In view of the imbalance of data categories in financial data mining, a two-stage classification algorithm was proposed based on SVM and KNN to classify imbalanced data. In the first stage, two one-class SVM classifiers are constructed, and the samples are divided into four types: majority class (MC), minority class (mC), boundary, and outlier while the KNN algorithm is introduced to classify boundary and outlier samples in the second stage. In addition, the effectiveness of the algorithm is verified by several imbalanced data sets in the financial field. The results show that the proposed algorithm has predominant execution in financial precision marketing analysis, and compared with other algorithms, the proposed algorithm achieves better performance in G-mean and AUC and F1 indexes. This research provides an effective way for the classification of imbalanced data in the financial field.

1. Introduction

In the progress of Internet finance, the characteristics of data and technology dependence are very obvious. A large number of valuable user data and transaction data interact constantly among various departments of banking, insurance, e-commerce, telecommunications, and Internet financial enterprises [1, 2]. These huge amounts of data put forward higher requirements for data processing and statistical work. Therefore, data mining technology has gradually replaced the traditional statistical methods and has been widely used in the financial field.

The application of data mining technology can quickly calculate the enterprise’s operating status, cost control effect, and other financial indicators. According to the market development of different product sales, industry development environment, and other data analysis, set up a financial early warning data system. With the technology of big data mining and analysis, the company automatically obtains the relevant historical information of the company, analyzes the correlation of relevant data from multiple levels and angles, accurately calculates the sales situation of the next stage, and lays a good foundation for the subsequent promotion, sales, and production of the company’s products.

At present, one of the main problems of financial data mining is the class imbalance of data. A large part of the data involved in the financial sector is imbalanced data, which is the data set with imbalanced categories; that is, the number of samples in a certain category is much more than or less than that of other categories. For example, in the relevant data of fraud detection, fraud in credit card transactions belongs to a small number of cases; in tens of thousands of transactions, there may be only one fraud transaction. While, in the loan default data, most of the customers can repay the loan normally, only a few of them default on the loan; in the customer credit rating data, the number of customers with good credit is much more than that with bad credit. In the telecom customer churn data, the lost customers often only account for a small part of the total customers. In addition, imbalanced data are widely used in...
financial market extreme risk prediction, enterprise bankruptcy risk prediction, and other practical applications [3, 4]. In the face of imbalanced data in the financial sector, traditional data mining algorithms aiming at maximizing the overall classification accuracy may have an obvious preference for most class samples, thus ignoring the learning of a few samples. However, the majority of samples are more valuable than the class data, and the misclassification of a small number of samples may cause more losses. Traditional data mining algorithms have been unable to meet the needs of imbalanced data classification (IDC) in the financial field. How to effectively classify imbalanced data has become a research hotspot in financial data mining.

In order to solve the problem of IDC, many scholars have proposed different methods. Among them, a one-class support vector machine (SVM) is widely used in the classification of imbalanced data [5]. One-class classification algorithm only needs one class of data as training samples, which can effectively reduce the computational cost of classification. However, there are still some shortcomings in the existing one-class classification algorithms that they cannot effectively deal with the overlapping problem of class distribution, poor noise immunity, and parameter sensitivity.

Therefore, this paper will study the classification of imbalanced data from the perspective of one-class classification by proposing a two-stage classification algorithm based on SVM and KNN to classify imbalanced data. The effectiveness of the algorithm is verified by multiple imbalanced data sets in the financial field. In view of the shortcomings of the existing one-class classification algorithm, this paper proposes an improved IDC algorithm, where several representative data sets are selected from financial fraud detection, financial precision marketing analysis, and credit default prediction, and the proposed algorithm is applied to these data sets, respectively, so as to effectively solve the problem of IDC in the financial field.

2. Related Works

2.1. Imbalanced Data Sets in the Financial Field. Imbalanced data exist widely in the financial field. In the fields of financial fraud detection, financial precision marketing analysis, and credit default prediction, the data often have the problem of category imbalance. For example [6], a data set provided by the data science competition platform Kaggle is imbalanced, which contains the transaction records of European credit card cardholders. Dong et al. [7] used the Fisher method to screen out important attribute features in the data set, and based on these features, combined with the support vector machine algorithm, fraud detection is realized. Xue [8] innovatively proposed a feature set based on user labels, including user ID, e-mail address, IP, and other features, and experiments show that the classification feature set based on user tags has good discrimination and generalization. Sahin et al. [9] proposed a decision tree classification algorithm based on cost-sensitive learning and applied it to financial fraud detection. In addition to the problem of imbalanced categories, the data from the financial fields also have problems such as complex source, high feature dimension, large sample size, complex sample distribution, and sensitive feature information, especially the problem of diversity of attribute types; that is, the data include not only numeric attributes but also category attributes. Due to the continuity and easy calculation of numerical attributes, the IDC has been deeply studied.

However, there are few researches on the classification of categorical imbalanced data and mixed imbalanced data. In many areas of financial practice, mixed imbalanced data with both numerical and categorical attributes are common. For example, in the commonly used demographic characteristics, “gender” and “education level” are classified attributes, while “age” is numerical attributes; in the credit risk assessment data, “work type” and “property status” are all category attributes, while “credit amount” and “current amount of all accounts” are numerical attributes. Therefore, the classification of mixed imbalanced data not only needs to consider the problem of category imbalance but also needs to consider the continuity of numerical attributes and the discreteness of category attributes, which is more challenging.

2.2. Classification Algorithm. In recent years, a one-class classification algorithm has been widely used to solve the problem of IDC. The one-class classification algorithm was originally used to solve such problems as fault detection, target detection, outlier mining, image recognition, disease detection, intrusion detection, and other abnormal data missing problems. The idea of solving them is basically the same. They neither aim at distinguishing different classes nor making an expected output for each sample but establish the corresponding data description boundary by using only one class of samples. Thus, a description of the training sample set is given, and the classification of unknown samples is to test whether the sample conforms to the description [10].

Traditional one-class classification algorithm often only uses most class samples in the application process, but, in practical problems, there are a large number of data sets without class labels. To solve this problem, Blum and Mitchell [11] proposed a text recognition algorithm based on collaborative learning for a few class samples or a large number of unlabeled samples. Manevitz and Yousef [12] proposed different implementation methods for a one-class SVM algorithm, and when constructing the separation hyperplane, it is assumed that the origin and the samples within a certain distance from the origin belong to the target class rather than the abnormal class. Yu et al. [13] thought that since no negative class samples are used, one-class SVM needs a large number of positive class training samples to produce an accurate classification boundary and proposed a one-class classification algorithm based on SVM. Because the one-class classification algorithm only needs one class of data as training samples, it does not need to balance the imbalanced data, which shortens the training time of the classifier, so it has great application prospects in many fields. In particular, it usually has a good effect in solving the problem of IDC.
3. IDC Algorithm Based on SVM-KNN

3.1. Problem Description. One-class SVM algorithm only needs one class of samples as a training set, which can avoid the problem of insufficient learning for a few classes in IDC task and can effectively reduce the computational cost, so it is widely used in IDC tasks. However, it does not consider the problem of class distribution overlap and noise in imbalanced data sets. In the field of finance and other practical applications, there may be overlapping in the distribution of different data sets. The overlapping problem of class distribution between mC and MC samples will bring great difficulties to the one-class SVM algorithm. In addition, the noise in the data set can seriously affect the performance of the one-class SVM classifier. As shown in Figure 1, if the traditional one-class SVM algorithm is used to process the data set, it will face great difficulties.

If the one-class classifier is used to classify the unknown samples, the blue area in Figure 1 is the sample area predicted as the MC. Obviously, the minority samples in the region where the distribution of MC and mC sample coincide will be wrongly classified as MC. The one-class SVM algorithm only learns one kind of sample information, and though it avoids the problem of insufficient learning for a few classes in the IDC task, it is unable to correctly classify the boundary samples in the overlapping area of class distribution. At the same time, outliers (or noises) may exist in the above data sets. One-class SVM algorithm is easily affected by noise, and its robustness is poor. Moreover, the one-class SVM algorithm is sensitive to outliers in the training set, and outliers will also reduce the classification performance of it.

Aiming at the shortcomings of traditional one-class SVM algorithm, this paper proposes a two-stage IDC algorithm based on one-class SVM and KNN.

Figure 2 shows the boundary samples and outlier samples in the dataset, where the red region is the sample area predicted as a mC, the blue region is the sample area predicted as the MC, and the green region is the region where the boundary samples are located. SVM-KNN algorithm uses a majority-class detector and minor-class detector to detect the above data set and combines the classification results together.

Outlier samples are usually a small part of the samples that deviate from the majority of samples, while the boundary samples are often distributed in the overlapping area of the decision boundaries of most class detectors and mC detectors, which are linearly inseparable in the feature space. It not only inherits the good performance of the one-class SVM algorithm in dealing with imbalanced data but also can avoid the influence of boundary samples and outlier samples on the performance of the one-class SVM algorithm, which provides a more reasonable mechanism for the processing of imbalanced data.

3.2. Algorithm Flow. SVM-KNN algorithm solves the problem that traditional one-class SVM algorithm cannot effectively deal with boundary samples and outlier samples through two stages of classification. The algorithm flow is shown in Figure 3.

One-class SVM algorithm is used to classify the first stage. An MC detector and an mC detector are constructed by one-class learning on the MC samples and mC samples of the training set, and the test samples are predicted according to the two detectors. If there is no divergence between the MC detector and the mC detector in predicting a certain sample, the classification result of the first stage is directly regarded as the final classification result of the sample; otherwise, the sample is regarded as an undetermined sample (i.e., boundary sample or outlier sample), and the next stage is classified; Second, for boundary samples and outlier samples generated in the first stage, classification in
the second stage is continued, where KNN is used to refine the classification of boundary samples and outlier samples, and the refined classification results are the final classification results of boundary samples and outlier samples. Thus, the divergence between the majority detector and the minority detector can be eliminated by the result of refining classification. In the first stage, the single-class SVM algorithm is used to fit the majority class samples and minority class samples in the training set. At the same time, only when there are outliers or boundary samples, the classification will enter the second stage, so as to resolve the disagreement between the majority detectors and minority detectors in the first stage.

3.3. First-Stage Classification. Since the one-class SVM algorithm can only process one class at a time, in the first stage of the SVM-KNN algorithm, the one-class SVM algorithm is used to fit the MC samples and the mC samples, respectively, in the training set, that is, the training twice, so as to construct two classifiers. The specific definitions are as follows.

Definition 1. Given an unbalanced data set $IT$, let $S_1 \subset IT$ and $S_2 \subset IT$ represent the majority and minority subset of class samples in $IT$, respectively, where $S_1 \cap S_2 = \emptyset$, $S_1 \cup S_2 = IT$. The classifier $ND_1$ trained by a single-class SVM algorithm on a subset $S_1$ is called a majority-class detector, and the classifier $ND_2$ trained by single-class SVM algorithm on subset $S_2$ is called minor-class detector.

After the MC detector and mC detector are constructed on the imbalanced data set, the test samples can be classified in the first stage. For a sample $t$ to be tested, the SVM-KNN algorithm uses the MC detector and mC detector to detect it and combines the classification results of these two detectors together. If the classification results of the MC detector and the mC detector are combined together, four combinations can be formed as shown in Figure 4.

Definition 2. Given an unbalanced data set $IT$, let $ND_1$ and $ND_2$ represent majority-class detector and minor-class detector constructed by single-class SVM algorithm on $IT$, respectively. For any sample $T$ to be tested, if both $ND_1$ and $ND_2$ predict $t$ to be $-1$ (that is, $ND_1$ considers $T$ not a majority sample, and $ND_2$ considers $T$ not a minority sample), then $T$ is said to be an outlier sample. If both $ND_1$ and $ND_2$ predict $t$ to be $1$ (that is, $ND_1$ considers $T$ to be a majority sample and $ND_2$ considers $T$ to be a minority sample), then $t$ is a boundary sample.

Outlier samples are excluded by both MC detector and mC detector, while boundary sample is accepted by both MC detector and mC detector. They are usually a small part of the samples that deviate from the majority of samples, while the boundary samples are often distributed in the overlapping region of the decision boundary between the MC detector and the mC detector, which presents a state of linear indivisibility in the feature space.

3.4. Second-Stage Classification. For the sample $t$ to be tested, whether $t$ is an outlier sample or a boundary sample can be determined by combining the prediction results of the MC detector and the mC detector. If $t$ is not an outlier sample or a boundary sample, the final prediction result of $t$ can be directly obtained according to the classification results in the first stage. However, if $t$ is an outlier or boundary sample, the second stage of the SVM-KNN algorithm is needed to further classify them, that is, to resolve the divergence between the MC detector and the mC detector in the first stage through the classification results of the second stage.

Due to the imbalance of the number of MC samples and mC samples in the whole dataset, the boundary samples and...
outlier samples formed by the first-stage classification may also have the problem of category imbalance. When the absolute number of samples in a few classes is too small, the traditional classification algorithm may be difficult to learn to form an effective classification boundary. KNN algorithm may be more robust than other complex classifiers in the face of overlapping class distribution and noise data. With the increase of class distribution coincidence degree, the KNN classifier with a lower $K$ value performs better than KNN classifier with a larger $K$ value; for abnormal samples such as noise, a slightly larger value of $K$ can bring better robustness. Therefore, the KNN algorithm is used to refine and classify the boundary samples and outlier samples generated in the first stage. The refined classification results are used as the final classification results of the SVM-KNN algorithm for boundary samples and outlier samples.

### 4. Experiments

#### 4.1. Data Set.

In order to verify the effectiveness of the proposed SVM-KNN algorithm in the classification of imbalanced data in the financial field, we selected four representative imbalanced data sets from financial fraud detection, financial precision marketing analysis, and credit default prediction, respectively: Final_CreditCard, Final_PaySim, Personal Loan, Bank Marketing. The specific information is shown in Table 1.

#### 4.2. Evaluation Index.

The SVM-KNN is compared with the following five representative classification algorithms: KNN; SVM with MCes; SVM with mCes; Borderline SMOTE-KNN algorithm; EasyEnsemble-SVM algorithm. The G-mean value, AUC value, and F1 are used as the performance evaluation indexes of these algorithms.

### 5. Results and Discussion

#### 5.1. G-Mean.

Figure 5 shows the G-mean values of the SVM-KNN algorithm and other five classification algorithms on the imbalanced data sets of four financial fields.

From the above experimental results, it can be seen that in addition to Final_PaySim, the G-mean of SVM-KNN was lower than Borderline-SMOTE-KNN, and in Bank Marketing, the G-mean was lower than EasyEnsemble-SVM while, in the other four data sets, the G-mean value of SVM-KNN algorithm is higher than that of other algorithms. In the personal loan data set, the G-mean value of the SVM-KNN algorithm is 0.048, 0.156, 0.155, 0.003, and 0.140 higher than the other five classification algorithms, respectively. Therefore, from the perspective of the G-mean, the SVM-KNN algorithm has better classification performance on
imbalanced data sets from the financial field. The data cleaning technology is used to remove the noise data, overlapping samples, and the boundary data of positive and negative samples in the oversampled dataset samples, so as to improve the quality of the synthesized samples.

5.2. AUC. Figure 6 shows the AUC values of the SVM-KNN algorithm and other five classification algorithms on the imbalanced data sets of four financial fields.

From the above experimental results, we can see that the change trend of AUC is similar to that of G-mean. In the personal loan data set, the AUC value of the SVM-KNN algorithm is 0.113, 0.081, 0.205, 0.102, and 0.087 higher than the other five classification algorithms, respectively. This may be due to the fact that the five classification algorithms are very sensitive to parameters in the first stage, which are difficult to fit the MC detector and the mC detector in the first stage, respectively, to the most appropriate state, which leads to the poor classification of imbalanced data.
5.3. F1. Figure 7 shows the F1 values of the SVM-KNN algorithm and other five classification algorithms on four imbalanced financial data sets.

It can be seen from the figure that the SVM-KNN algorithm has good classification performance on multiple imbalanced data sets from different fields. Especially for the traditional one-classification SVM algorithm, in Final_CreditCard dataset, its F1 value is 0.6 lower than that of the SVM-KNN algorithm. In addition, on the personal loan dataset, the performance of easyEnsembles SVM is good, but still inferior to the algorithm proposed in this paper. The traditional one-class SVM algorithm does not consider the problem of class distribution overlap and noise in imbalanced data sets. In the field of finance and other practical applications, there may be overlapping in the distribution of different classes of imbalanced data sets, and the problem of class distribution coincidence between mC samples and MC samples will bring great difficulties to the one-class SVM algorithm.

From the above experimental results, it can be seen that, compared with the existing representative classification algorithms, the proposed SVM-KNN algorithm has better performance in solving the problems of financial fraud detection, financial precision marketing analysis, credit default prediction, and so on, which provides an effective way for IDC in the financial field.

6. Conclusion

In view of the imbalance of data categories faced by data mining research in the financial field, this paper studies from the perspective of single category classification. To solve the problem that the traditional one-class SVM algorithm cannot effectively deal with the overlapping problem of class distribution and poor noise immunity; this paper proposes a two-stage IDC algorithm based on SVM-KNN. The comparison results of G-mean, AUC, and F1 values of different algorithms show that it can effectively solve the problem of IDC in the financial field and achieve more accurate results in financial fraud detection, financial precision marketing analysis, and credit default prediction. However, this paper did not select a specific enterprise as the carrier to verify the algorithm, and the subsequent research will focus on the specific application of the model, and the economic benefits obtained.

Data Availability

The dataset can be accessed upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The study was supported by the following projects: (1) “Research on Financial Support for the Transformation and Upgrading of Small and Medium-sized Private Enterprises in Guangxi”, the project number is 2018KY0585, and (3) “University Bi-lingual Model Courses-International Finance”, Beibu Gulf University, the project number is QY16SY02.

References

[1] J. Yu, “Data mining and application research in the era of big data,” Science and Technology and Innovation, vol. 120, no. 24, pp. 166–167, 2018.
[2] J. Niu, “Internet financial innovation and transformation strategy of traditional banks in the era of big data,” China Business Theory, vol. 4, pp. 81–83, 2021.
[3] Y. Peng, G. Wang, G. Kou, and Y. Shi, “An empirical study of classification algorithm evaluation for financial risk prediction,” Applied Soft Computing, vol. 11, no. 2, pp. 2906–2915, 2011.
[4] F. Lu, F. Tang, J. Huang, and B. Wang, “Research on financial fraud account detection for imbalanced data sets,” Computer Engineering, vol. 47, no. 6, 2021.
[5] Y. Wu, W. Li, M. Ni, and Z. Xu, “Anomaly detection model based on one-class support vector machine fusion depth autoencoder,” Computer Science, vol. 49, no. 3, pp. 144–151, 2022.
[6] M. S. Kumar, V. Soundarya, and S. Kavitha, “Credit card fraud detection using random forest algorithm,” in Proceedings of the 2019 3rd International Conference on Computing and Communications Technologies (ICCCT), pp. 149–153, IEEE, Chennai, India, February 2019.
[7] Y. Dong, X. Liu, and B. Li, “A click fraud detection method based on user behavior feature selection,” Computer Science, vol. 43, no. 10, pp. 145–149, 2016.
[8] F. Xue, Design and Implementation of Credit Card Fraud Online Detection System Based on User Tag, Xi’an University of Electronic Science and Technology, Xi’an, China, 2012.
[9] Y. Sahin, S. Bulkan, and E. Duman, “A cost-sensitive decision tree approach for fraud detection,” Expert Systems with Applications, vol. 40, no. 15, pp. 5916–5923, 2013.
[10] Li Nan, Anomaly Detection Method Based on One-Class Learning and its Application in Heavy Equipment Condition Recognition, Shanghai Jiaotong University, Shanghai, China, 2014.
[11] A. Blum and T. Mitchell, “Combining labeled and unlabeled data with co-training,” in Proceedings of the eleventh annual conference on Computational learning theory, pp. 92–100, Madison, WI, USA, July 1998.
[12] L. M. Manevitz and M. Yousef, “One-class SVMs for document classification,” Journal of Machine Learning Research, vol. 2, no. 12, pp. 139–154, 2001.
[13] H. Yu, J. Han, and K. C. C. Chang, “PEBL: positive example based learning for web page classification using SVM,” in Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 239–248, Edmonton, Canada, August 2002.