Automatic Bank Fraud Detection Using Support Vector Machines

Djeffal Abdelhamid[1], Soltani Khaoula[1], Ouassaf Atika[2]
[1]Computer science department, LESIA Laboratory, Biskra University, Algeria
[2]Economic science department, Economic science Laboratory, Biskra University, Algeria
Adelhamid_Djeffal@yahoo.fr, khaouladady@yahoo.fr, a_ouassaf@yahoo.fr

ABSTRACT

With the significant development of communications and computing, bank fraud is growing in its forms and amounts. We try in this paper to analyze the various forms of fraud to which are exposed banks and data mining tools allowing its early detection using data already accumulated in a bank. We propose the use of supervised learning methods called support vector machines to build models representing normal and abnormal customers behaviors and then use it to check new transactions. We also propose a hybridization of the two SVM methods, binary and single class to enhance the detection of fraudulent transactions.
The obtained results from databases of credit card transactions show the power of these techniques in the fight against banking fraud comparing them to others in the same field.

KEYWORDS

Support vector machines, Bank fraud detection, single class SVM, Binary SVM.

1 INTRODUCTION

The development of our country is the main occupation of our society. This development begins with the modernization of our businesses and administrations by introducing new management techniques, monitoring and analysis based on the use of large amounts of accumulated data. These techniques can help to minimize risk and improve the quality of services offered to customers in order to succeed in a competitive world. Fraud is a very important risk facing financial companies and banks in particular and traditional prevention techniques such as PINs, passwords and identification systems have become inadequate and heavy in modern banking systems [7]. Fraud in banks may be faced in several activities. The remote use of credit cards is a very fashionable tool of fraud; just have some information to make a purchase by other's card via the Internet.
Data mining can play a very important role in the fight against these types of fraud. It is a set of techniques for extracting relevant information from large amounts of data to assist in decision-making. The SVM method is used particularly in this context, due to its precision and its variants fitting different learning situations [1,2,10]. In the literature, several techniques have been used for the detection of fraud, including credit cards fraud. Among these techniques, there are neural networks [3], Bayesian networks [6], Markov chains [13] , sequence alignment [5] ... etc. The objective of this work is to provide to bankers an automatic fraud detection system that enables them to detect fraudulent transactions based on machine learning by support vector machine that has shown its power in several other areas such as face recognition, fingerprints, voice, ... etc.
Our goal is, therefore, to study the problem of fraud in banks and its resolution by the SVM techniques. We present an analysis of the bank fraud problem and its different forms and propose for each form, the variant of SVMs that can be used for its resolution and the necessary adaptations. We also propose a hybridization of two SVM methods: binary and single class to enhance the detection of fraudulent transactions. A system summarizing the use of the proposed solutions is designed and built in this work.
The rest of the paper is organized as follows: we first present the various forms of fraud in banks as well as the indices used to discover it, then we discuss the types of data mining solutions that can be used. In the third section, we discuss the use of support vector machines to meet the needs of detection of each form of fraud. The fourth section

presents the validation of the proposed solutions by testing them on bank databases. We conclude the article by a conclusion and envisaged perspectives.

2 FORMS AND INDICES OF BANK FRAUD

Fraud in banks has many forms; it can be internal i.e. committed by employees of the bank itself or external, committed by clients, persons or institutions foreign to the bank. We are interested in this paper to external fraud that may exist in three main forms:

2.1 Fraud by credit cards

The remote use of credit cards is a very fashionable fraud tool. It is sufficient to have just some information to make a purchase by the card of others via the Internet. The detection of credit card fraud is often based on a number of forecast indicators that are generally concluded from the transaction information retrieved from the historical database [15]. We calculate from this base, indices such as: frequency of use, the remaining unpaid balance of each cycle, the frequency of the uncovered, the maximum number of late days, shopping frequency, average number of consumption, daily transactions, the largest number of transactions in historic database ... etc. These indices or features are extracted for each transaction and are recorded for discovering patterns of fraudulent transactions.

2.2 Money laundering

Money laundering is also a well-known form of fraud, international lute against this activity is conducted by different states to discover and prosecute criminal activities that occur. The fight against money laundering in the financial industry is based on the analysis and processing statements regarding suspicious transactions detected by financial institutions [4]. Generally, only a few suspicious transactions are really money laundering operations, but the number of operations to be analyzed by financial institutions require a long time . In the literature, artificial intelligence methods are used to improve the ability of financial institutions to automatic processing of suspect data. However, the search for efficient methods for identifying suspicious transactional behaviors of money laundering remains a very active research field.

Nowadays, it is difficult to determine all the indices and variables characterizing a money laundering operation, because generally such unofficial activities are generated by complex social and economic conditions. Among the money bleaching indices used in the literature include:

- The amount of the transaction (withdrawal/payment) if it exceeds a predetermined amount by the bank, the transaction not justified, is then suspicious. For example, in Algeria this amount is set at 10,000 Euros,
- Billing: If the customer in his profession, has no accounting as in public works, agriculture, ... etc, then a transaction with large amount is considered suspicious,
- The source of transfer,
- The date of the transaction,
- Type of customer: a transaction with a high amount of passenger customer is suspected,
- The change of address,
- The speed of circulation of money in the account,
- The time of the transaction: transactions made at night with a large amount are suspect.
- ... etc.

2.3 Mortgage Fraud

At each attribution of credit to a customer by the bank, it holds a mortgage or guarantee to ensure repayment of the credit in case of repayment difficulties by the customer. Several customers present to the bank false mortgages or with overestimated value not allowing the bank a refund of its credit. This form of fraud presents to banks a significant portion of their loss. The indices used for the detection of this type of fraud are personal and professional information of customers as well as the presented mortgage.

3. FRAUD DETECTION SYSTEM

The system we propose, for the detection of bank fraud, is shown in figure 1.

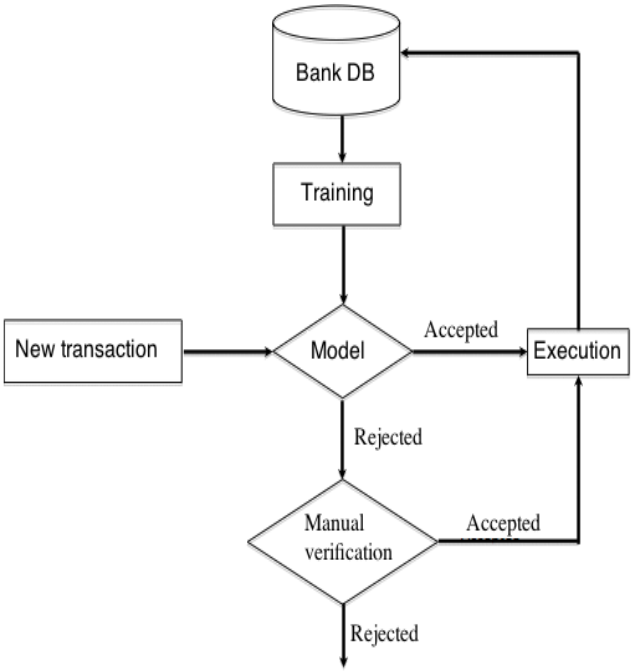

Figure 1. Fraud detection system

The system takes the data base accumulated in the bank and performs a learning for extracting a model (functions, rules,...) representing the characteristics of the data. The model is used to decide about new transactions, a transaction accepted by the model is executed then added to the database to improve the model. Transactions rejected by the model (suspicious) pass to a manual check; if they are considered normal, they are executed and then added to the database otherwise the transaction is rejected.

4 MACHINE LEARNING FOR BANK FRAUD DETECTION

The discovery of models followed by fraudsters through the analysis of their behavior is impossible due to the complexity of the operation and secondly the rapid change and development of the techniques used by fraudsters. In this context, machine learning from examples of fraud discovered by the bank can play a very important role in the fight against banking fraud.

In the literature, the two known forms of automatic learning are used: supervised and unsupervised. The supervised learning methods have been used such as association rules [11], Bayesian networks [3, 9]. These methods assume a prior knowledge of the nature of transactions, fraudulent or sane, the learning in this case consists in building a model separating the space into two parts according to the available examples then classify new examples based on their membership to one of these two classes. Unsupervised methods such as sequence alignment [12], HMM [5], neural networks [6], [14] …etc, require no prior classification of training examples, it is rather based on the detection of strange transactions.

In this work, we propose to reinforce the two types of learning by hybridizing them. Supervised learning is performed to separate fraudulent transaction of those sane then another unsupervised learning on the sane transactions only to detect strange transactions, which can take at the same time information about fraudulent transactions and information about strange data (Figure 2).

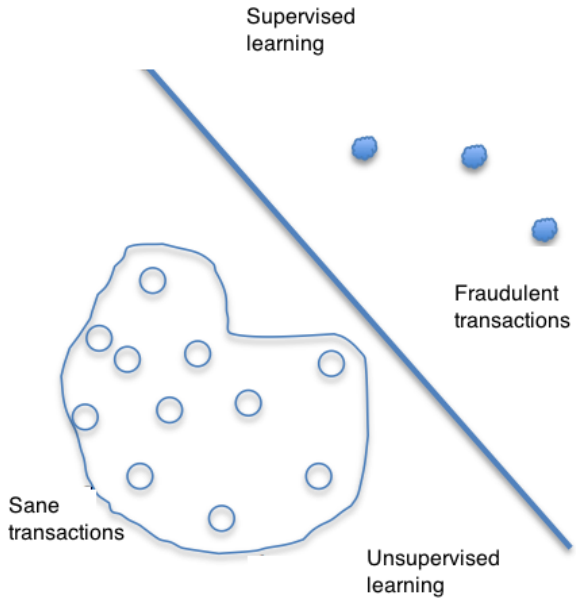

Figure 2. Hybridization of supervised and unsupervised learning for fraud detection

This technique avoids the case of a false generalization in use of the binary learning by filtering the positive part with single class learning. In figure 3, the upper-hatched area is excluded by supervised learning while the lower hatched area is excluded by unsupervised training. It is clear that the use of binary learning only leaves the lower part, representing half of the space, as part of the sane transactions, which creates a false over-generalization. While the use of single class learning only can extend the generalization to contain fraudulent trans- actions. Hybridization can well reduce sane transactions space, and therefore, extend the space of fraudulent transactions. This can pose obstacles to the rapid development of techniques of fraud by fraudsters.

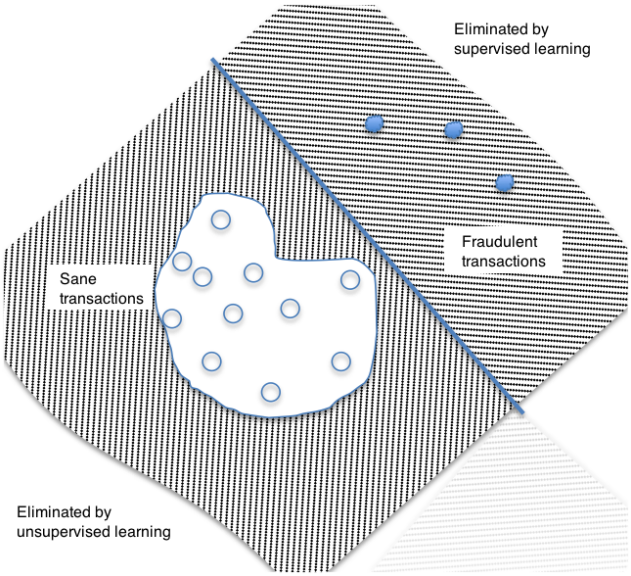

Figure 3. Advantages of hybridization of supervised and unsupervised learning

The proposed system that allows to use both types of learning side by side, is structured as in Figure 4.

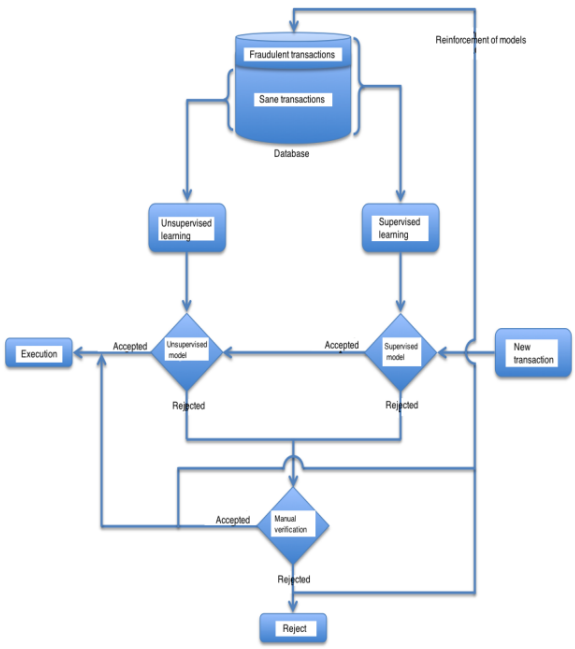

Figure 4. Proposed system

Database operations with their old sane or fraudulent classification are passed to the system. Two independent operations are launched: supervised and unsupervised learning. The first uses the entire operations and the second uses only sane transactions. Each operation provides a decision model. On arrival of a new transaction, the supervised model tests it, if it is accepted, a second test is performed by the unsupervised model, otherwise it is passed to the manual verification. If the unsupervised model accepts the transaction, it is executed; otherwise it is passed to the manual verification. After manual inspection, if the transaction is accepted it is executed, otherwise it is rejected. In both cases the transaction and the decision are added to the database to build the model with the new information.

4.1 Support vector machine

The binary SVM solves the problem of separating two classes represented by n examples of m attributes each. Consider the following problem:

$$\{(x_1,y_1),..,(x_n,y_n)\},\ x_i \in R^m,\ y_i \in \{-1,+1\}$$

Where x_i are learning examples and y_i their respective classes. The objective of the SVM

method is to find a linear function f (equation 1) called hyperplane, which allows to separate the two classes:

$$f(x)=(x \bullet w)+b \qquad (1)$$

Where x is an example to classify, w is a vector and b is a bias. We must therefore find the widest margin between the two classes, which is equivalent to minimizing $\frac{1}{2}w^2$. In the case where the training data are not linearly separable, we allow deviations ξ_i of examples relative to the boundaries of the margin of separation with a penalty parameter C, and the problem becomes a convex quadratic programming problem:

$$\begin{cases} \textit{Minimiser} & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \\ \textit{sous contraintes} & \\ & y_i(w^T x_i + b) \geq 1 - \xi_i; i = 1..n \\ & \xi_i \geq 0 \end{cases} \qquad (2)$$

The problem of the equation 2 can be solved by introducing Lagrange multipliers in the following dual problem:

$$\begin{cases} \textit{Minimiser} & \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j K(x_i, x_j) - \sum_{i=1}^{n}\alpha_i \\ \textit{sous contraintes} & \\ & \sum_{i=1}^{n}\alpha_i y_i = 0 \\ & 0 \leq \alpha_i \leq C \end{cases} \qquad (3)$$

From which we can have the following decision function (hyperplane):

$$H(x) = \sum_{i=1}^{n}\alpha_i y_i K(x_i, x) + b \qquad (4)$$

The K function is called the kernel, it is a symmetric function that satisfies Mercer conditions [12]. It can represent a transformation of the original space in which the data may be non-linearly separable to a new space with more dimensions where a linear separator exists. Solving the problem of equation 3 passes through optimization especially in the case where the number of samples is high. Among the most used optimization methods include the SMO (Sequential Minimal Optimization) where the problem is broken into several sub-problems, each must optimize tow α_i [8].

This technique is also used for the detection of outliers through a version called single class SVM. We provides to the method a set of examples with the same class, it produces a decision function that is positive for examples resembling to the training ones and negative for strange ones.

4.2 Credit card fraud and money laundering

If the bank has a historical database on fraudulent transactions and those sane, the system given in Figure 3 is used. In cases where the bank has no such historical database and all transactions are considered sane, single class learning is only appropriate. In both cases, the construction of the decision model involves three steps:

1. Features Extraction: allows to convert all transactions of each account in a features vector (vectors which will be used by the training and testing phases). The feature vector contains statistics on customer behavior such as the number of transactions, the amount handled, times and dates of transaction per day, week, month and year. This phase concerns the credit card fraud and money laundering.

2. Training: is to build the decision model.

3. Test and validation: allows to test and validate the learned model.

These three phases allow to build a model according to the scheme of Figure 5.

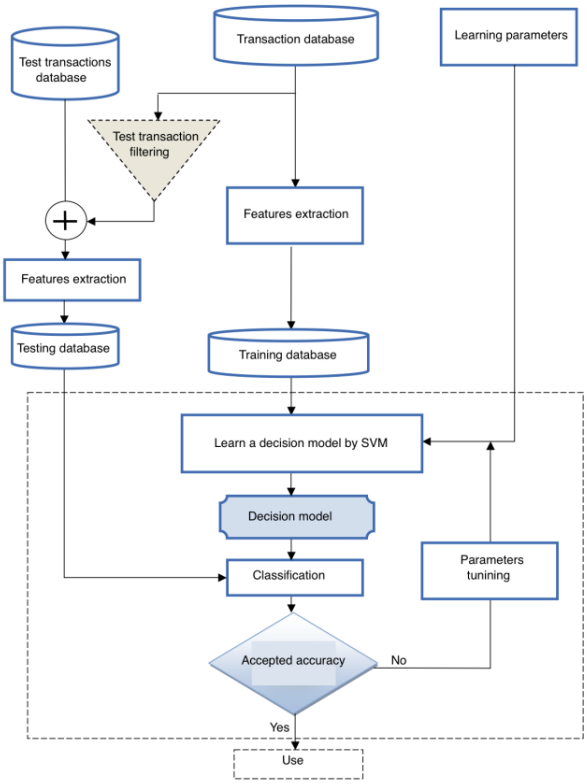

Figure 5. System for Credit card fraud detection

4.3 Mortgage fraud

In the case of this type of fraud, the detection is not based on historical transactions, but rather on the information provided by previous customers on their mortgages. Figure 6 shows the construction of the decision model.

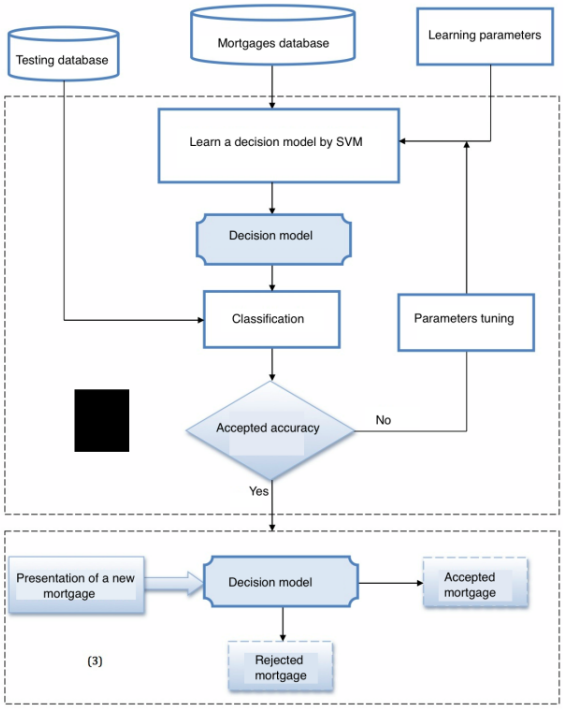

Figure 6. System for mortgage fraud detection

5 TESTS AND RESULTS

5.1 Used data

In reality, it is very difficult to obtain real data that describe the behavior of bank customers, due to the confidential nature of the data, however there are standard databases used in the literature to test fraud detection methods.

We used in our tests, three databases of different types, GeneralLedger, PayablesData and RevenueData, corresponding to credit card fraud, money laundering and mortgage fraud respectively. Databases are available from the repository ”Fraud Detection with ActiveData for Excel”. It contains data belonging only sane transactions, and consistent we used to prove the convenience of SVM method (single class) for fraud detection.

To test our proposition of hybridization, we used the German and Australian databases of credit cards using.

ISBN: 978-0-9891305-5-4 ©2014 SDIWC

5.2 Results

Results obtained on the databases of detection of banking fraud are shown in Table 1. Parameters of the single class SVM method are taken as follows:

- V=0.5
- C=100
- Sigma=0.1

The used validation method is split of 70% training and 30% test. Table 1 presents recognition rates obtained on the test databases. Preliminary results show the power of the SVM method for detecting fraud in banks. Indeed, the authors of [6] used Bayesian networks combined with neural networks, which enabled them to obtain an accuracy of about 70%. The use of hidden Markov models have achieved only accuracy of about 80% according to [13].

Table 1. Obtained accuracy by single class SVM method

Table	General Ledger Data	PayablesData	Revenue Data
Precision	94%	100%	85%

The sequence alignment method used in [5] could not get only accuracy rates below 80%. Comparing to these results, support vector machines can make a considerable improvement to accuracy of fraud detection methods in banks.
The results obtained by the method we proposed are shown in table 2.

Table 2. Results obtained by the proposed method

Table	Binary SVM	Single class SVM	Hybrid SVM
Australian	83.56%	54.67%	83.85%
German	72.4%	67.2%	72.4%

The preliminary results of the proposed method show a slight improvement over the binary method, but on bases of credit scoring and not on databases accumulating customer's behavior. We intend in the near future, as part of a PNR project to which we participate, acquire data from a national bank for use in the analysis of our proposition.

6 Conclusion

The fight against fraud is a current need for multiple sectors and banks in particular. It is in this context that we propose a system for detecting bank fraud based on support vector machines technique, depending on the application in the bank. We studied in this context, three cases of fraud in banks: credit card fraud, money laundering and mortgage fraud. We proposed, in this context a method based on the hybridization of single class and binary SVM methods.
The performance of the proposed system has been tested on the benchmarks GeneralLedger, PayablesData, Revenue- Data, Australian and German databases. The precision obtained for the single class SVM method, was of about 80%, which represents a significant improvement in comparison to similar works. For the proposed method the slight improvement on credit scoring databases was because of the difficulty of obtaining real databases. The results can be improved by studying the influence of various parameters used by the SVM method.

7 References

[1] Tareq Allan and Justin Zhan. Towards fraud detection methodologies. In Future Information Technology (FutureTech), 2010 5th International Conference on, pages 1–6. IEEE, 2010.

[2] S Benson Edwin Raj and A Annie Portia. Analysis on credit card fraud detection methods. In Computer, Communication and Electrical Technology (ICCCET), 2011 International Conference on, pages 152– 156. IEEE, 2011.

[3] Rdiger W Brause, T Langsdorf, and Michael Hepp. Neural data mining for credit card fraud detection. In Tools with Artificial Intelligence, 1999. Proceedings. 11th IEEE International Conference on, pages 103–106. IEEE, 1999.

[4] FATF-GAFI.ORG. Financial action task force on money laundering. Rapport 1996-1997 sur les typologies du blanchiment de l'argent, Groupe d'Action Financire (GAFI), Fvrier 1997.

[5] Amlan Kundu, Shamik Sural, and Arun K Majumdar. Two-stage credit card fraud detection using sequence alignment. In Information Systems Security, pages 260–275. Springer, 2006.

[6] SamMaes,KarlTuyls,BramVanschoenwinkel,andBernardManderick. Credit card fraud detection using bayesian and neural networks. In Proceedings of the 1st international naiso congress on neuro fuzzy technologies, 2002.

[7] Md Delwar Hussain Mahdi, Karim Mohammed Rezaul, and Muham- mad Azizur Rahman. Credit fraud detection in the banking sector in uk: a focus on e-business. In Digital Society, 2010. ICDS'10. Fourth International Conference on, pages 232–237. IEEE, 2010.

[8] Edgar Osuna, Robert Freund, and Federico Girosi. Animprovedtraining algorithm for support vector machines. In Neural Networks for Signal Processing [1997] VII. Proceedings of the 1997 IEEE Workshop, pages 276–285. IEEE, 1997.

[9] Suvasini Panigrahi, Amlan Kundu, Shamik Sural, and Arun K Majum- dar. Credit card fraud detection: A fusion approach using dempster– shafer theory and bayesian learning. Information Fusion, 10(4):354–363, 2009.

[10] Jon TS Quah and M Sriganesh. Real-time credit card fraud detection using computational intelligence. Expert Systems with Applications, 35(4):1721–1732, 2008.

[11] DanielSa´nchez,MAVila,LCerda,andJose´-Mar ıaSerrano.Association rules applied to credit card fraud detection. Expert Systems with Applications, 36(2):3630–3640, 2009.

[12] Bernhard Scholkopf and Alexander J Smola. Learning with kernels: support vector machines, regularization, optimization, and beyond. The MIT Press, 2001.

[13] Abhinav Srivastava, Amlan Kundu, Shamik Sural, and Arun K Majum- dar. Credit card fraud detection using hidden markov model. Dependable and Secure Computing, IEEE Transactions on, 5(1):37–48, 2008.

[14] Wen-Fang Yu and Na Wang. Research on credit card fraud detection model based on distance sum. In Artificial Intelligence, 2009. JCAI'09. International Joint Conference on, pages 353–356. IEEE, 2009.

[15] Gao Zengan. Application of cluster-based local outlier factor algorithm in anti-money laundering. In Management and Service Science, 2009. MASS'09. International Conference on, pages 1–4. IEEE, 2009.