Hindawi
Journal of Mathematics
Volume 2022, Article ID 9135117, 10 pages
https://doi.org/10.1155/2022/9135117

Research Article

Financial Data Anomaly Detection Method Based on Decision Tree and Random Forest Algorithm

Qingyang Zhang

Business & Tourism Institute, Hangzhou Vocational and Technical College, Hangzhou, Zhejiang 310018, China

Correspondence should be addressed to Qingyang Zhang; 2007010025@hzvtc.edu.cn

Received 24 December 2021; Revised 18 March 2022; Accepted 21 March 2022; Published 16 April 2022

Academic Editor: Miaochao Chen

The fast-developing computer network not only brings convenience to people but also brings security problems to people due to the appearance of various abnormal flows. However, various current detection systems for abnormal network flows have more or less flaws, such as the most common intrusion detection system (IDS). Due to the lack of self-learning capabilities of market-oriented IDS, developers and maintenance personnel have to update the virus database of the system in real time to make the system work normally. With the emergence of machine learning and data mining in recent years, new ideas and methods have emerged in the detection of abnormal network flows. In this paper, the random forest algorithm is introduced into the detection of abnormal samples, and the concept of abnormal point scale is proposed to measure the abnormal degree of the sample based on the similarity of the samples, and the abnormal samples are screened out according to this scale. Simulation experiments show that compared with the other two distance-based abnormal sample detection techniques, the random forest-based abnormal sample detection has greater advantages than the other two methods in terms of improving the accuracy of the model and reducing the computing time.

1. Introduction

With the listing of the Science and Technology Innovation Board, the registered listed company system has made China's financial and economic system more open and diversified. Investors are paying close attention to how to judge the operating level and ability of listed companies and to ensure the compliance of public financial statements of listed companies, which also directly affects the vital interests of small and medium shareholders. In the process of business operations, both systemic risks and risks caused by their own operating conditions may bring many problems. In order to maintain the stock price of their own companies, listed companies have a certain incentive to hide risks. Through the supervision of the China Securities Regulatory Commission, it has been researched and discovered. As an important research direction of artificial intelligence, machine learning has naturally become a very important candidate for prediction methods on the issue of predicting financial irregularities. Machine learning and deep learning include many kinds of learning models, such as supervised learning, unsupervised learning, semisupervised learning, and so on. These different research categories differ in the degree of labeling of data samples. Supervised learning needs to make early judgments on the sample (such as whether the academic performance is good), unsupervised learning is automatically judged through the classification of data, and semisupervised learning is between the two and requires a certain degree of data labeling. This paper uses the decision tree model and random forest model in supervised learning to predict and judge the financial violations of listed companies. The reason is that decision trees and random forest algorithms have certain advantages in the processing of financial data [1–10].

How to identify abnormal samples in a dataset has become a hot topic in data mining research in recent years, for example, mining abnormalities in a large number of bank enterprise credit data, removing noise when establishing a credit evaluation model, and improving the accuracy of the model. In addition, data mining technology can also detect

possible abnormal credit fraud. This paper introduces a new learning algorithm that can better tolerate noise—random forest, introduces a method to measure the similarity of samples, combines the similarity of samples, and proposes the concept of abnormal point scale to quantify the abnormality of samples and filter out abnormal samples based on this scale.

The technology that has been commercialized in the network abnormal flow detection method is intrusion detection. The earliest intrusion detection technology was proposed by James P. Anderson in 1980. The idea is to detect computer attacks by recording and auditing. The general model of intrusion detection was proposed by Dorothy Denning in 1987. The early intrusion detection models were based on the host. It was not until 1990 that L. T. Heberlein proposed a network-based intrusion detection model. Later, more commercial intrusion detection systems continued to appear, such as Cisco's NetRanger, ISS's RealSecure, and Snort, but due to the limitations of technological development, the products were not mature [11–15].

After decades of development, intrusion detection technology has achieved rapid development. At present, the most popular intrusion detection model is network intrusion detection based on pattern matching. This model has a characteristic rule base of the attack flow. When the intercepted data flow has the rule characteristics defined in the rule base, the data flow is considered to be attack flow. The accuracy of this detection method is extremely high, but it cannot automatically discover new attack flows, so the false negative rate is relatively high, and the system's feature rule base needs to be continuously updated to work normally.

In recent years, there have been many other research studies on network abnormal flow identification, which can be summarized into the following three types: abnormal flow detection methods based on characteristics or behaviors, abnormal flow detection methods based on statistical theory, and machine learning and the abnormal flow monitoring method of data mining [16–20].

In the abnormal flow detection method based on characteristics or behavior, the detection system needs to store the message characteristics or behavior characteristics of the abnormal flow. The message characteristics include the load characteristics and encapsulation characteristics of the message, and the behavior characteristics include the transmission characteristics and the connection establishment characteristics. The detection system compares these characteristic attributes in the network traffic with the signature database to detect abnormalities in the network flow. This type of abnormal flow detection method can be divided into fuzzy matching and complete matching according to different matching rules. Fuzzy matching can identify the type of network flow by matching part of the characteristics of the network flow through the regular matching method, while complete matching requires the characteristics of the network flow data. The abnormal type of the flow can be determined only when it matches the features in the feature library completely. Some documents have implemented an abnormal traffic detection method based on traffic characteristics by extracting payload characteristic data of abnormal traffic and tested the system through DDoS attack flows and network worms. There are also documents that realize a network worm detection algorithm by automatically extracting network worm flow characteristics and process flow behavior characteristics and speed up the detection speed of worms and improve the detection ability of slow worms by rating host data packets [21–25].

Different from the abnormal flow detection method based on characteristics or behavior, the abnormal flow detection method based on statistical theory does not need to know the characteristic attributes of the abnormal flow in advance. The method first obtains the flow data on the network flow time series, performs statistical analysis on the flow data to obtain a statistical result, and completes the detection of abnormal flows according to the statistical results. There have been many research studies on abnormal flow detection methods based on statistical theory at home and abroad. For example, someone introduced an adaptive filtering theory into abnormal flow detection and proposed an adaptive AR abnormal flow detection model; this model uses time series analysis technology, analyzes and models the collected historical data through the model, and finds a threshold that can distinguish abnormal flow from normal flow. Some people transform the network traffic in the time series to the frequency domain or wavelet domain and then realize the detection of abnormal flow based on the transformed spatial characteristics. In addition, Lakhinaa et al. decomposed the high-latitude structure space of the data flow between the source host and the target host by PCA to obtain 3 principal components and reconstruct the characteristics of the network flow through 3 composite variables [26–29].

Anomaly flow detection methods based on machine learning and data mining have been studied in depth with the development of machine learning and data mining, as shown in Figure 1. This kind of method can be divided into abnormal flow detection based on classification algorithm and abnormal flow detection based on aggregation algorithm. Typical representatives of abnormal flow detection based on classification algorithms include abnormal flow detection based on Naive Bayes classification algorithm, abnormal flow detection based on neural network, and abnormal flow detection based on decision tree classification algorithm. The main representative of abnormal flow detection based on clustering algorithm is the abnormal flow detection based on K-means algorithm. The abnormal flow detection based on the classification algorithm needs to use the sample data to train the algorithm in advance and then use the trained pattern to detect the network flow, while the aggregation algorithm is to mine the attribute feature set in the network flow and classify and aggregate the network flow. In this way, the aggregated flow of the abnormal characteristic pattern is obtained, and the abnormal behavior of the network traffic is detected according to the characteristic of the aggregated flow pattern. Someone used a decision tree classification algorithm to detect anomalous flows, which used cross entropy to represent anomalies in network traffic. Some people also use the classification ability

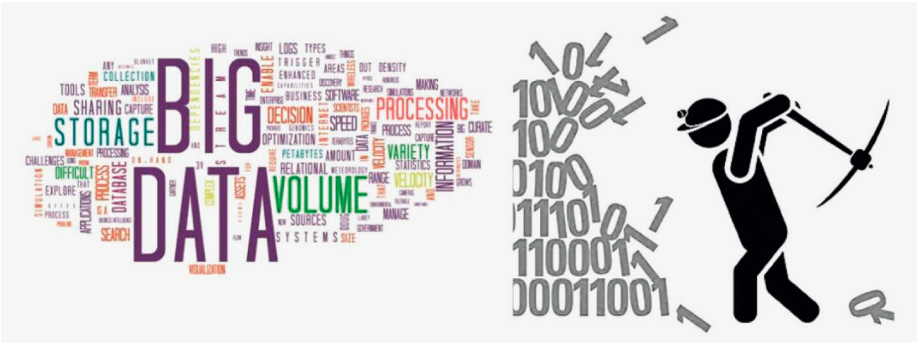

Figure 1: Data mining.

of support vector machines to convert the abnormal flow classification problem into an SVM classification decision-making problem and then judge the abnormal situation of the network flow according to the different distribution of attribute entropy values. Some people have also proposed an intrusion detection classification method based on Bayesian Yin-Yang learning and K-means clustering algorithm. The literature first uses Bayesian model selection ability to determine the number of clusters in the sample set and then uses Bayesian competition network. A large amount of original sample data is mapped to a small number of data nodes, and the output unit is used as the input of the K-means algorithm for cluster analysis, and finally the intrusion traffic is detected according to the cluster analysis result [30–33].

The random forest algorithm is an integration and improvement based on the decision tree algorithm and is the result of the integrated learning of the decision tree algorithm. This paper uses two machine learning algorithms, decision tree and random forest, to construct a financial statement analysis-based system for judging financial irregularities of listed companies, which can predict and analyze financial irregularities, thereby helping to discover more potential unknown financial violations and risk of violations and promote the steady development of the financial system.

2. Decision Tree Algorithm

When judging whether a company has disclosed violations, different data contents also have different degrees of importance. When judging whether a violation has occurred, government agencies such as the China Securities Regulatory Commission will also analyze their data based on certain priority conditions. This way of judgment has a certain degree in common with the business logic of the decision tree. Therefore, we decided to use the machine learning model of decision tree to judge and analyze corporate financial data. Decision tree is a typical machine learning algorithm, and its role is to solve the classification problem in supervised learning.

Decision tree classification algorithm is a kind of machine learning algorithm, which is a process of automatically mining a set of regular patterns from training samples that are also effective for data other than training samples. In addition to decision tree classification algorithms, machine learning algorithms also include Naive Bayes classification algorithms, classification algorithms based on support vector machines, neural network algorithms, K-means clustering algorithms, and fuzzy classification methods.

The decision tree constructed by the decision tree classification algorithm is a classification model. Each branch in the model represents the mapping relationship between an attribute of the object and a certain value or value type of the attribute. In the decision tree, each non-leaf node represents a judgment condition, each judgment condition corresponds to an object attribute, and each branch path represents an attribute value that satisfies the judgment condition. Each leaf node L in the tree represents a value set, and each value in the set satisfies each judgment condition in the path from the root node to the leaf node L. The decision tree is constructed from the root node. First, select appropriate attributes to divide the sample set into several subsets, and each subset forms a branch node, and then divide each branch node until the types of all samples in the node are the same or satisfy a certain end condition. The decision tree model is an algorithm model with a tree structure. Its idea is to simulate a structured thinking mode that people think in daily life. When we have the financial data of a company, we will first pay attention to whether there is any profit during the reporting period.

Decision tree construction generally includes the following two steps: (1) decision tree generation, that is, the process of using training sample sets to generate a decision tree; (2) decision tree pruning: after the decision tree is generated, it needs to be verified, corrected, and revised. The algorithm proposed in this paper uses the method of multidecision tree integration. Each decision tree is a weak decision tree, and there will be no overfitting phenomenon, so it does not involve the problem of decision tree pruning. The input sample set form of the decision tree construction algorithm is as follows:

$$I = \left\{\left(A_{00} \ldots, A_{0j} \ldots, A_{0n}, T_0\right) \ldots \left(A_{i0} \ldots, A_{ij} \ldots, A_{in}, T_i\right) \ldots\right\}, \quad (1)$$

where A_{ij} represents the value of the j-th attribute of the i-th sample in the set and T_i is the type mark of the i-th sample. The result of decision tree construction is a binary tree or multibranch tree. The binary tree is generally used for data collection whose attributes are all Boolean logic judgments. The general process of decision tree construction is shown in

Figure 2. As can be seen, the structure of the decision tree is similar to a real tree (Figure 2(a)), hence the name.

Different decision tree classification algorithms use different judgment conditions to select split attributes. The two most important judgment conditions are information gain and information gain rate. Split attribute selection based on information gain. Suppose the training sample set is S, and the attribute set is

$$P = \{p_1, \ldots p_i, \ldots p_m\}. \tag{2}$$

Then, the proportion of samples belonging to the j-th category in the sample dataset is

$$P(C_j) = \frac{|S_{ij}|}{|S|}. \tag{3}$$

At this time, the information entropy of the sample dataset S is

$$\text{Entropy}(S, p_i) = \sum_{j=1}^{n} - P(C_j)\log_2 P(C_j). \tag{4}$$

Suppose that in the sample dataset, the value range corresponding to the attribute p_i is v_i, and $S_i(v)$ represents the subset of samples whose attribute pi takes the value v. Then, the information gain of the sample set S to the attribute p_i is

$$\text{Gain}(S, p_i) = \text{Entropy}(S, p_i) - \sum_{v \in v_i} \frac{|S_i(v)|}{|S|}\text{Entropy}(S, p_i). \tag{5}$$

After the information gain of the sample set S is calculated, the split information of S on the attribute pi is calculated as

$$\text{SplitGain}(S, p_i) = - \sum_{v \in v_i} \frac{|S_i(v)|}{|S|}\log_2 \frac{|S_i(v)|}{|S|}. \tag{6}$$

Then, the information gain rate of the sample dataset S relative to the attribute p_i is

$$\text{GainRatio}(S, p_j) = \frac{\text{Gain}(S, p_j)}{\text{SplitInfo}(S, p_j)}. \tag{7}$$

3. Random Forest Algorithm

If a certain company has been profitable for many years and has been operating steadily, we can make a preliminary classification and attribute it to the candidate database of excellent companies. If the company is poorly operated and has past losses, then we should be more vigilant, continue to pay attention to other situations in the report, and continue to strengthen our comprehensive understanding of the company. In constructing a random forest model, it is particularly important to establish multiple independent and differentiated decision tree individuals. It is difficult to achieve a diversified decision tree model if you simply use raw data to judge the data. It is difficult to reflect the advantages of integrated learning. Therefore, when building a random forest model, we need to sample the data according to certain rules. Each decision tree uses a part of the total sample for training. This can reduce the number of repetitions of the training data and make each data basis of individual decision trees different, which increases the diversity of individual decision trees.

Random forest is a combined classifier method, and the basic classifier that constitutes random forest is decision tree, as shown in Figure 3, which also shows the combination of the random forest and decision tree. A decision tree is a hierarchical structure composed of nodes and directed edges. The tree contains three types of nodes: root nodes, internal nodes, and end nodes. The decision tree has only one root node, which is the entire training dataset. Each internal node in the tree is a splitting problem, which divides the samples arriving at that node into blocks according to a specific attribute. Each end point (also called a leaf node) is a data collection with a classification label. A path from the root node of the decision tree to the leaf node forms a discriminant rule. The decision tree algorithm uses a top-down greedy algorithm. Each internal node selects the attribute with the best classification result to divide the data arriving at that node into 2 or more blocks, and continue this process until the tree can be accurately classified. The core problem of the decision tree algorithm is to choose a better splitting attribute. There are many criteria for selecting split attributes, such as information gain, information gain ratio, Gini index, and so on. The decision tree algorithms corresponding to different attribute selection methods include ID3, C4.5, CART, and so on.

Integrated learning is a learning method that combines multiple modules to independently judge. In our model, if a decision tree is compared to an experienced professional analyst, then random forest is a large and professional team of analysts able to make more accurate judgments on decision results. The decision tree algorithm in this article is similar to the CART algorithm, and the choice of splitting attributes is based on the Gini index. Gini index is an impurity splitting method. It can be applied to fields of category, binary, continuous value, etc. The specific algorithm idea is as follows: assuming that the data sample set r at a certain node t contains records of k categories, then the Gini indicators are

$$\text{Gini}(t) = 1 - \sum_{j=1}^{k} [p(j|t)]^2, \tag{8}$$

where p is the probability from the category j to the node t. When the minimum Gini is 0, all samples at this node belong to the same category, which means that the maximum useful information can be obtained; when the category field is uniformly distributed, Gini(t) is the largest, and the useful information at this time is the smallest. If the set is divided into l parts, then the Gini index for this division is

$$\text{Gini}(T) = \sum_{i=1}^{l} \left(\frac{n_i}{n}\right)\text{Gini}(i). \tag{9}$$

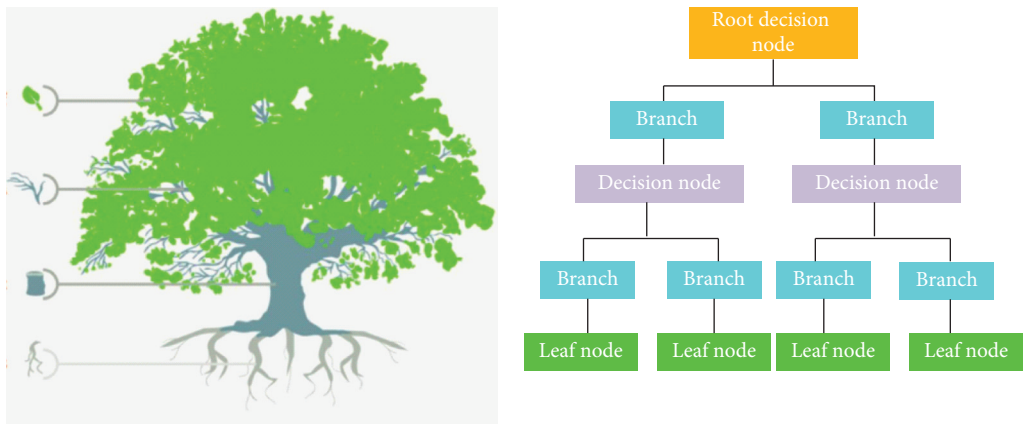

Figure 2: Decision tree.

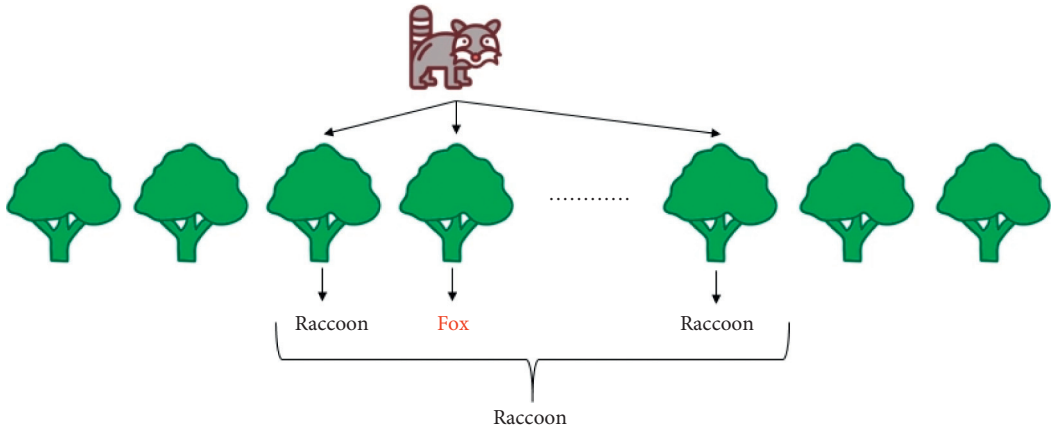

Figure 3: Random forest.

After understanding the profitability, we can continue to analyze and judge other aspects of the company, such as the main business, assets and liabilities, inventory turnover speed, and gross profit margin, to further understand the company's operations. Where l is the number of child nodes, n_i is the number of samples at the child node i, and n is the number of samples at the parent node. The basic idea of the Gini index is as follows: for each attribute, all possible segmentation methods must be traversed. If the minimum value can be provided, it will be selected as the criterion for splitting at this node; at this time, it will be split according to the corresponding attribute value, and create branches according to each attribute value; further divide the sample down until the stop condition is met, for example, the original leaf node belonging to the same type or the purity of the leaf node (that is, the frequency of the node containing a certain type of sample) meets a certain threshold range. The threshold is set in advance, and the division is stopped when the purity of the leaf node exceeds the threshold. This process is equivalent to pruning the column tree. The predicted value is shown in Figure 4.

Random forest repeats the above tree building process to build a combination of multiple decision trees. First, assume that there are M trees in the forest, that is, there are M decision tree classifiers, and the total number of samples of all training data is N. The bagging method is used to form a training set of a single decision tree by randomly sampling N samples from all training samples with replacement. Repeat this sampling process M times to obtain learning samples of M decision trees. In addition to the fact that learning samples of a single decision tree are randomly generated, random forest also adds randomness to the generation process of each tree. Suppose the sample has Q attributes in total, and $q < Q$ is usually given in advance (q is usually the square root of Q). When selecting the split attribute of each node, not all attributes are compared, but q is randomly selected from all attributes The attributes are compared, and the attributes with better classification results are selected for splitting. This can increase the degree of difference between each tree, thereby increasing the generalization error of the forest. There is no pruning during the construction of a single decision tree. After the forest is formed, for a new sample, each tree draws a corresponding classification conclusion, and finally all the trees pass a simple majority vote to determine the classification result. Compared with other combined classification techniques, when the number of

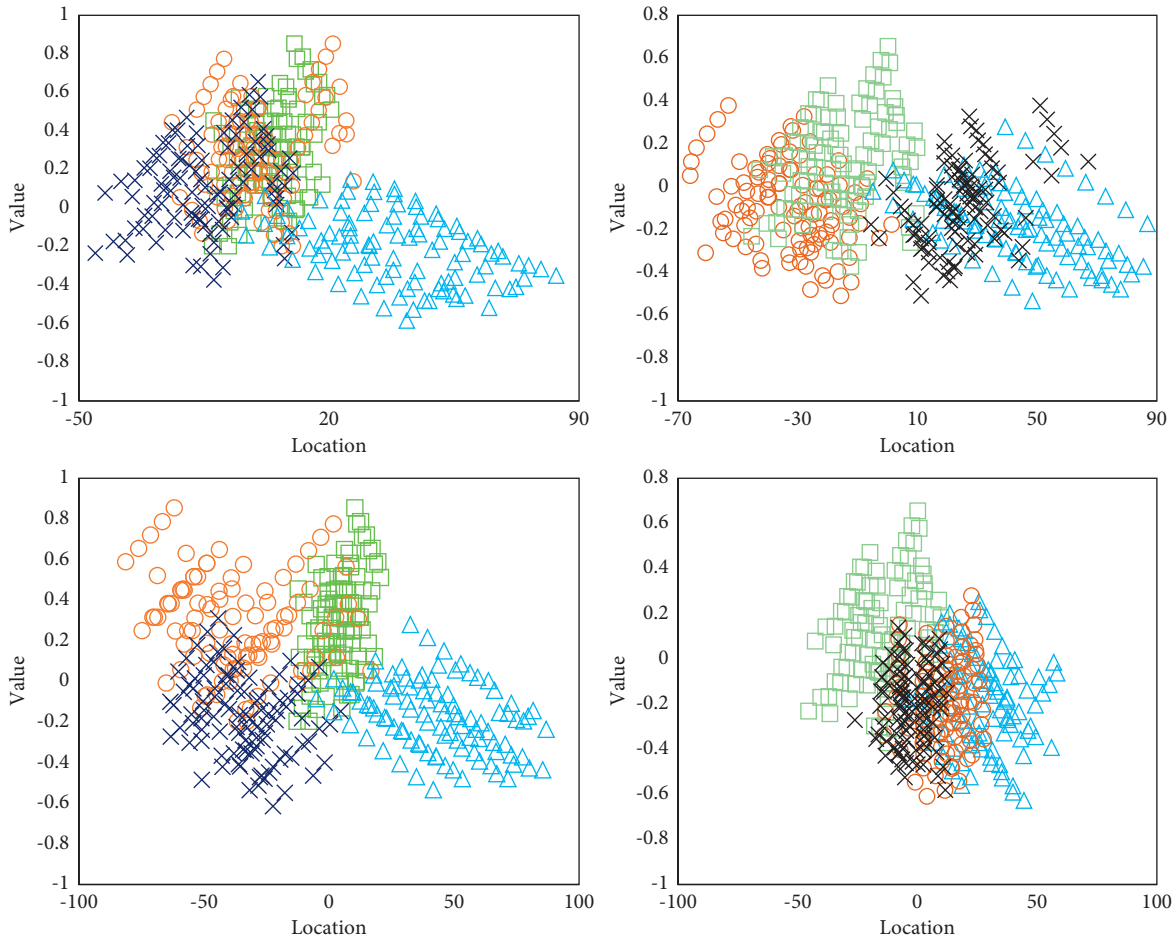

Figure 4: Predicted value.

trees is quite large, random forests are not prone to overfitting. It can be proved that the upper bound of its generalization error is less than

$$\frac{\overline{\rho}\left(1-s^2\right)}{s^2}, \tag{10}$$

where ρ is the average correlation coefficient between trees (representing the degree of correlation between the classification trees) and s is the classification efficiency of a single tree.

For the training set with the number of samples IV, first generate an $N \times N$ zero-element matrix Prox. Use the tree generated each time to discriminate all samples, and each sample will reach a certain leaf node of the tree; for any two samples n and k, if sample n and sample k appear in the same leaf of the tree at the node, add 1 to the corresponding nth row and kth column of the Prox matrix; repeat this process until all the hydrazine trees are built, and get the corresponding matrix. Divide each element in the matrix by the number of trees. Normalize processing to get the final Prox matrix, which is a symmetric matrix with diagonal elements of 1, and the element Prox(n, k) in the nth row and kth column can be defined as the sample of sample n and sample k. When the random forest was built, the Prox matrix was also obtained. It is not difficult to see that if there are a large number of samples in a certain category in the dataset, the rows corresponding to the samples in this category usually contain more elements close to 1, and those corresponding to the rows contain more elements close to zero. We have more reasons to believe that they are less similar to other samples. As a result, it naturally leads to a measure of the degree of abnormality of the sample—the scale of abnormal points.

For sample n, define its original abnormal point scale as

$$\text{rawom}(n) = \frac{n\text{sample}}{\overline{P}(n)}. \tag{11}$$

In the same class, if the $p(n)$ value of a sample is low, its rawom value will be large. For all samples of each class, calculate the mean and variance of the original outlier scales of all samples of this class and get the final outlier scale of each sample after standardization:

$$\text{outliermeasure}(n) = \frac{[\text{rawom}(n) - \overline{\text{rawom}}]}{\sigma}. \tag{12}$$

Random forest is a combination of multiple decision trees. The frequency of two samples at the same node of each tree can be used to measure the similarity between the two samples or the probability that the two samples belong to the same class. The above transformation avoids the numerical

difference caused by the large difference in the number of various types of samples, so as to facilitate the comparison of the abnormal point scales of the various types of samples. After the random forest is built, we get the scale of abnormal points of all samples according to the above calculation process. If the scale of abnormal points of a sample is larger, the similarity between this sample and other samples is small, and there may be anomalies. If the confidence interval is predicted, the samples whose abnormal point scale exceeds a certain threshold can be regarded as abnormal points. For example, each dataset used in the simulation experiment in this article is known to contain 5% of abnormal samples. Therefore, it is possible to sort the abnormal point scales of all samples and consider the largest top 5% of the samples as abnormal samples. The prediction is shown in Figure 5.

4. Simulation

Involving larger numerical calculation tasks, the algorithm is required to be as simple and fast as possible. However, the distance-based detection methods have the problems of long calculation time and high memory consumption. UC (University of California Irvine) machine learning database is a well-known machine learning verification database, which is widely used in the modeling and verification of learning algorithms. This article uses 6 commonly used standard datasets for simulation experiment comparison, and it is known that the number of abnormal samples in these 6 datasets accounts for 5% of the total number of samples. Compare the abnormal sample detection method based on random forest technology proposed in this article with the two distance-based detection methods (RHM and robust Mahalanobis distance) introduced above and compare the predictions of the built model after removing the abnormal samples from the three methods. In addition, the robustness of the three methods is compared through a model built using support vector machine (SVM) technology. The evaluated value is compared in Figure 6.

First, three methods are used to delete 5% of the "abnormal samples" in each dataset, and then the deleted dataset is used to build a random forest model. According to the total number of samples in the 6 datasets, a random forest with a scale of 500–1000 trees is established, and the number of candidate split attributes at each node q is set to the square root of the total number of attributes in the dataset, and 5-fold cross-checking is performed, respectively.

In order to further compare the robustness of the three different detection methods, we also conducted the following experiment: each dataset was cross over five times to obtain [training set i], [test set i]($i = 1, 2, \ldots, 5$). Use the training set to build the SVM model, and then use the test set to test and average all 5 test results to get the accuracy of 5-fold crossover; next, use 3 methods to eliminate abnormal samples for each [training set i], use the deleted [training set i one], establish the SVM model, and then use the [test set i] of the unremoved abnormal samples for testing, and the accuracy of the 5-fold crossover is obtained. SVM modeling uses the IibSVM toolbox, which uses the Gaussian kernel function, and uses the grid method to select the optimal penalty coefficient C.

After the three methods delete the abnormal samples in the dataset, the accuracy of the model has been improved to varying degrees, indicating that the three abnormal sample detection methods are all effective. The comparison found that the RF-based abnormal sample detection method has greater advantages than the other two methods, which greatly improves the accuracy of the model and at the same time has stronger robustness.

4.1. Comparison of Model Accuracy and Calculation Time of 3 Methods to Delete Abnormal Points in the Entire Dataset. In the modeling of sonar, wine, and zoo datasets, the three methods all improve the accuracy of the model. Among them, the RF method improves the accuracy of the model most significantly, increasing by 2% to 6%, respectively. On the breast cancer, heart, and medium-volume datasets, the RF method highlights its superiority. It takes less time than the other two methods and can locate abnormal samples more accurately. The accuracy of the model is increased by 8%, which is significantly better than the other two methods, indicating that the RF method has identified most of the abnormal samples in the dataset. The convergence is shown in Figure 7, where it can be seen that the proposed method exhibits a better performance compared with the conventional method.

Due to the existence of abnormal samples, the calculation of the mean and covariance matrix of large datasets requires more iterations. For example, in the robust Mahalanobis distance method, in order to obtain a robust mean and covariance matrix, the number of iterations of the algorithm operation exceeds 150. It takes a long time to calculate. In addition, the calculation of Mahalanobis distance requires a long calculation cycle, and its huge matrix calculation takes up a lot of memory space. During the simulation experiment, it was found that the time used by the RHM algorithm to run the program reached 102 times that of the RF algorithm, and the memory occupied was much larger than that of the RF algorithm. If the dataset has more attribute values, this problem becomes more prominent. The main running time of the RF method is the modeling process, and this process usually requires less time. For example, it only takes more than ten seconds to construct a random forest with a scale of 1,000 trees for a dataset with a capacity of 5,000 samples. Once the random forest model is established, the abnormal point scale of each sample can be quickly obtained. The calculations involved are simple counting and standardized calculations after modeling and have nothing to do with the size of the attribute value matrix. Therefore, the RF algorithm can significantly reduce the calculation running time than the other two methods in processing large datasets.

4.2. Comparison of Robustness of 3 Methods to Delete Abnormal Samples. Modeling and testing are performed by deleting only the abnormal samples in the training set but

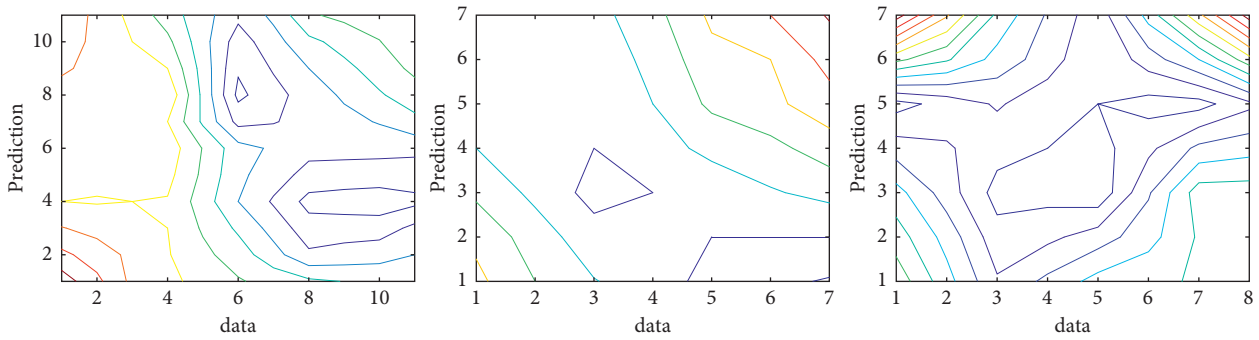

Figure 5: Prediction.

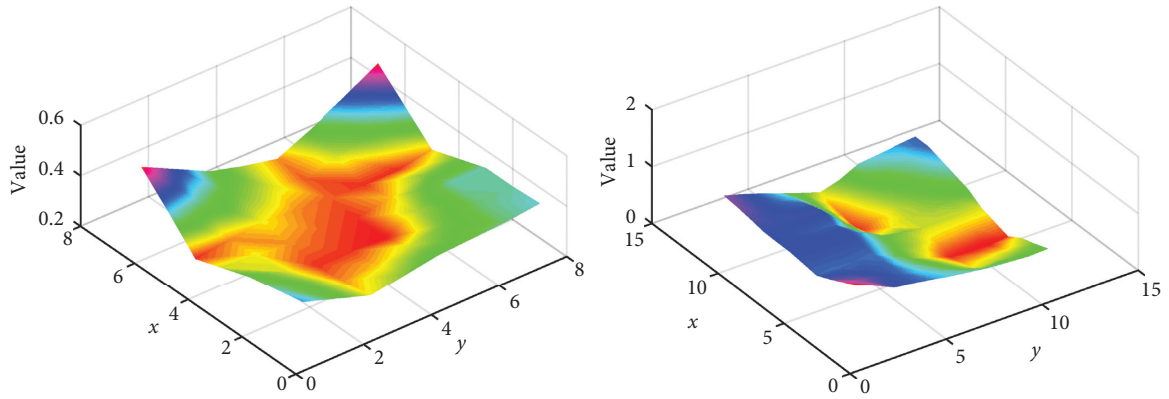

Figure 6: Evaluated value.

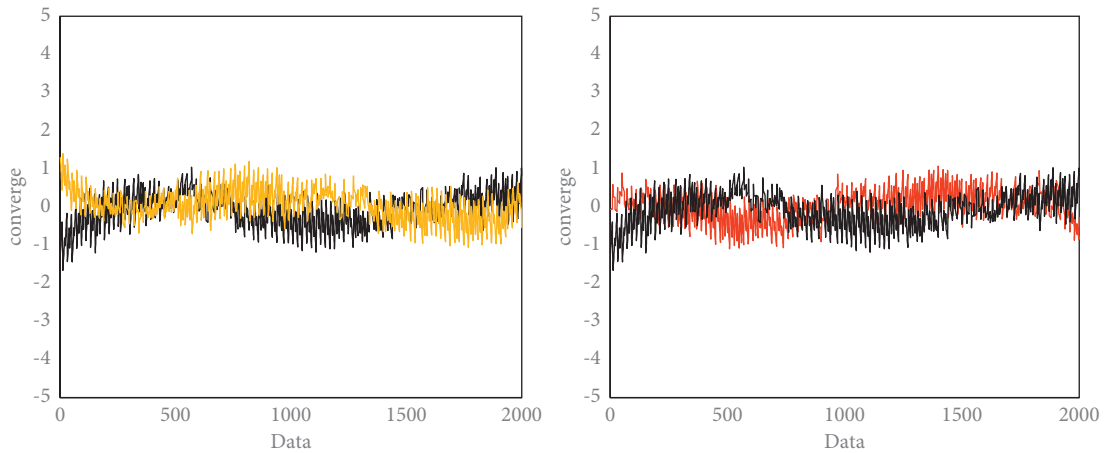

Figure 7: Convergence.

not the abnormal samples in the test set, which has higher requirements for the generalization ability of the model. The RHM method and the robust Mahalanobis distance method need to calculate the inverse of the covariance. However, if the covariance matrix is singular, a pseudo-inverse is required, which impairs the robustness of the algorithm. As discussed before, the robustness of the RF method is better than that of the RHM and the robust Mahalanobis distance method. RF guarantees the generalization ability of the model on all six datasets. The other two methods are not as robust as the RF algorithm and even drop in accuracy on the heart dataset. Similar to (1), the memory occupation and time-consuming of large datasets become a bottleneck problem that affects the detection of abnormal samples, and the RF algorithm does not have such a problem. Such advantages make the wide application of random forest-based abnormal sample detection methods possible. Figure 8 compares the predicted value.

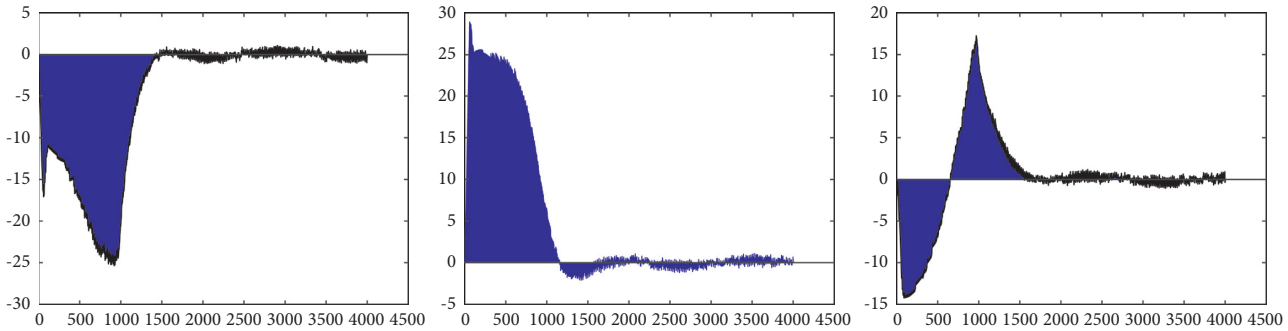

Figure 8: Comparison of predicted value.

5. Conclusion

In this paper, the random forest algorithm is introduced into the detection of abnormal samples; combined with the sample similarity, the concept of abnormal point scale is proposed to measure the degree of abnormality of the sample, and abnormal samples are screened according to this scale.

In addition to being used for outlier detection, sample similarity can also be used to establish dataset prototypes, describe dataset coordinates, and supplement missing values in training and test sets. The sample similarity provided by random forest has more and broader potential in mining the characteristics of the dataset itself. However, the scale threshold selection of abnormal samples involves only experimental results, and there is no quantitative judgment standard, which is worthy of further research.

Data Availability

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

Conflicts of Interest

The author declares that there are no conflicts of interest.

Acknowledgments

This study was supported by Hangzhou Vocational and Technical College.

References

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, Las Vegas, NV, USA, June 2016.

[2] S. Khouri, L. Elexa, M. Istok, and A. Rosova, "A study from Slovakia on the transfer of Slovak companies to tax havens and their impact on the sustainability of the status of a business entity," *Sustainability*, vol. 11, no. 10, pp. 2803–2813, 2019.

[3] S. Kumar Dwivedi, R. Amin, and V. Satyanarayana, "Blockchain-based secured event-information sharing protocol in internet of vehicles for smart cities," *Computers & Electrical Engineering*, vol. 86, no. 1, pp. 1–9, 2020.

[4] M. M. Khyareh, "Entrepreneurship and economic growth: the mediation role of access to finance," *JANUS NET E-Journal of International Relation*, vol. 1, no. 11, pp. 98–111, 2020.

[5] J. M. Cairney, K. Rajan, and D. Haley, "Mining information from atom probe data," *Ultramicroscopy*, vol. 159, no. 1, pp. 324–337, 2020.

[6] J. Yu and P. Lu, "Learning traffic signal phase and timing information from low-sampling rate taxi GPS trajectories," *Knowledge-Based Systems*, vol. 110, no. 1, pp. 275–292, 2016.

[7] L. Zhu, M. Li, and N. Metawa, "Financial risk evaluation *Z*-score model for intelligent IoT-based enterprises," *Information Processing & Management*, vol. 58, no. 6, pp. 1–9, 2021.

[8] Z. Wang, H. Ren, Q. Shen, W. Sui, and X. Zhang, "Seismic performance evaluation of a steel tubular bridge pier in a five-span continuous girder bridge system," *Structures*, vol. 31, no. 1, pp. 909–920, 2021.

[9] He Han, Y. Hong, S. Lin, and W. Liu, "Marine financial development and China's marine economic strategy," *Journal of Coastal Research*, vol. 94, no. 1, pp. 585–588, 2019.

[10] H. Shao, W. H. K. Lam, and M. L. Tam, "A reliability-based stochastic traffic assignment model for network with multiple user classes under uncertainty in demand," *Networks and Spatial Economics*, vol. 6, no. 3, pp. 173–204, 2019.

[11] A. Chen, J. Kim, and S. Lee, "Stochastic multi-objective models for network design problem," *Expert Systems with Applications*, vol. 37, no. 2, pp. 1608–1619, 2020.

[12] H. Wang, W. H. K. Lam, X. Zhang, and H. Shao, "Sustainable transportation network design with stochastic demands and chance constraints," *International Journal of Sustainable Transportation*, vol. 9, no. 2, pp. 126–144, 2015.

[13] S.-M. Hosseininasab and S.-N. Shetab-Boushehri, "Integration of selecting and scheduling urban road construction projects as a time-dependent discrete network design problem," *European Journal of Operational Research*, vol. 246, no. 3, pp. 762–771, 2015.

[14] S.-M. Hosseininasab, S.-N. Shetab-Boushehri, S. R. Hejazi, and H. Karimi, "A multi-objective integrated model for selecting, scheduling, and budgeting road construction projects," *European Journal of Operational Research*, vol. 271, no. 1, pp. 262–277, 2018.

[15] Y. Hong, "Research on anomaly detection algorithms for financial data based on angle," *Journal of Physics: Conference Series*, vol. 1345, no. 6, pp. 062038–062050, 2019.

[16] C. Soviany, "AI-powered surveillance for financial markets and transactions," *Journal of Digital Banking*, vol. 3, no. 4, pp. 319–329, 2019.

[17] X. Wang, X. Yu, L. Guo, F. Liu, and L. Xu, "Student performance prediction with short-term sequential campus behaviors," *Information*, vol. 11, no. 4, p. 101, 2020.

[18] Q. Guo, Z. Zhu, Q. Lu, D. Zhang, and W. Wu, "A dynamic emotional session generation model based on seq2seq and a dictionary-based attention mechanism," *Applied Sciences*, vol. 10, no. 6, pp. 1–10, 2020.
[19] S. Roberta, C. Paola, and A. Tomaso, "Information theoretic causality detection between financial and sentiment data," *Entropy*, vol. 23, no. 5, pp. 621–629, 2021.
[20] J. Arús-Pous, T. Blaschke, S. Ulander, J.-L. Reymond, H. Chen, and O. Engkvist, "Exploring the GDB-13 chemical space using deep generative models," *Journal of Cheminformatics*, vol. 11, no. 1, pp. 20–29, 2019.
[21] J. C. Long, "Detection of financial statement fraud using deep learning for sustainable development of capital markets under information asymmetry," *Sustainability*, vol. 13, no. 17, pp. 9879–9889, 2021.
[22] N. Pourdamghani and K. Knight, "Neighbors helping the poor: improving low-resource machine translation using related languages," *Machine Translation*, vol. 33, no. 3, pp. 239–258, 2019.
[23] L. Bote-Curiel, S. Muñoz-Romero, A. Gerrero-Curieses, and J. L. Rojo-Álvarez, "Deep learning and big data in healthcare: a double review for critical beginners," *Applied Sciences*, vol. 9, no. 11, pp. 1–11, 2019.
[24] Dubravka, "Application of data mining techniques in the detection of financial statement fraud," *Journal of Accounting and Management*, vol. 8, no. 2, pp. 97–1114, 2020.
[25] Y. Chen, Y. Ma, X. Mao, and Q. Li, "Multi-task learning for abstractive and extractive summarization," *Data Science and Engineering*, vol. 4, no. 1, pp. 14–23, 2019.
[26] P. Zhou and Z. Jiang, "Self-organizing map neural network (SOM) downscaling method to simulate daily precipitation in the Yangtze and Huaihe river basin," *Climatic and Environmental Research*, vol. 21, no. 5, pp. 512–524, 2016.
[27] X. Xiao, "Analysis on the employment psychological problems and adjustment of retired athletes in the process of career transformation," *Modern Vocational Education*, vol. 5, no. 12, pp. 216-217, 2018.
[28] H. Han, Y. Hong, W. Liu, and S. A. Kim, "Data mining model for multimedia financial time series using information entropy," *Journal of Intelligent and Fuzzy Systems*, vol. 39, no. 4, pp. 1–7, 2020.
[29] Y. Zhou and B. Yang, "Sports video athlete detection using convolutional neural network," *Journal of Natural Science of Xiangtan University*, vol. 39, no. 1, pp. 95–98, 2017.
[30] A.-H. K. Gubran and M. Pritheega, "Financial fraud detection applying data mining techniques: a comprehensive review from 2009 to 2019," *Computer Science Review*, vol. 40, pp. 1–10, 2021.
[31] G. Querzola, C. Lovati, C. Mariani, and L. Pantoni, "A semi-quantitative sport-specific assessment of recurrent traumatic brain injury: the TraQ questionnaire and its application in American football," *Neurological Sciences*, vol. 40, no. 9, pp. 1909–1915, 2019.
[32] C. Patricia, A. Kim, and L. Stefan, "Deep learning for detecting financial statement fraud," *Decision Support Systems*, vol. 139, pp. 113421–113432, 2020.
[33] G. Ma, "Research on the design of juvenile football players' sports injury prediction model," *Automation Technology and Application*, vol. 277, no. 7, pp. 141–144, 2018.