression as well as support vector machine in credit scoring. Lohokare et al. [6] explained the application of artificial neural networks in credit scoring based on social media data collected from SMS in smartphone. He et al. [7] compared logistic regression, decision trees and neural networks to verify the feasibility and validity of credit scoring models applied in credit evaluation. However, in these literature, there has been little focus to date on the summation of the application of computational technologies in the modern credit investigation; thus, this paper proposes a modern generic credit investigation process which summarizes the application of various modern computational technologies and algorithms, and on the basis of which, how to develop the modern investigation under a governance framework in the interdisciplinary approach is explored.

This paper makes several principal contributions:

1. This paper describes and compares the applications of modern computing technologies in credit investigation, and these technologies are divided into four categories: data collection techniques, encrypted data transmission and data access, data processing and analysis, and privacy computing architectures. Technological trends and approaches of credit investigation and credit assessment in the context of big data are highlighted during the discussion.

2. This paper outlines various computational algorithms such as decision-support algorithms, data cleaning algorithms, and data preprocessing algorithms applied to modern credit, and evaluation techniques of algorithm applications.

3. This paper proposes a unified general credit investigation process on the basis of computing technologies and algorithms, illustrating the workflow and different stages of modern credit reference.

4. This paper further discusses the bottlenecks of applying modern computing in the process of credit investigation and provides legal countermeasures from an interdisciplinary perspective.

2. Computational Technologies Applied in Modern Credit Investigation

The traditional credit system of credit investigation was dominated by the central banks and financial industries due to institutional and technical constraints; thus, the data resources were relatively limited and had low-involvement compared to the large population in each country.

2.1. Data Collection Technology. The data used in modern credit investigation business is composed by data from different scenarios, credit databases, and information distribution channels. And data collection technology is the key to integrate different data sources.

2.1.1. Credit Information Databases. The world’s well-known credit information databases include World Base of D&B [8], Credit Risk Database (CRD) in Japan [9], ZestFinance in US, and the Credit Reference Center of the People’s Bank of China (PBCCRC) [10]. Most traditional credit databases have centralized and multibackup management of data.

Currently, distributed credit databases (DDB) characterized by separated physical distribution, logical integrity, and site autonomy are becoming a new trend [11]. At the
physical level, with DDB, data could be stored and distributed in different units and can form a unified database network through logical relationships, effectively solving the problem of limited centralized database capacity and scale storage of a single entity; DDB could also optimize the access mechanism of the whole database and allow greater access concurrency; most importantly, the distributed credit information database solves the connection problem between multiple business sectors and multi-region databases, overcoming the difficulties of centralized database construction and the difficulty of unifying nonrelational data. At the same time, the distributed architecture of DDB avoids the single point of failure bottleneck of the centralized database and can quickly call the backup node after a node goes down to ensure data integrity and nonloss, with stronger disaster recovery capability.

2.1.2. Automated Web Crawlers. Web crawler is a program or Internet robot that extracts target information and data automatically. As illustrated in Figure 2, the automated web crawler often starts from one or several initial web pages, gets the URL of the initial web page, during the process of grabbing the web page, constantly extracting new URLs from the current page until a certain stop condition of the system is met. Web crawler needs to filter the links that are not relevant to the topic according to a certain web analysis algorithm, keep the useful links and put them into a queue of URLs waiting to be crawled, then it selects the next web page URL to be crawled from the queue according to a certain search strategy and repeat the above process until it stops when a certain condition of the system is reached. All web pages crawled will be stored by the system for the purpose of analyzing, filtering, and indexing for later query and retrieval; for focused crawlers, the analysis results obtained are still possible to give feedback and guidance on the subsequent crawling process.

Today, in addition to the traditional credit information databases run by central banks and credit bureaus, increasing credit data (such as business registration data, judicial enforcement data, social networks, and identity traits) are grasped by web crawler technology [12]. It is simple and inexpensive to implement, while avoiding administrative costs.

2.1.3. Credit APIs. APIs stands for Application Programming Interfaces, and they are mainly used to build standardized connections between different credit databases, investigation platforms, and data collection terminals, and are considered as the foundation of innovation in modern application-driven credit investigation. Increasing sites are opening their resources by external APIs to developers, users, and small- and medium-sized websites, making the content more relevant between sites, and feeding back the credit mechanism of the whole database and allow greater access concurrency; most importantly, the distributed credit information database solves the connection problem between multiple business sectors and multi-region databases, overcoming the difficulties of centralized database construction and the difficulty of unifying nonrelational data. At the same time, the distributed architecture of DDB avoids the single point of failure bottleneck of the centralized database and can quickly call the backup node after a node goes down to ensure data integrity and nonloss, with stronger disaster recovery capability.

2.2. Encrypted Data Transmission and Access

2.2.1. Encrypted Transmission of Credit Data. Modern credit investigation requires secure data transmission, which also implies the extensive use of trusted keys or trusted certificates, secure transmission protocols, and strong cryptography. Encryption, as the major element of cryptography, has already been used as a fundamental tool for achieving secure data transmission in credit investigation businesses. By performing certain arithmetic processing on plaintext data or files, the data can become segments that cannot be read directly, thus ensuring that data and information are protected from plagiarism, tampering, or unauthorized access by others. Cryptography in credit investigation mainly achieves three purposes: firstly, privacy protection, preventing the content of the transmitted information or user identification from being read by illegal users; secondly, data security protection, preventing data from being displaced or altered; and thirdly, identity signature, ensuring that the data are identifiable (source subject) and verifiable (for a specific target), enabling authentication and signature functions [13].

2.2.2. Credit Data Access Control. Credit access control is to control the user's access ability and scope to credit information data through certain ways and means after identifying the user's legal identity, to prevent illegal access to unauthorized system resources. Access rights are restricted through different access control policies to achieve the

![Figure 2: Implementation of web crawler in credit investigation.](image)
maximum data resource sharing within the scope of legal authorization. Usually, these access control policies include discretionary access control (DAC), mandatory access control (MAC), role-based access control (RBAC), and attribute-based access (ABAC) (Table 1).

DAC is an access control technique commonly used in multiuser environments like credit reference. The core idea of DAC is the subject who owns the data resources can autonomously grant access to other subjects, specifying which devices and data resources can be accessed by each legitimate user and what type of access could be performed.

MAC means that the subject who owns the data resource could predefine the trustworthiness level of the user or data demander and the security level of the resource, and when the user or the demander makes an access request to a determined data resource, the system compares the two to determine whether the access is legitimate. For example, if the security level of the data demander is not higher than the security level of the data resource, the access operation could be performed; if the security level of the data demander is not lower than the security level of the data resource, the system could perform operations such as rewriting or deleting existing data and sending control commands.

Recently, with the development of diverse credit databases, multiply data access, and different types of data inquiry requirements, the original DAC and MAC gradually cannot fully meet the practical needs of system with multiuser or multi-node in big data credit investigation, so the role-based access control mechanism (RBAC) and attribute-based access (ABAC) emerged. The core idea of RBAC is to simplify authorization management in various environments by mapping users to different roles such as data holders and data demanders, and using roles as the basis to achieve user access rights management.

ABAC takes attributes (a collection of four elements: user data, operation mode, data access mode, and data resource information) as the minimum authorization unit, and it could realize a more flexible authorization mechanism considering the complexity and variability of the environment, which can solve the problem of dynamic expansion of large-scale users and fine-grained access control in complex systems.

2.3. Data Processing and Analysis. Another core of modern credit investigation is the reliable and secure measurement of massive amounts of data and hyper-credit data in credit assessment and credit management, and computational algorithm is the core of data processing and analysis technology.

As mentioned before, the large amount of data collected or crawled puts enormous pressure on storage, and personalized computing tasks put higher demands on data scheduling, caching, and computing, therefore online processing systems are created, and usually they could be divided into two categories depending on the specific type of tasks: online analytical processing (OLAP) and online transaction processing (OLTP).

2.3.1. Online Analytical Processing (OLAP). OLAP can organize and display data in a variety of formats to suit the different needs of different customers [14]. Specifically, it can provide online real-time complex analysis of diverse and high-dimensional data for different users to facilitate their needs. It can mainly realize the functions of data slicing and dicing, data pivot and drill, in this sense, OLAP is also called decision-support system (DSS).
2.3.2. Online Transaction Processing (OLTP). OLTP, on the contrary, generally provides only real-time query access and command response services, supports fast response to transactions and large concurrency, reflects the state of the business, and does not do complex calculations such as analysis (see Table 2).

2.4. Privacy Computing Architecture. The purpose of data encryption is to prevent data from being interpreted or used beyond its scope during transmission and computation. The new computing architecture is supported by secure multi-party computation (MPC) [15], Federated learning (FL) could further guarantee the absolute control of data by credit data subjects, to reduce the risk of eavesdropping during data transmission, and ensure the validity and correctness of joint computation.

2.4.1. Secure Multiparty Computation. MPC aims to solve the problem of smooth performance of joint computation tasks in multiparty models without divulging information about any party. It allows multiple data owners to perform collaborative computations without trusting each other and output the computation results, ensuring that no party has access to any information other than the computation results to which it is entitled.

MPC architecture usually has a third-party collaborator or control platform, and the computation tasks are often performed locally by the data holder, the data holder must update their own models based on the gradients/losses aggregated by the third party, and the need to encrypt the transmitted data using cryptographic techniques when interacting with each other. The well-known techniques include Garbled circuit [16], secret sharing (Shamir [17] and Blackley [18] independently proposed the concept based on interpolating polynomials and projective geometry), and homomorphic encryption (a new cryptographic tool that supports arbitrary function operations on encrypted messages, and the result obtained after decryption is the same as the result of performing the corresponding operation on the plaintext [19]).

2.4.2. Trusted Execution Environment. Trusted execution environment (TEE) is a standalone execution environment that runs in parallel with the normal operating system. The TEE offers safe execution of authorized security software (trusted applications), which enables it to provide end-to-end security. In TEE, the security application is executed by an isolated, scalable execution environment independent of the mobile device operating system, providing a secure zone for the device to protect the execution of authenticated code, system integrity and data access rights.

There are many differences between MPC and TEE. Generally speaking, MPC computation occurs in each independently distributed data source, and the computation is performed according to the computation tasks issued by the unified control platform, while each computing unit interacts with each other through a secure protocol (homomorphic encryption) [20] to finally complete their respective computing tasks; while TEE technology still keeps the computation tasks in the unified platform, while acquiring the encrypted data of each unit and transmitting it to the secure TEE container; if each participant passes the security authentication, the CPU calculates with the data decryption operation and completes the calculation task in the TEE platform with the decrypted data.

2.4.3. Federated Learning. Federated Learning (FL) is an emerging fundamental AI technology [21], first proposed by Google in 2016. Its design goal is to carry out efficient machine learning among multiple participants or multiple computing nodes under the framework of legal compliance, and to guarantee information security, endpoint data, and user's privacy at the same time of data exchange. There are actually various implementations of Federated Learning, such as federated computing, shared intelligence, knowledge federation, and federated intelligence.

TEE performs data computation without data interaction between individual data holders, but only between the third-party collaborators/trusted computing centres and the data holders. Moreover, TEE is based on third-party hardware, and data are shared in a trusted execution environment created by the hardware, which is more dependent on the hardware vendor.

Federated learning (FL), based on cryptographic techniques such as MPC, is actually a cryptographic distributed machine learning technique, which is a form of application of MPC in the field of AI and machine learning, allowing various participants to co-construct machine learning
models without appropriating the underlying data. For different data sets, federated learning is divided into horizontal federated learning and vertical federated learning. The cryptographic training process of vertical federated learning is essentially the same as the MPC architecture, but the task of vertical federated learning is AI learning, which is not limited to MPC, while vertical federated learning adds effect incentives to replace the "conscious honesty" requirement of MPC for participants, reducing the emergence of malicious attackers.

The architecture of horizontal federated learning is similar to that of TEE (Table 3), but the computational task of horizontal federated learning is carried out by the individual data holders, individual data holders calculate the training gradients locally, mask the selected gradients by means of encryption, differential privacy, or secret sharing techniques, and then send the masked results to a third-party collaborator server; the third-party collaborator is responsible for secure aggregation and will send the aggregated results to the individual participants, without needing to know any participant’s information.

3. Computing Algorithms Applied to Modern Credit Investigation

Algorithms applied to modern credit investigation could be divided into two categories: transactional algorithms and decision-support algorithms. Transactional algorithms mainly undertake data cleaning task, data preprocessing task, etc. However, decision-support algorithms are mainly responsible for credit service decisions and are classified into individual, sequential, and joint decisions depending on the policies and scenarios.

3.1. Decision-Support Algorithms

3.1.1. Individual Decision-Support Algorithm. The goal of individual decision-support algorithms is to determine the unique solution by supervised learning or unsupervised learning methods to support decision-making. The task of supervised learning is to use the training data to build learning so that the model can give the corresponding predicted output for any new data input. Logistic regression, KNN, Naïve Bayes, decision trees, support vector machine (SVM), and neural network (NN) are the most representative algorithms.

Firstly, the linear regression model expresses the objective as a combinatorial function of all data dimensions (attribute characteristics) and can be expressed as

\[ H(x) = b + w_1x_1 + w_2x_2 + \ldots + w_nx_n = W^T X + b. \]

(1)

The presence of covariance between data dimensions and the presence of outliers has a considerable impact on the results, and data preprocessing and normalization are essential. Moreover, the algorithm has no strategy for handling missing values, slow computing speed, and is suitable for small sample data sets in credit reference tasks. To avoid the shortcomings of linear models, a nonlinear function can be mapped from the input to the output, and logistic regression (LR) models use log odds with nonlinearized linear regression by

\[ \ln \frac{H(x)}{1 - H(x)} = W^T X + b. \]

(2)

Here, the classification likelihood is directly modelled and approximate probability predictions are obtained without prior assumptions about the data distribution, avoiding the problems associated with assuming inaccurate distributions. Wiginton [22] applied logistic regression to the study of personal credit assessment, and this study showed that the log-likelihood ratio function makes it possible to overcome the drawbacks of linear discriminant analysis methods by not requiring the credit variable data to obey a normal distribution.

Secondly, K-nearest neighbor algorithm (KNN) operates within a given dataset of a determined class of instances, for new instances, KNN classifies the training instances according to their K-nearest neighbors by majority voting [23].

Thirdly, Naïve Bayes, as a classification method based on Bayes’ theorem and the attribute conditional independence assumption, is widely used as well. For a given training dataset, the joint probability distribution of input and output is first learned based on the attribute conditional independence assumption, and then the output class with the highest posterior probability is found based on this distribution for new instances using Bayes’ theorem.

Fourthly, decision tree model is a multilayered, subtractive decision model, and it can automatically filter variables and thus is suitable for handling high-dimensional data. The function of the decision tree model is information entropy H:

<table>
<thead>
<tr>
<th>Type</th>
<th>OLTP</th>
<th>OLAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Users</td>
<td>Executives and lower-level managers</td>
<td>Decision makers and senior managers</td>
</tr>
<tr>
<td>Application</td>
<td>Impose mandatory requirements that the accessing users must or must not access a resource or perform action</td>
<td>High security</td>
</tr>
<tr>
<td>scenarios</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Design</td>
<td>Application-oriented</td>
<td>Subject-oriented</td>
</tr>
<tr>
<td>philosophy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data features</td>
<td>Up-to-date, detailed, discrete</td>
<td>Historical, multiple, high-dimensional, unified</td>
</tr>
<tr>
<td>Access volume</td>
<td>A dozen of items (read/write)</td>
<td>Millions of items (read)</td>
</tr>
<tr>
<td>Processing units</td>
<td>Simple tasks</td>
<td>Complex inquiries</td>
</tr>
</tbody>
</table>

Table 2: OLTP and OLAP comparison.
The amount of change in a classification, \( p \), is the probability of selecting that classification, and the optimization functions on which SVMs depend are:

\[ H = - \sum_{i=1}^{n} p(x_i) \log_2 p(x_i). \]  

In each classification, the dimension (attribute) that can cause the largest amount of change (decrease) in information entropy \( H \) is selected, \( n \) is the number of classifications, \( p(x_i) \) is the probability of selecting that classification, and the amount of change in \( H \) is the information gain. Lee et al. [24] explores the performance of CART and Mars algorithms on the task of credit scoring with higher accuracy than traditional discriminant analysis, logistic regression, neural networks, and support vector machine (SVM) and other methods.

Fifthly, support vector machine (SVM) models can be considered as linear or nonlinear models depending on the kernel method. The optimization functions on which SVMs depend are:

\[ \min L = \frac{1}{2} ||w||^2. \]

That is, to find the classification hyperplane (line) when the sum of the distances from any sample point to the classification line (face) is minimized, that is, the classification face with the maximum geometric interval. SVM only considers the nearest points on both sides of the classification face, does not depend on the data distribution, and is not affected by other distant points, so the model has better generalization ability and theoretically better sample differentiation ability.

Bellotti and Crook [25] compared SVM with traditional methods on a large credit card database and found that SVM has some advantages for this problem.

And finally, there are neural network (NN) algorithms represented by BP neural network, CNN, RNN, etc. These algorithms generally require high data preprocessing, no missing values, and all attributes are numerical attributes. Thus, for simply processed data samples, the classification results are generally worse than other nonlinear algorithms.

CNNs also require data to meet specific dimensional criteria, which is because the data processed by CNNs are generally feature matrices, so CNNs are more often applied in the image domain; while RNN models are mainly applied in natural language processing, which requires data to have a temporal dimension. Jensen [26] used BP in his credit evaluation study neural network and obtained an accuracy of nearly 80% on the prediction of customer classification problem.

### 3.1.2. Sequential Decision-Making Algorithm

Sequential decision-making algorithms include Q-learning (QL), Markov decision process (MDP), and voting, and they are typical cumulative models of individual decisions in the temporal dimension through continuous self-evaluation and improvement during the learning process. Sequential Decision is a multistage decision-making method, also known as dynamic decision-making method, in which decisions are arranged in time order to obtain various decision strategies in sequence; that is, there is a sequential difference in time. Each stage of multistage decision-making requires a decision to be made, and the selection of each stage decision is not arbitrarily decided, it depends on the current state faced and does not give influence to the later development, so that the whole process is optimal. When the decisions at each stage are determined, a sequence of decisions or strategies for the problem is formed, called the decision set.

Markov decision process is a combination of Markov process theory and deterministic dynamic programming, and it contains a set of interactive objects: agent and environment. The agent in the MDP that perceives the state of the external environment to make decisions, acts on the environment and adjusts its decisions through feedback from the environment. The environment refers to the set of all things external to the agent in the MDP model, whose state is changed by the actions of the agent, and the above changes can be fully or partially perceived by the agent. The
environment may give back to the agent the corresponding reward after each decision.

The agent perceives the environment, \( s_i \) perceives the environment and performs actions according to a policy \( a_i \) to implement actions, where \( i \in [1,T] \). The environment is affected by the new action and thus enters a new state.

\[
p(s_{i+1}|s_i, a_i, \ldots, s_1, a_1) = p(s_{i+1}|s_i, a_i),
\]

and gives a reward \( r_i \) back to the agent; thus causing the agent to reach the new state of

\[
A_T = \{s_1, a_1, s_2, a_2, \ldots, s_T, a_T\}.
\]

In 2018, Yang and Zhou [27] used Markov decision process to calculate the credit line of enterprise group based on the historical data of commercial banks which had a great significance for the credit risk evaluation of commercial banks.

Reinforcement learning (RL) also known as reactive learning, evaluative learning, enables an agent to learn by “trial and error,” guiding behavior through rewards obtained by interacting with the environment, with the goal of maximizing the reward of the agent. Reinforcement learning is represented by Q-learning (QL) [28].

In 2019, Herasymovych et al. [29] developed a dynamic reinforcement learning system based on real data from international consumer credit companies and show that an adaptive reinforcement learning system can be effectively used to optimize the credit scoring acceptance threshold in a dynamic consumer credit environment.

3.1.3. Joint Decision-Making Model. Joint decision-making is a hybrid strategy that combines the optimal decision model through individual and sequential decisions based on different algorithms, including Boosting, Bagging.

Boosting is a family of algorithms that can boost weak learners to strong learners. Bagging is a voting algorithm that first generates different training data sets using bootstrap, then gets multiple base classifiers based on these training data sets separately, and finally gets a relatively better prediction model by combining the classification results of the base classifiers. random forest (RF) [30] algorithm is based on bagging method, which means that multiple tree classifiers are built without any connection between classifiers, and the final result is given by voting and other algorithms to reduce the variance of prediction results of individual tree models; while GBDT (gradient boosting decision tree) is a boosting method based on boosting, by building a series of single decision tree weak classifiers to gradually enhance the effect, each new decision tree is built based on the previous decision tree, and the process of gradual iteration can avoid the prediction bias caused by a single decision tree classifier.

In 2018, Wang and Liao [31] showed that the support vector machine-based bagging integration algorithm has a low misclassification rate and can be better adapted to SME credit assessment. Wang and Yao [32] proposed a GBDT-based personal credit assessment method. The comparison experiments on two UCI public credit audit data sets showed that GBDT significantly outperformed SVM as well as LR for credit assessment, with better stability and generalizability.

As a conclusion of this section, strengths and defects of the abovementioned decision-support algorithms applied in credit investigation could be observed in Table 4.

3.2. Data Cleaning Algorithm. While there are many ways to construct a computational model, the most important thing is to find an algorithm (or a combination of algorithms) that provides the best differentiation, more stable computational structure output, and more accurate evaluation of the target data. This requires a series of descriptive analysis methods including principal component analysis (PCA), cluster analysis, and graph processing (relational computation method).

3.2.1. Principal Component Analysis. PCA is one of the most commonly used dimensionality reduction methods; for sample points in the orthogonal attribute space, a hyperplane (a high-dimensional generalization of a straight line) is used to represent all samples appropriately, and this hyperplane needs to satisfy the nearest reconfigurability (the sample points are all close enough to this hyperplane) and maximum separability (the projections of the sample points on this hyperplane can be separated as much as possible).

Input: sample set \( D = \{x_1, x_2, \ldots, x_m\} \);
Dimension to low-dimensional space \( d' \).
Process:
(1) Centralize all samples: \( x_i \leftarrow x_i - 1/m \sum_{i=1}^m x_i \);
(2) Calculate the covariance matrix of the sample \( XX^T \);
(3) Make an eigenvalue decomposition of the covariance matrix \( XX^T \);
(4) Get the eigenvector \( w_1, w_2, \ldots, w_{d'} \) corresponding to the largest \( d' \) eigenvalues.

PCA can project the new samples into the low-dimensional space by simple vector subtraction and matrix vector multiplication by simply retaining the mean vector of \( W \) and samples. The low-dimensional space discards the smaller \( d-d' \) eigenvectors of the eigenvalues, allowing the samples to be sampled at a higher density and providing some degree of denosing effect.

3.2.2. Cluster Analysis. Cluster Analysis attempts to divide the samples in a data set into a number of usually disjoint subsets, each of which is called a “cluster.” Through this division, each cluster may correspond to some underlying concept (category). Clustering can be used either as a standalone process to find the inherent distribution structure of the data, or as a precursor to other learning tasks such as classification. Formally, suppose the sample set \( D \) contains \( m \) unlabeled samples, and the clustering algorithm divides the set of samples \( D \) is divided into \( k \) disjoint clusters \( \{C_i\} \), where \( l \neq l', C_i \cap C_l = \emptyset \), and \( D = \bigcup_{l=1}^k C_i \).

Discrete Dynamics in Nature and Society
<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Strengths</th>
<th>Defects</th>
</tr>
</thead>
</table>
| Logistic regression | 1. Fast speed and generally high interpretability;  
2. Simple and easy to understand; directly notice the weight of each feature;  
3. Convenient to update models and absorb new data easily.  
4. Dynamic adjustment of classification thresholds if a probabilistic framework is desired | 1. Linear and complex feature handling.  
2. Requiring normalization and more feature engineering methods |
| Decision trees | 1. Low computational complexity; output results are easy to understand and well interpreted.  
2. Insensitive to missing intermediate values, being able to handle uncorrelated nonlinear feature data.  
3. High classification accuracy with simple generated models  
4. Good robustness to noisy data | 1. Limited nonlinear expressiveness and weak classification capability for single trees;  
2. Having difficulties in dealing with continuous variables;  
3. Prone to overfitting |
| Naïve bayes | 1. Solid mathematical basis; stable classificational efficiency and easy to interpret.  
2. Few parameters to be estimated and less sensitive to missing data  
3. Not requiring a complex iterative solution framework and being suitable for data sets of enormous size.  
4. Quite complex surfaces can be fitted | 1. Sensitivity to the way in which the input data are presented, and easy underfitting.  
2. Independence assumptions between attributes often do not hold  
3. Existence of error rate in classification decisions |
| SVM | 1. Suitability for small sample data sets.  
2. Possibility of improving generalization performance.  
3. Ability to solve high-dimensional problems.  
2. No universal solution for non-linear problems.  
3. High computational complexity and time-consuming computation of large-scale data |
| KNN | 1. Suitability for modeling more complex non-linear relationships  
2. Low cost of retraining.  
3. Applicability to sample sets to be classified with more crossover or overlap of class domains.  
4. Suitability to deal with multimodal classification and multi-label classification problems | 1. Calculation is relatively large and slow, and requires editing of sample points.  
2. Sample imbalance can lead to large prediction bias.  
3. Tend to be sensitive to dimensional disasters.  
4. Category scoring is not specific |
| Neural network | 1. Suitable for modeling more complex non-linear relationships  
2. Data structure in neural networks is essentially very flexible for learning relationships for any type of characteristic variables  
3. Studies have shown that providing the network with more training data can improve the network performance | 1. Training of the network can be very challenging and computationally intensive, requiring a large number of parameters.  
2. The model is more complex and the results are difficult to interpret.  
3. The high performance of the network requires a large amount of data to achieve.  
4. Nonlinear, easy to overfit and implied noise |
| MDP | 1. Quantitative analysis of uncertainty events in complex systems is possible  
2. Ability to calculate the probability of systems with maintenance capability and multiple degradation states | 1. Probability of a change of state needs to be assumed to be fixed.  
2. Need to know the various probabilities of state changes.  
3. The knowledge about matrix operations is relatively complicated |
| Q-learning | 1. Few parameters required.  
2. No model of the environment is required.  
3. Without being limited to episodes task | 1. Maximization deviation problem  
2. Slow update speed may occur  
3. Foreseeability is not strong |
| Bagging | 1. Training can be highly parallelized  
2. The decision tree nodes can be randomly selected to divide the features, and the model can still be trained efficiently when the sample features are of high dimension.  
3. Random sampling, small variance of the trained model, and strong generalization ability  
4. Insensitive to partial feature | 1. High classification accuracy.  
2. Various methods can be used to construct weak classifiers without doing feature screening  
3. Without overfitting problems |
| Boosting | 1. High classification accuracy  
2. Various methods can be used to construct weak classifiers without doing feature screening  
3. Without overfitting problems | 1. Execution results depend on the selection of weak classifiers, and the number of iterations and the number of weak classifiers are not easy to set  
2. Long training time  
3. Susceptible to noise interference and data imbalance leading to classification accuracy degradation |
In 2013, Gao and Cheng [33] proposed a new dynamic credit scoring model based on clustering integration to address the problem of inability to dynamically predict customer credit as well as population drift in customer credit scoring. It has been proven that the dynamic model not only has a lower misclassification rate than the static model, but also is able to predict bad credit lenders as early as possible.

3.2.3. Graph Processing. Graph is an important data structure that consists of nodes \( V \) (or called vertices, i.e., individuals), with edges \( E \) (i.e., connections between individuals), which we generally denote as \( G(V, E) \). Relationships such as social networks and commodity recommendations expressed by graph data are a good representation of the correlation between data related to creditworthiness.

In the financial model, there are a large number of different types of inter-entity relationships, some of which are relatively static, i.e., equity relationships between enterprises and kinship relationships between individual customers, while others are constantly changing dynamically, such as transfer relationships, trade relationships. In the credit investigation industry, in the process of data analysis and mining for a certain financial business scenario, the differences and differences between individuals (i.e., enterprises, individuals, and accounts) are usually analyzed from the perspective of the individuals themselves, and rarely from the perspective of the correlation relationships between individuals. It is in this regard that graph computing and graph-based cognitive analytics make up for the shortcomings of traditional analytic techniques by analyzing problems from the perspective of the economic behavioral relationships between entities and entities.

For example, commercial banks are facing the credit risk problem of escalating nonperforming rate of enterprise customers’ loans. In order to improve the prediction ability of banks on the transmission of enterprise nonperforming risks, the use of graph computing technology can completely portray the social and economic relations between enterprise customers and between enterprises and natural persons, and build an all-round risk association network to realize the dynamic and complete presentation of risk elements. When credit risk occurs in a certain enterprise in the network, cross-correlation analysis can be conducted using information such as customer portraits and economic behavior trajectories of risky customers in the risk correlation network to predict the transmission path and spread of risk, helping banks to take effective measures to block the source of risk contagion and isolate risk, thus improving the reliability and accuracy of risk management.

3.3. Data Preprocessing Algorithm. Modern computing algorithms generally rely on large training data sets to learn the behavioral characteristics of good and bad users, respectively, but data imbalance means that the algorithm will not be able to capture the key information about defaults.

The data imbalance problem involves two main aspects. On the one hand, the amount of data will affect the solution of the data imbalance problem, generally divided into large data sets or small data sets; on the other hand, the degree of data imbalance, here divided into three degrees: slight (the difference between the number of positive and negative samples is within one order of magnitude), moderate (the difference between the number of positive and negative samples is within two orders of magnitude), and severe (the difference between the number of positive and negative samples is more than two orders of magnitude).

When the acquired credit data has imbalance data, the processing methods can be selected according to the degree of imbalance and the quantity of data. The most commonly used processing methods can be divided into three layers: data layer, algorithm layer, and logic layer.

3.3.1. Data Layer. In the data layer, the idea of the solution is to solve the data imbalance problem by processing and altering the original data set so that the proportion of data in each category is maintained at a reasonable ratio. It is generally divided into two ways: data sampling and data synthesis.

Data sampling is divided into oversampling and under sampling. Oversampling maintains a reasonable proportion of data in each category by repeatedly replicating data from a small number of data categories, but models trained based on such sampling methods are prone to overfitting and can incorporate slight random perturbations each time new data are generated. This method is generally used to solve small data sets with minor data imbalance problems.

Under sampling is to maintain a reasonable proportion of data in each category by filtering out some data from most data classes, but this sampling method may lose key data, but it can be solved by multiple random under sampling. This method is commonly used to solve large data sets with moderate data imbalance problems.

Data synthesis is to generate more new samples by using the feature similarity of existing samples, this method is generally used to solve small data sets with moderate degree of data imbalance. SMOTE algorithm (synthetic minority oversampling technique) is one of the representatives, the basic idea is to analyze the minority-class samples, and based on the minority-class samples data to synthesize new samples to add to the dataset, the algorithm process is as follows.

(1) For each sample \( x \) in the minority class, calculate its distance to all samples in the minority-class sample set by Euclidean distance to obtain its \( k \)-nearest neighbors.

(2) Set the sampling multiplier \( N \) according to the sample imbalance ratio, for each minority-class sample \( x \), randomly select \( N \) samples from its \( k \)-nearest neighbors \( \bar{X} \). Assume that the selected nearest neighbors are \( \bar{X}_n \) set.

(3) For each of the nearest neighbors \( \bar{x} \) in the set \( \bar{X}_n \), a new sample is constructed from the original sample according to the formula, respectively,

\[
x_{\text{new}} = x + \text{rand}(0, 1) * (\bar{x} - x).
\]
3.3.2. Algorithm Layer. In the algorithm layer, the solution is to appropriately use a cost function (cost) or an incentive function (incentive), to increase the calculated weights of the lesser class of samples, which is actually used to reset the new data distribution without altering the operation on the dataset itself. This method is generally applicable to large data sets with slight data imbalance.

Commonly used algorithms are extreme gradient boosting (i.e., eXtreme Gradient Boosting, XGBoost), adaptive boosting (AdaBoost), etc. They all use the idea of Boosting to adjust the training sample distribution based on its performance using the initial training out base learner, which increases the weights of a few samples and decreases the weights of most samples along the way when applicable to solve data imbalance, thus mitigating the effect of imbalanced data sets on the model effectiveness.

Song et al. [34] used XGBoost (eXtreme gradient boosting) algorithm for the identification of abnormal customers in power grids, and their comparison used XGBoost classifier, KNN (k-nearest neighbor) classifier, BP (back-propagation) neural network classifier and random forest classifier for abnormal customer identification under balanced and unbalanced sample sets, respectively, and the experimental results show that XGBoost has higher recognition rate runs faster and has obvious performance improvement under unbalanced data sets. Guo et al. [35] proposed a novel AdaBoost improvement algorithm by processing the weights and labels of classification difficult samples in large class processing, so that the classifier can obtain better check accuracy and check all rates at the same time, and its can effectively improve the classification performance on unbalanced data sets.

3.3.3. Logic Layer. The logic layer converts the original classification and discrimination problem into a one-class classifier or anomaly detection problem, etc. The focus of these methods is not on capturing the differences between classes, but on modeling one of the classes, which is suitable for large/small data sets with heavy degree of data imbalance problems.

When there is a heavy imbalance of positive and negative sample ratios, relying on simple sampling and data synthesis is no longer a good solution to the problem. Because although the above method solves the problem of positive and negative sample ratio of training data, it seriously deviates from the real distribution of the original data, which will lead to the model training result does not really reflect the actual situation and will have a great deviation.

One-class SVM can be a suitable solution at this point, the basic idea is as follows: use Gaussian kernel function to map the sample space to the kernel space, find one sphere in the kernel space that can contain the data, and when discriminating, if the test data lies in this high-dimensional sphere, it will be classified as majority class, otherwise it will be classified as minority class [36]. Because the model may go to match these outliers, the training set of One-class SV should not be mixed with points from the minority class, so the number of minority-class data will not affect how good the One-class model is, and is well suited to address data imbalance when the ratio of positive to negative samples is heavily imbalanced.

3.4. Evaluation Techniques for Algorithm Applications. Data cleaning algorithms, descriptive analysis algorithms, and preprocessing algorithms are not directly related to decision support as predecessor algorithms, so they cannot be proven or falsified by observation or simulation, right now there is no mature method or system to evaluate these algorithms, and generally these algorithms are chosen according to operator preference and the easiness of implementation. In contrast, for relational computing problems such as social networks, the first thing to consider is the coupling and coordination analysis of hardware, environment resources, and application requirements, i.e., high access to storage ratio, poor data localization, diverse types and operations, large-scale and irregular structure.

Usually, different algorithms can match different application scenarios, while multiple methods or combinations can be chosen for the same scenario. For a specific credit business, the first and foremost consideration is the accuracy of the algorithm results, followed by the generalization ability of the algorithm model (the ability to differentiate between different samples) and the stability of the algorithm output results. The factors that affect these results are not only the algorithm itself but also the model parameters on which the algorithm relies.

When used as a binary classification problem, the output (prediction) of each algorithmic model can be divided into four subsets: true positive (TP), false positive (FP), true negative (TN), and false negative (FN), based on the correspondence between the true labeling of the sample and the prediction of the model. And the four subsets can form a confusion matrix as shown in Table 5.

For classification problems, accuracy and error rate are the most commonly used discriminatory criteria, and the accuracy and error rate are calculated by the proportion of correctly classified results and incorrectly classified results to the overall data, respectively.

\[
\text{accuracy} = \frac{TP + TN}{TP + FP + FN + TN},
\]

\[
\text{errorRate} = \frac{FP + FN}{TP + FP + FN + TN} = 1 - \text{accuracy}.
\]

When the task requirements are extended from the common classification problem to the credit assessment domain, the problem of data imbalance arises and the accuracy and error rates cannot meet the proper credit task requirements and need to be replaced by accuracy (precision), recall (recall) and F1 metrics. The accuracy rate indicates the proportion of correctly predicted positive samples to all predicted positive samples, the recall rate indicates the proportion of correctly predicted positive samples to all actual positive samples, and the F1 metric combines the accuracy rate and recall rate. From the
In modern credit investigation, it could be observed that the application of modern computing brings prosperity to the credit reference industry, but also brings many risks. This section discusses the current bottlenecks and provides corresponding countermeasures in the background of credit investigation practices in China.

5. Further Discussion on Legal Governance

In modern credit investigation, it could be observed that the application of modern computing brings prosperity to the credit reference industry, but also brings many risks. This section discusses the current bottlenecks and provides corresponding countermeasures in the background of credit investigation practices in China.

5.1. Data Quality Control and Process Management. Credit investigation industry requires extreme attention to the data quality, because data collection and evaluation may incorporate a high degree of inaccurate information [38], and fake and false information (i.e., “click farming” on e-commerce platforms, “credit hype” to subject reputation, and “cash out” in credit card fraud) will greatly undermines confidence of data-driven decision-making in credit investigation.

For this reason, the Chinese government and Chinese central bank require that Internet platforms who have cooperation (i.e., loan assistance and joint lending) with financial institutions must describe in detail the data collection, data processing, processing data subjects, data flow as well as funds flow in each business line. Additionally, according to the Notice issued in 2021 by Credit Bureau of the People’s Bank of China, the credit Internet platforms will not be allowed to provide (1) data submitted by individuals, (2) data generated within the network of platform, or (3) data obtained from external sources—directly to financial institutions, in order to strengthen the process management of data flow of credit bureaus. And finally, APIs and Web crawlers must not collect data from illegal channels in order to ensure the legality, legitimacy, and necessity of data collecting activities required by law.

5.2. Privacy Protection. The individual’s credit data usually include financial account data, personal identification data, and precise location data and thus are often considered as sensitive data of which the leakage may easily damage the data subject’s reputation or physical and mental health. In the process of modern credit investigation, attention should be paid not only to the creditworthiness and authenticity of data, but also to the privacy protection of data subject.

After the enforcement of Chinese Personal Information Protection Law in 2021, data collection in credit investigation industry must meet the following requirements: firstly, data which may reveal racial or ethnic origin, religious or philosophical beliefs, genetic data, biometric data, medical data, etc. are prohibited to be collected. Secondly, data on individuals’ income, deposits, securities, commercial insurance, real estate and taxation, precise location data, etc. must be collected in very restricted manner guaranteed by opt-in mechanism. Thirdly, the collection of ordinary personal data must be guaranteed by the so-called “lawful use
mechanism” or consent of the data subjects. And fourthly, public and open data could be collected in the open field but could not infringe on other’s individual rights. In this sense, it could be observed that the so-called privacy computing technologies (i.e., differential privacy, homomorphic encryption, multiparty secure computing, zero-knowledge proof, trusted execution environment, and federal learning) will be further applied in the credit investigation industry in order to enhance the big data analysis capabilities and fulfill the requirement of privacy protection law.

5.3. Algorithmic Governance. It is known that the use of automated decision-making models may lead to the so-called “algorithmic bias,” and there are two main types of causes of algorithm bias, namely, “data bias” and “algorithm design bias”. The former mainly includes biases arising from the data input layer and the training data layer, for example, inaccurate data collection methods may lead to off-truth data descriptions and structural biases (i.e., gender). The latter often arise from algorithm programming bias in data labeling and data classification.

And there are three main countermeasures for algorithm bias: firstly, at the technical level, developing algorithms which are able to regulate algorithms in the approach of “governance-by-design.” Secondly, at the risk control level, introducing an ex-ante assessment system, strengthening the interpretability of algorithms and the transparency of decision-making operations, and introducing third-party assessment systems. And thirdly, at the compliance level, allowing the ex-ante and ex-post remedies, giving data subjects the right of rejection in the automatic decision-making of credit operations, and strengthening algorithmic accountability mechanisms. For instance, in March 1 of 2022, China has launched new Regulation on the Administration of Algorithmic Recommendation of Internet Information Services, and the Chinese credit investigation platforms are encouraged to develop more comprehensive and transparent algorithmic strategies in order to optimize interpretability and fairness of algorithms.

6. Conclusion

In summary, modern credit investigations supported by computational technologies and algorithms have many advantages compared to traditional credit investigation. The process of modern credit investigation can be unified into a generic and integrated process which brings higher demands on legal governance such as data quality control, privacy protection, and algorithmic governance. In the big data era, modern computing will ultimately drive the transformation of credit investigation into a more intelligent system and legal compliance will undoubtedly further unlock the future development potential of the credit investigation industry.

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 Tsinghua University National Social Science Foundation of China: Research on Rule of Law of Internet Economy (No. 18ZD149).
References


