A Behavior-cluster Based Imbalanced Classification Method for Credit Card Fraud Detection

Qi Li†
School of Computer and Information Technology
Beijing Jiaotong University
Beijing, China
17120384@bjtu.edu.cn

Yu Xie †
Department of Computer Science
Tongji University
Shanghai, China
1710844@tongji.edu.cn

ABSTRACT

Credit card fraud detection has been paid more and more attention by researchers. The credit card transactions are represented by highly imbalanced data sets. The number of genuine transactions is far more than fraudulent transactions, which will greatly affect the detection of fraud. Existing methods mainly consider how to balance the two classes only based on data volume, without considering the complexity of user behavior in credit card transactions, that is, the behavior noise. In this paper, we propose a behavior-cluster based imbalanced classification method. The main idea is to divide user behaviors into several group behaviors, remove behavior noise, and then hierarchical sampling. Experiments on a large scale credit card transaction data provided by a financial institution and 18 UCI data sets show that our method is superior to the existing method.

CCS CONCEPTS

•Information systems→Information systems applications→Data mining

KEYWORDS

Credit Card Fraud Detection, Imbalanced Classification, Behavior Noise

ACM Reference format:

Qi Li and Yu Xie. 2019. A Behavior-cluster Based Imbalanced Classification Method for Credit Card Fraud Detection. In *Proceedings of 2019 2nd International Conference on Data Science and Information Technology (DSIT'19). Seoul, Republic of Korea, 6 pages.* https://doi.org/10.1145/3352411.3352433

1 Introduction

Credit card fraud costs billions of dollars a year. The number rose to more than $24.71 billion worldwide in 2016 [1]. Therefore, the research on credit card fraud detection is important. In credit card fraud detection, the main problems are class imbalance [2][3], concept drift [4][5], verification latency [6], etc. Among the problems above, class imbalance is our primary consideration. The number of genuine transactions is much more than the number of fraud, which makes the traditional machine learning methods side with genuine transactions, seriously affecting the detection performance. Consequently, the application of imbalanced classification method is critical. In fraud detection, the current methods to solve class imbalance are mainly at the data-level and model-level. Data-level methods consist of sampling methods and cost-sensitive methods, while model-level methods use the ensemble classifiers. which improves the classification performance to recognize minority samples by combining sampling technique with ensemble learning. Random over-sampling increases the weight of minority samples in classification by random repeated sampling of minority class. [7] and [8] gave a detailed description of the over-sampling and under-sampling. SMOTE [9] is an improved version. Since the ROS adopts the strategy of simple replication to add samples of minority class, it may generate the problem of overfitting, that is, the information learned from the model is too specific and not generalized enough. The basic idea of SMOTE is to analyze the minority class and synthesize new samples.

Ensemble methods, e.g. Bagging [10], Easy Ensemble [11], Multi-classifier Meta-learning Approach(Chan) [12], AsymBoost [13] often use ensemble models for classification. [11] and [13] divide majority class into several subsets. Each subset is combined with minority samples to train several classifiers. The difference is that C4.5 is used as the base classifier in AsymBoost, while in Easy Ensemble, Adaboost is used for increasing the complexity of the model.

Generally, the above methods consider how to balance the two classes only based on data volume, without considering the complexity of user behavior in credit card transactions. It is well recognized that some users have similar transaction behavior. Nonetheless, some users' transaction behavior which are represented by some independent points in the high-dimensional feature space is not similar to any behavior. Their transactions may even behave like transactions that are contrary to their labels. For example, if a user's credit card information is stolen, fraudsters will conduct multiple large transactions in a short time to maximize

† Both authors contributed equally to this research.

DSIT '19, July 19–21, 2019, Seoul, Republic of Korea

ACM ISBN 978-1-4503-7141-4/19/07…$15.00
https://doi.org/10.1145/3352411.3352433

the benefits, while some normal users may also conduct large transactions in a short period of time for some reasons, that is, the individual behavior of the normal user is close to the fraud. We define this type of transactions behavior noise. In fraud detection, especially to solve the problem of class imbalance, the existence of behavior noise will make the fraud detection system judge some genuine transactions as fraud, resulting in false positives, or judge some fraudulent transactions as genuine transactions, resulting in false negatives. As the transaction that alarms the detection system will be finally confirmed by the bank staff, this will greatly increase the workload and cause verification latency [6].

In order to solve the problems above, we propose a behavior-cluster based under-sampling method. The main idea is to separate normal transactions from fraudulent transactions, and then use clustering algorithm to divide the two classes of user transactions into multiple subsets, that is, multiple group behaviors. Noise reduction is performed for each cluster, and then hierarchical under-sampling is implemented for each cluster without noise. Considering the integrity and diversity of user behavior, our method can recognize behavior noise well.

The proposed method is applied on a large-scale data provided by a financial institution, including more than 5 million credit card transactions, and 18 UCI data sets. Experimental results show an effective improvement of the performance in fraud detection when compared with several existing imbalanced classification methods.

The main contributions of this paper mainly include:

1. We propose a new behavior-cluster based imbalanced classification method, which combines clustering with under-sampling to ensure the integrity of user behavior.
2. We propose the concept of behavior noise. Considering the diversity of user behavior, the proposed method can effectively mine and eliminate behavior noise.
3. We conducted extensive comparative experiments on more than 5 million credit card transactions and 18 UCI data sets with several existing imbalanced classification methods. The experimental results show that our method is superior to the existing methods.

The rest of this paper is organized as follows: Section 2 analyses the work related to this paper. This is followed by the introduction of behavior noise and the proposed method in Section 3. Section 4 describes the experiments followed by the conclusion in Section 5.

2 Related Work

Researchers have done a lot of research on using machine learning algorithm to solve class imbalance in credit card fraud detection. Weighted Support Vector Machine (WSVM) adjusts the boundary by setting different weights on genuine transactions and fraudulent transactions to achieve effective detection. The bagging classifier [14] use J48 algorithm based on C4.5 model as the single classifier to construct the integration. Cost-based convolutional neural network [15] considers that fraudulent transactions near decision boundary are beneficial to training classifiers, so when over-sampling, more samples at the boundaries are collected, and then CNN is used to classification. [16] proposes a new ensemble method which combines bagging and boosting technology. Bagging reduces the variance of model by re-sampling the original data sets, while boosting can reduce the deviation. Cost-sensitive methods impose different costs on different categories of classification errors. [17] uses Bayesian minimum risk as the cost. The cost-sensitive method works well in some scenarios, but the disadvantage is that the cost matrix is difficult to deter- mine. KNN is used in [18]. Each new instance is compared with the existing one by distance, and the label is assigned to the new one by using the latest existing instance. It is mentioned that KNN performs stably in various credit card fraud detection methods. [19] proves that Bagging is better than Naive Bayes (NB) [20], Support Vector Machine (SVM) [21], K-Nearest Neighbor (KNN) in credit card transaction data sets.

3 Proposed Method

3.1 Behavior Noise

It is well recognized that the users' transaction behavior are diverse. From the credit card transactions with the same label, we can summarize some group behaviors, that is, the transactions with the same label in the high-dimensional feature space are a cluster formed by points close to each other. But there are also some users whose transaction behavior are not similar to that of any group, which are represented by some independent points in the feature space, and even some are similar to that of the opposite label.

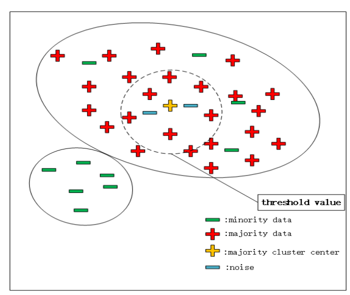

Figure 1: Behavior Noise in Majority Class

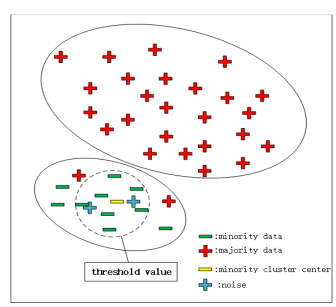

Figure 2: Behavior Noise in Minority Class

For instance, some genuine transactions are close to some fraudulent transactions, while some fraudulent transactions are close to the genuine ones in feature space. For the scenario of fraud detection, we define it as behavior noise. In the previous methods, the influence of user behavior noise on imbalanced classification methods was not considered. However, from the perspective of user behavior, it can better analyze the diversity of user behavior in credit card transactions, so as to remove the influential behavior noise.

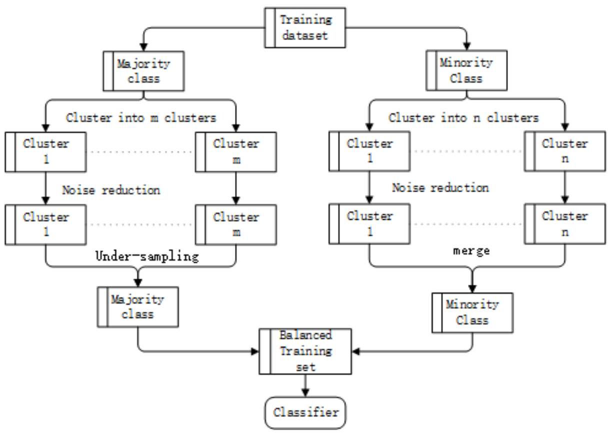

Figure 3: Flow Diagram

The behavior noise is shown in Figure 1 and 2. For simplicity, behavioral noise is pictured for one cluster in the majority class minority class respectively.

The definition of behavior noise is as follows: The data set is divided into majority class and minority class based on the label and clustered in the high-dimensional feature space respectively. The purpose is to divide user behavior into multiple group behaviors. For each majority cluster and minority cluster, the farthest distance $d_{i_{max}}$, $d_{r_{max}}$ from the transactions in cluster to the cluster centers C_{maj_i}, C_{min_r} are obtained. For each majority cluster, the region with C_{maj_i} as the center and $d_{r_{max}} * \mu$ as the radius is delineated. $\mu \in \{0, 1\}$ is a hyper-parameter that represents the range of the region. Then, the distances between the minority instances and C_{maj_i} are calculated, and all the minority instances in the defined area are identified and labeled, which are behavior noise in majority class. Similar operations are performed on each of the minority clusters. The algorithm is described in Algorithm 1.

3.2 Algorithm

In this section, we will introduce the proposed imbalanced classification method. The flow diagram is shown in Figure 3.

Firstly, we separate the positive and negative instances from the original data and use clustering algorithm to divide the positive and negative class into multiple clusters in the high-dimensional feature space. KMeans is used as clustering algorithm [22], not only because KMeans is a widely used clustering technology, it can also put the transactions with similar behavior characteristics in one group according to the distance in the feature space, and divide the transactions with different behavior characteristics into different groups. Then, for each cluster of positive and negative class, the behavior noise will be mined according to Algorithm 1 aiming to get several subsets of non-behavior noise.

Algorithm 1 Noise Reduction

Require: $X_{N_{maj}}$: The majority set
$X_{N_{min}}$: The minority set
m : The number of majority set clustering
n : The number of minority set clustering
N_{maj_i} : The number of majority set
N_{min_r} : The number of minority set
X_{maj_i} : The majority cluster
X_{min_r} : The minority cluster
C_{maj_i} : The center of majority cluster
C_{min_r} : The center of minority cluster
μ : Threshold value
Noise Reduction:
for $i = 1$ to m **do**
  $d_{i_{max}} = max(EuclideanDistance(X_{maj_i}, C_{maj_i}))$
  for $j = 1$ to N_{min_r} **do**
    $d_{ij} = EuclideanDistance(X_{N_{min}}[j], C_{maj_i})$
    if $d_{i_{max}} * \mu \geq d_{ij}$ **then**
      $DeleteX_{N_{min}}[j]$
      $returnX'_{N_{min}}$;
    end if
  end for
end for
for $r = 1$ to n **do**
  $d_{r_{max}} = max(EuclideanDistance(X_{min_r}, C_{min_r}))$
  for $l = 1$ to N_{maj_i} **do**
    $d_{rl} = EuclideanDistance(X_{N_{maj}}[l], C_{min_r})$
    if $d_{r_{max}} * \mu \geq d_{rl}$ **then**
      $DeleteX_{N_{maj}}[l]$
      $returnX'_{N_{maj}}$;
    end if
  end for
end for
Output: $X'_{N_{min}}$: The minority set after noise reduction;
$X'_{N_{maj}}$: The majority set after noise reduction;

For each positive subset, behavior-sparsity based hierarchical under-sampling is used, as is shown in Figure 4. The closer the area is to the cluster center, the more samples we take. This is because a large number of instances near the clustering center can better represent the user behavior, and in order to ensure the integrity of user's information, we also sample from areas with low user behavior concentration. The number of sampled instances in each positive cluster is related to the proportion of positive and negative transactions. Then the sampled positive instances are aggregated to form a positive set. For negative samples, all the negative samples will be aggregated to form negative set. Combining positive and negative sample training sets, a balanced training set is formed for model training.

4 Experiments

4.1 Experimental Setup

To verify the effectiveness of our proposed method, extensive experiments were conducted on a large scale credit card transaction data provided by a financial institution and 18 UCI data sets. The data provided by financial institutions includes 5125,107 credit card transactions with an imbalance ratio of 33.7. The 18 UCI data sets, with imbalance ratios ranging from 1.3 to 115 and data volumes ranging from 198 to 12960, are all binary data sets. The detailed information of these data sets is shown in Table 1 and Table 2. The data sets are divided into 80% training sets and 20% test sets by 5-fold cross validation.

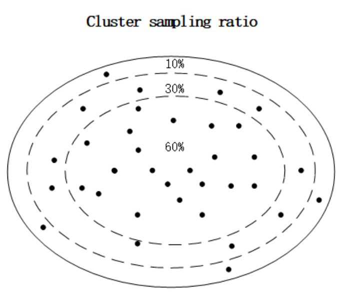

Figure 4: Each Dotted Line Is Divided by the Number of Samples (The Density of User Behaviors).

Table 1: Description of the Attributes in Credit Card Transaction Data

Attributes name	Description
Common_phone	Customer ' s usual mobile phone number
Pay_bind_phone	Customer ' s number bound on the electronic payment platform
Pre_trade_result	Customer' s verification results of the last
Is common_ip	Whether this transaction is a common IP
Trade_amount	Amount of a transaction
Pay_single_limit	Limit on the amount of a single transaction
Pay_accumulate_limit	Total daily transaction amount limit
Account_number	Credit card number
Client_mac	MAC address of a transaction
Trade_date	Date of transaction
Trade_time	Exact time of transaction
White_list_mark	Whether the account is in the trusted list
Card_balance	Account balance before payment
Transaction_object	Is the receiver a person or a business
Receiver_number	Receiver number
Last_trade_time	Account' s last transaction time

4.2 Performance Measure

In this paper, several measures are used as follows.

- $Precision = \frac{TP}{TP + FP}$
- $Recall = \frac{TP}{TP + FN}$
- $F_1 Score = 2\frac{Precision * Recall}{Precision + Recall}$

Table 2: Description of Data Sets

Datasets	Samples	Attr	min/max	Ratio
Abalone	4177	8	36/4141	115.0
Arrhythmia	452	280	13/439	33.8
Balance-scale	625	4	49/576	11.8
Cmc	1473	10	333/1140	3.4
Flag	194	9	17/177	10.4
German	1000	28	300/700	2.3
Glass	214	20	17/197	11.6
Haberman	306	3	81/225	2.8
Heart-stalog	270	14	120/150	1.3
Hepatitis	155	20	32/123	3.8
Housing	506	13	106/400	3.8
Ionosphere	351	35	126/225	1.8
Nursery	12960	8	328/12632	38.5
Phoneme	5404	5	1586/3818	2.4
Pima	768	8	268/500	1.9
Satimage	6435	36	626/5809	9.3
Vehicle	846	18	212/634	3.0
Wpdc	198	33	47/151	3.2
Credit card fraud data	5125107	16	147829/ 4977278	33.7

$Precision$ and $Recall$ are calculated by confusion matrix which is shown in Table 3. $F_1 Score$ takes into account both the $Precision$ and $Recall$ of the classification model. The AUC [23] represents the area under the receiver operating characteristic curve.

Table 3: Confusion Matrix of Binary Classification

	Positive prediction	Negative prediction
Positive class	True positive (TP)	False negative (FN)
Negative class	False positive (FP)	True negative (TN)

4.3 Experimental Results and Discussion

This section will be divided into two parts. In the first part, we analyze the experimental results in the credit card transaction data. In the second part, we compare our method with nine existing imbalanced classification methods (C4.5, RUS, ROS, SMOTE, Chan, EasyEnsemble, Asym, IRUS [24], Adasyn [25]) on 18 UCI data sets.

The experimental results of credit card transaction data are shown in Table 4. It can be seen that the Recall and AUC of under-sampling are better than those of over-sampling, but the $F_1 Score$ and $Precision$ are lower. The reasons for this are that the over-sampling methods will increase the fitness of fraudulent transactions, so the Precision will be improved. However, in the scenario of fraud detection, the goal is to find as many fraudulent transaction as possible without too much false alarm. In other words, Recall can more accurately describe the performance of fraud detection. Moreover, although the Accuracy of the over-sampling methods have been greatly improved, it will produce a large number of artificial fraud samples, if which not only greatly increases the time of mode detection, but also results in the overfitting of the existing fraud, affecting our judgment of unknown fraud. Experimental results show that our algorithm outperforms the existing popular imbalanced classification algorithms in different models.

Table 4: Experimental Results of Credit Card Transaction Data

model	accuracy	recall	precision	f1	auc
RF CNMP	0.985	0.979	0.655	0.785	0.994
RF RUS	0.984	0.979	0.653	0.783	0.993
RF EE	0.985	0.980	0.654	0.784	0.993
RF ROS	0.995	0.919	0.919	0.919	0.986
RF AD	0.995	0.905	0.927	0.916	0.987
RF SM	0.995	0.921	0.908	0.915	0.987
DT CNMP	0.974	0.976	0.521	0.680	0.975
DT RUS	0.973	0.974	0.517	0.676	0.974
DT EE	0.973	0.976	0.517	0.676	0.974
DT ROS	0.994	0.901	0.898	0.899	0.949
DT AD	0.993	0.901	0.863	0.882	0.952
DT SM	0.994	0.910	0.872	0.891	0.956
LR CNMP	0.707	0.832	0.077	0.141	0.862
LR RUS	0.684	0.821	0.071	0.130	0.822
LR EE	0.685	0.821	0.071	0.131	0.823
LR ROS	0.685	0.823	0.071	0.131	0.829
LR AD	0.780	0.985	0.114	0.205	0.960
LR SM	0.689	0.826	0.072	0.133	0.841

Table 5: Auc of 18 UCI Data Sets

Datasets	C4.5	RUS	ROS	SMOTE	Chan	EasyEnsemble	Asym	IRUS	CNPM
Abalone	0.711	0.736	0.800	0.794	0.856	0.860	0.853	0.855	**0.882(0.000)**
Arrhythmia	0.900	0.885	0.940	0.907	0.973	0.972	0.974	**0.977**	**0.977(0.000)**
Balance-scale	0.500	0.523	0.627	0.540	0.544	0.612	0.565	0.588	**0.636(0.000)**
Cmc	0.681	0.667	0.673	0.699	0.709	0.706	0.716	**0.736**	0.732(-0.004)
Flag	0.719	0.778	0.749	0.695	**0.807**	0.751	0.795	0.804	0.736(-0.071)
German	0.704	0.697	0.705	0.714	0.728	0.782	0.728	**0.766**	0.735(-0.031)
Glass	0.645	0.718	0.776	0.791	0.796	0.780	0.805	0.803	**0.812(0.000)**
Haberman	0.619	0.620	0.650	0.683	0.668	0.681	0.664	0.673	**0.722(0.000)**
Heart-stalog	0.852	0.841	0.850	0.852	0.853	0.884	0.840	0.888	**0.892(0.000)**
Hepatitis	0.795	0.789	0.782	0.781	0.828	0.848	0.836	0.838	**0.875(0.000)**
Housing	0.748	0.742	0.759	0.767	0.800	**0.817**	0.789	0.811	**0.817(0.000)**
Ionosphere	0.926	0.938	0.940	0.935	0.943	**0.974**	0.931	0.954	0.955(-0.019)
Nursery	**1.000**	0.982	0.998	**1.000**	0.999	0.999	0.999	0.999	0.994(-0.006)
Phoneme	0.920	0.900	0.926	0.918	0.924	**0.956**	0.927	0.923	0.943(-0.013)
Pima	0.778	0.765	0.777	0.777	0.801	0.809	0.769	**0.812**	0.806(-0.006)
Satimage	0.918	0.915	0.920	0.925	0.947	**0.956**	0.949	0.951	**0.956(0.000)**
Vehicle	0.825	0.785	0.824	0.820	0.839	**0.860**	0.833	0.853	0.793(-0.067)
Wpdc	0.642	0.663	0.696	0.700	0.698	0.699	0.712	0.732	**0.767(0.000)**
Average	0.771	0.775	0.800	0.794	0.817	0.830	0.816	0.831	**0.835**

The experimental results on 18 UCI data sets are shown in table 5 and table 6. Table 6 shows the of 18 UCI data sets on each algorithm while Table 7 shows the AUC. The bold type indicates the optimal result, and the value in parentheses represents the difference between our result and the optimal result. Table 6 point out that our method achieves the best results in 13 datasets with F_1Score, while AUC achieves the best results in 10 datasets in Table 7. Generally, our method is superior to the existing imbalanced processing methods, which also supports our assumption of behavior noise.

5 Conclusions

Aiming at the class imbalance in credit card fraud detection, we propose a new behavior-cluster based imbalanced classification method. We divide user behavior into several groups by clustering, and ensure the integrity of user information through hierarchical sampling. Behavior noise is defined. By comparing our method with the existing popular imbalanced classification methods on multiple data sets, it is proved that eliminating behavior noise in fraud detection is necessary and the proposed method is superior to the existing method in solving the problem of class-imbalance in fraud detection. In the following work, we will focus on the combination of the ensemble method and noise reduction, hoping to further improve the detection performance.

REFERENCES

[1]. Janet Attard. Credit Card Fraud: How to Protect Your Business. Available:https://www.businessknowhow.com/money/credit- card-fraud.htm

[2]. Pozzolo A D, Caelen O, Bontempi G. When is Undersampling Effective in Unbalanced Classification Tasks?// 2015.

[3]. Pozzolo A D, Caelen O, Johnson R A, et al. Calibrating Probability with Undersampling for Unbalanced Classification// Computational Intelligence, IEEE Symposium. 2016.

[4]. Alippi C, Boracchi G, Roveri M. Just-In-Time Classifiers for Recurrent Concepts. IEEE Transactions on Neural Networks and Learning Systems, 2013, 24(4):620-634.

[5]. Joo Gama, Bifet A, Pechenizkiy M, et al. A survey on concept drift adaptation. Acm Computing Surveys, 2014, 46(4):1-37.

[6]. G. Krempl and V. Hofer. Classification in presence of drift and latency. In 11th Data Mining Workshops, pages 596603. IEEE, 2011.

[7]. He H, Garcia E A. Learning from Imbalanced Data// Encyclopedia of Information Science and Technology. 2018.

[8]. Liu X Y, Wu J, Zhou Z H. Exploratory Undersampling for Class- Imbalance Learning. IEEE Transactions on Systems Man & Cybernetics Part B, 2009, 39(2):539-550.

[9]. Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 2002, 16(1):321-357.

[10]. Breiman L. Bagging Predictors. Machine Learning, 1996, 24(2):123-140.

[11]. Liu X Y, Wu J, Zhou Z H. Exploratory Undersampling for Class- Imbalance Learning. IEEE Transactions on Systems Man & Cybernetics Part B, 2009, 39(2):539-550.

[12]. P.K. Chan, S.J. Stolfo, Toward scalable learning with non-uniform class and cost distributions: a case study in credit card fraud detection, in: Proceedings of the 4th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY

[13]. Li G Z , Meng H H , Lu W C , et al. Asymmetric bagging and feature selection for activities prediction of drug molecules. BMC Bioinformatics, 2008, 9 Suppl 6(Suppl 6):S7.

[14]. Zareapoor M, Shamsolmoali P. Application of Credit Card Fraud Detection: Based on Bagging Ensemble Classifier. Procedia Computer Science, 2015, 48:679-685.

[15]. Fu K , Cheng D , Tu Y , et al. Credit Card Fraud Detection Using Convolutional Neural Networks// International Conference on Neural Information Processing. Springer, Cham, 2016.

[16]. Zhang Z, Krawczyk B, Garca, Salvador, et al. Empowering One-vs-One Decomposition with Ensemble Learning for Multi-Class Imbalanced Data. Elsevier Science Publishers B. V. 2016.

[17]. Bahnsen A C, Stojanovic A, Aouada D, et al. Cost sensitive credit card fraud detection using Bayes minimum risk//2013 12th international conference on machine learning and applications.IEEE, 2013, 1: 333-338.

[18]. Zareapoor M, Seeja KR, Alam AM. Analyzing credit card: fraud detection techniques based on certain design criteria. International Journal of Computer Application. 2012;52(3):3542.

[19]. Zareapoor M, Shamsolmoali P. Application of Credit Card Fraud Detection: Based on Bagging Ensemble Classifier. Procedia Computer Science, 2015, 48:679-685.

[20]. Excell D. Bayesian inference the future of online fraud protection. Computer Fraud & Security, 2012, 2012(2):8-11.

[21]. Lu Q. Research on Credit Card Fraud Detection Model Based on Class Weighted Support Vector Machine. Journal of Convergence Information Technology, 2011, 6(1):62-68.

[22]. Jain A K, Murty M N, Flynn P J. Data clustering: a review. ACM Computing Surveys, 1999, 31(3):264-323.

[23]. Fawcett, T. An introduction to ROC analysis. Pattern Recognition Letters, 2006, 27(8): 861874.

[24]. Tahir M A, Kittler J, Yan F. Inverse random under sampling for class imbalance problem and its application to multi-label classification. Pattern Recognition, 2012, 45(10):3738-3750.

[25]. Barua S, Islam M M, Murase K. A novel synthetic minority oversampling technique for imbalanced data set learning// Inter- national Conference on Neural Information Processing. 2011.