Research on Credit Card Fraud Detection Model Based on Distance Sum

Wen-Fang YU
Computer Science and Information Engineering College
Zhejiang Gongshang University
Hangzhou, China
Ywf_1@163.com

Na Wang
Computer Science and Information Engineering College
Zhejiang Gongshang University
Hangzhou, China
n_wang@163.com

***Abstract*—Along with increasing credit cards and growing trade volume in China, credit card fraud rises sharply. How to enhance the detection and prevention of credit card fraud becomes the focus of risk control of banks. This paper proposes a credit card fraud detection model using outlier detection based on distance sum according to the infrequency and unconventionality of fraud in credit card transaction data, applying outlier mining into credit card fraud detection. Experiments show that this model is feasible and accurate in detecting credit card fraud.**

Keywords-distance sum; outlier; credit card; fraud detection

I. INTRODUCTION

In recent years, with the improvement of China's economy and culture and fast pace of people's life, credit card market has a great development. By the end of 2007, there are more than 70 million credit cards in mainland China [1]. But with the increasingly growing credit card market, crimes of credit card fraud have risen in recent years, which disturbs the financial order and makes banks and cardholders suffer great losses. It also endangers the healthy development of credit services of banks. How to enhance the detection and prevention of credit card fraud becomes the focus of risk control of banks.

Traditional detection methods mainly depend on database system and the education of customers, which usually are delayed, inaccurate and not in-time. After that, methods based on discriminate analysis and regression analysis are widely used [2], which can detect fraud by credit rate for cardholders and credit card transactions, however, with a shortcoming of a large amount of data. In recent years, the prevailing data mining concerns people with credit card fraud detection model based on data mining.

Data mining is a process of finding right, creative and useful knowledge and rules which can be understood from the large amount of data. Outlier mining is one of important research fields of data mining, used to finding a small part of data objects in data set which significantly differ from other objects in common behaviors and data mode. It is widely used in financial and internet fields[3-5]. Contrasting with the whole credit card trade, credit card fraud transactions are the few anomalies. Outlier detection method, one type of data mining [6], which mines fraud transactions as outliers by a fraud detection model, is useful to reveal the valuable hidden knowledge of credit card transactions and help bank governors make right decision on fraud prevention and risk control.

This paper has 4 sections. The second section describes the basic problems. The third section simply describes the outlier mining algorithm based on distance sum and detection process. The fourth section describes the experiments of this algorithm and data analysis. Finally this paper gives the conclusions.

II. THE DEFINITIONS OF DISTANCE-BASE OUTLIER

The concept of distance-based outlier is initiated by E.M.Knorr and R.T.Ng[7]. S.Ramaswany et al[8] and S.D.Bay et al[3] improves it respectively. This method judge whether it is outlier or not according to the nearest neighbors of data objects.

1st definition: outlier: Data set $T = \{t_1, t_2, \cdots, t_n\}$， U is one data object, If the p parts of data set named S in data set is far away from object U, $S \in T$, $U \in T$, then U is named outlier[9].

This definition differs from others, but all are based on distance, that is, the same measurement for distance between objects and different for outlier's measurement. Outlier mining has the following advantages than mining based on statistics: First, this method can effectively find outliers, unknown the distribution of data set. Besides, it overcomes the shortcoming that outlier mining detection can only detect single attribute.

2th definition: distance: DIS (p，d) denotes the distance between p parts of data set named S and outlier U, where d denotes the radius of a circle with outlier U as the center, U denotes outlier based on distance.

3rd definition: degree of dispersion: Let data set $T = \{t_1, t_2, \cdots, t_n\}$， where c% of data is outlier, c is called degree of dispersion and $c = k/n$, where k denotes the number of outliers, n denotes the number of all the points in data set T.

III. CREDIT CARD FRAUD DETECTION ALGORITHM BASED ON OUTLIER MINING

A. Outlier mining algorithm based on distance sum

This algorithm judges outliers not depending on p and d but on distance sum of distance.

Outlier mining based on distance sum is described as follows [10]：

978-0-7695-3615-6/09 $25.00 © 2009 IEEE
DOI 10.1109/JCAI.2009.146

Let universe of discourse $X = \{x_1, x_2, \cdots, x_n\}$ be the object to detect. Every object has m indexes, namely $x_i = \{x_{i1}, x_{i2}, \cdots, x_{im}\}, i = (1,2, \cdots, n)$ .The following data matrix can express it as follows:

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix} \quad (1)$$

Now outlier sets of n objects are required to figure out.

In order to judge dispersion degree of every object in X, compute d_{ij} which denotes the distance between any two objects and composes distance matrix R, described as follows,

$$R = \begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1m} \\ d_{21} & d_{22} & \cdots & d_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \cdots & d_{nm} \end{bmatrix} \quad (2)$$

It is vital important to select distance function. This paper selects Euclidean distance.

$$DIS(U,\ f) = \sqrt{\sum_{i=1}^{n}(o_i - f_i)^2} \quad (3)$$

Let

$$P_i = \sum_{j=1}^{n} d_{ij} \quad (4)$$

Where P_i is the sum of ith row in matrix R. The bigger P_i is, the longer the distance between i object and other object is. Then P_i is the candidate item of outlier set.

$$\lambda_i = \frac{P_i - P_{min}}{P_{min}} \times 100\% \quad (5)$$

Where, λ denotes threshold, the objects with $\lambda_i \geq \lambda$ are taken as outlier set.

B. Design solution

Outlier mining algorithm based on distance sum is applies into credit card fraud detection. Anomaly detection process of this algorithm is as Figure 1 shows:

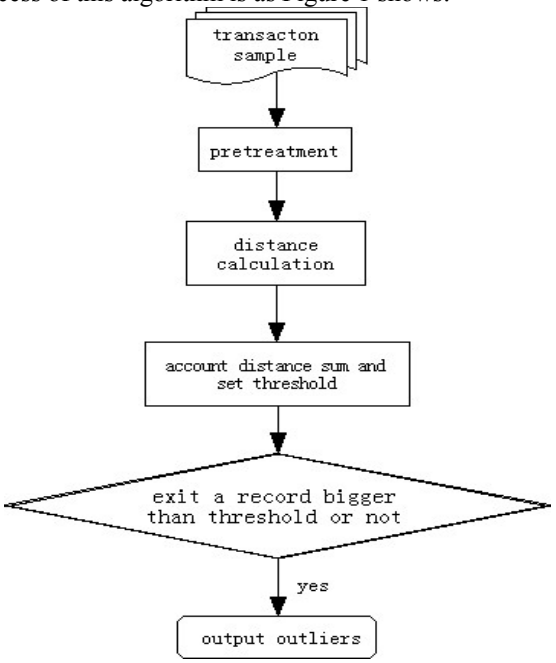

Figure 1. Credit Card Fraud Detection Model Structure

This model detects outlier sets by computing distance and setting threshold of outliers. It features that threshold varies in different situations of application as an input parameter. Therefore, outlier set detection in MODM and comprehensive evaluation is very effective.

C. Data standardization

Considering the specific attributes of the data in original data set and the influence of different measurement for attributes on the computed distance, before computing the distance, different attributes of transaction sample should be standardized which is used to compute distance. This paper gets the attribute values in standardized interval by standardization method of standard deviation in order to avoid the influence of the comparison between attributes with large range and attributes with small range on results. The standardization is as follows:

$X = \{X_i | X_i = (x_{i1}, x_{i2}, \cdots, x_{ij}, \cdots, x_{im}), i = 1,2, \cdots, n; j = 1,2, \cdots, m\}$, X is the data set for outlier mining. X_i denotes the ith object (n objects in all). X_{ij} denotes the jth attribute value of the ith object (m attributes in all).

Let $\overline{X_j}$, R_j and S_j denote the mean of the jth attribute, Mean Absolute Deviation and standard deviation respectively, described as follows:

$$\overline{X_j} = \frac{1}{n}\sum_{i=1}^{n} x_{ij}\ ; R_j = \frac{1}{n}\sum_{i=1}^{n}\left|x_{ij} - \overline{X_j}\right|; \quad (6)$$

$$S_j = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \overline{X_j})^2} \quad (7)$$

Standardized data is described as follows:

$$x_{ij}^{\ *} = \frac{x_{ij} - \overline{x_j}}{R_j} \text{ 或} x_{ij}^{\ *} = \frac{x_{ij} - \overline{x_j}}{S_j} \quad (8)$$

As for outlier mining, standardization should help isolate outliers, but attribute value and mean isn't squared while computingRj, which has limited influence on isolating outliers. Therefore, this paper selects standard deviation Sj to standardize data, described as follows: $x_{ij}^{\ *} = \frac{x_{ij} - \overline{x_j}}{S_j}$

IV. EXPERIMENT DESIGN

This paper selects real credit card data of one certain domestic commercial bank as the target object for research. Transaction records of cardholders in recent 10 years are selected, 16,584 in all, in which there are 15,135 none-fraudulent transactions and 1,449 fraudulent transactions. The frauds have different consuming habits from cardholders. And account and transaction records of cardholders can largely reflect their consuming habits. Therefore, account and transaction records are selected as object attributes [11]. Meanwhile, consuming habits correlates largely cardholders' characteristics. It is hard to judge fraudulent transactions only according to related information of transactions. Therefore, some characteristics of cardholders can be select as object attributes.

Every sample has 51 attributes. And finally 28 attributes are decided as inputs of the model by discarding the attributes which have nothing to do with fraud according to previous experiences, showed as Table1:

TABLE I. ATTRIBUTES OF TRAINING SAMPLE SET

Attribute number	Attribute
1	Customer income
2	Customer age
3	Customer profession
4	Customer position
5	Marriage status
6	Working years
7	Number of card used
8	Housing type
9	Credit card type
10	Credit grade
11	Credit line
12	Book balance
13	Times of using card
14	Times of overdraft
15	Time bracket
16	Times of overdraft
17	Times of bad debt
18	Times of overdraft but not bad debt
19	using card frequency
20	Overdraft rate
21	Growth rate of shopping
22	Average of book balance
23	Average daily spending
24	Average daily overdraft
25	Average amount per transaction
26	Average number of days per overdraft

A. *Experiment process*

Experiment process has three steps:

- ①Input a group of data of credit card transactions, every transaction record with m attributes, then standardize data and get the sample finally.
- ②Compute d_{ij} which denotes the distance between two credit card transaction records of data set and get distance matrix, then according to the equation $P_i = \sum_{j=1}^{n} d_{ij}$ get the sum. The bigger the sum is, the longer the distances between the credit card transaction record i and other objects is. The record i is the candidate item of outlier set.
- ③Compute the threshold of outliers according to this equation $\lambda_i = \frac{P_i - P_{min}}{P_{min}} \times 100\%$, and set parameter λ denoting threshold. All the objects with $\lambda_i \geq \lambda$ are taken as the output of outlier set.

B. *Analysis of experiment result*

In this section, the ratio of real frauds to predicted frauds as performance index is tested.

In order to test the performance of this algorithm, visual C+ + is used to conduct this algorithm. The preceding credit card transaction database of bank customers is utilized to conduct experiment, which contains all the attributes showed in Figure 2.

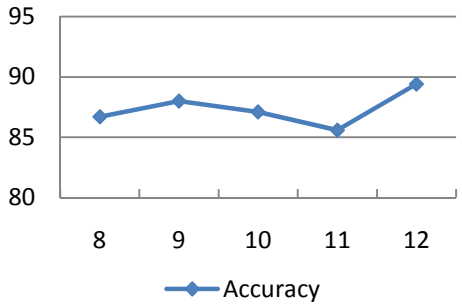

Figure 2. Accuracy of different thresholds

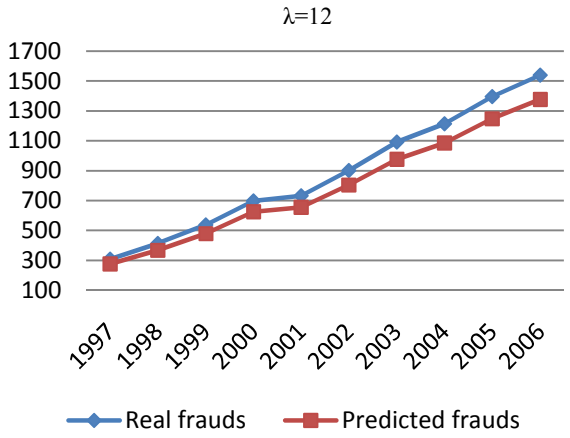

Figure 3. Comparison between predicted frauds and real frauds

Figure 2 shows different accuracy selecting different thresholds. When λ=12, the accuracy is the highest, 89.4% .While Figure 3 shows the comparison between predicted frauds and real frauds when λ=12. Experiments show that this algorithm is effective to detect the risk of malevolent overdraft and predict fraud, and also it can help governors understand the risk of malevolent overdraft and prevent

losses from malevolent overdraft in the process of sale, service and management of credit card.

V. CONCLUSION

This paper analyzes the feasibility of credit card fraud detection based on outlier mining, applies outlier detection mining based on distance sum into credit card fraud detection and proposes this detection procedures and its empirical process. And finally this method proves accurate in predicting fraudulent transactions through outlier mining emulation experiment of credit card transaction data set of one certain commercial bank. The experiment shows that outlier mining can detect credit card fraud better than anomaly detection based on clustering when anomalies are far less than normal data. If this algorithm is applied into bank credit card fraud detection system, the probability of fraud transactions can be predicted soon after credit card transactions by the banks. And a series of anti-fraud strategies can be adopted to prevent banks from great losses before and reduce risks.

ACKNOWLEDGMENT

Wenfang Yu thanks the China National Science Foundation (70671094), Zhejiang Technology Project (2008C14061), and Foundation of National Doctor Fund (20050353003) for support.

Na Wang thanks the Zhejiang Xinmiao Talent Project (2008R40G2050025), the Project of Graduate Student Science Innovation of Zhejiang Gongshang University (1130XJ1508083) for support.

REFERENCES

[1] Wang Xi. Some Ideas about Credit Card Fraud Prediction China Trial. Apr. 2008, pp. 74-75.

[2] Liu Ren, Zhang Liping, Zhan Yinqiang. A Study on Construction of Analysis Based CRM System. Computer Applications and Software. Vol.21, Apr. 2004, pp. 46-47.

[3] S D Bay, M Schwabacher. Mining Distance-Based Outliers in Near Linear Time with Randomization and a Simple Pruning Rule[C]. In:SIGKDD 03, Washington.DC.USA, 2003.

[4] J.Laurikkala, M Juhola, E Kentala. Informal Identification of Outilers in Medical Data[C]. In:5th International Workshop on Intelligent Data Analysis in Medicine and Pharmacology,(IDAMAP-2000),2000.

[5] K Yamanishi, J Takeuchi. A Unifying Framework for Detectiong Outliers and Change Points from Nen-Stationary Time Series Data[C]. In:SIGKDD 02 Edmonton, Alberta, Canda, 2002

[6] Han J W,Kamber M.Data Mining: Concepts and Techniques. Beijing: Higher Education Pr. and Morgan Kaufmann Publishers, 2007

[7] E M Knorr, R T Ng, V Tucakov. Distance-Based Outliers: Algorithms and Application[J]. VLDB Journal: Very Large Databases, 2000:237-253.

[8] S Ramaswamy, R Rastogi, K Shim. Efficient Algorithm for Mining Outliers from Large Data Sets[C]. In: Proceedings of the ACM SIGMOD Conference,2000:473-438.

[9] F Angiulli, C Pizzuti. Fast Outlier Detection in High Dimensional Spaces[C]. In: Proceedings of the 6th European Conference on the Principles of Data Mining and Knowledge Discovery, 2002-06

[10] Lu Shenglian, Lin Shimin. Research on Distance-based Outliers Detection. Computer Engineer and Applications.Vol.33, 2004, pp.73-76.

[11] Huges. The Complete Database Marketer: Second Generation Strategies and Techniques for Tapping the Power of Your Customer Database[M]. Chicago, Irwin Professional, 1996.