International Journal of Scientific Engineering and Technology (ISSN : 2277-1581)
www.ijset.com, Volume No.1, Issue No.2 pg:139-142 01 April 2012

Credit Card Transaction Fraud Detection by using Hidden Markov Model

Nitin Mishra, Ranjit Kumar, Shishir Kumar Shandilya
N.I.I.S.T. Bhopal (M.P.) India
nkmishra0701@gmail.com,ranjit_kumaruitbu@yahoo.co.in, shishirsayshi@yahoo.com

Abstract: this paper proposes a HMM (Hidden Markov Model) based fraud detection system for credit card fraud detection. The method works on the statistical behavior of user's transactions. Since the original transactions are not available due to privacy policies of bank we used here synthetically generated data for a credit card user, and then HMM model is trained using different size of sample of generated labeled data we also discuss the performance of the HMM model on this data set in terms of detection accuracy and earliness of fraud detection. The system has been tested on a Pentium 4 PC with 2 GB of RAM, the test program is coded in MATLAB 7.5.

Keywords: Fraud detection, HMM (Hidden Markov Model).

1. Introduction

With the rise of economic culture standard and the rapid of people's life rhythm, Credit card market has a great development. Meanwhile, crimes involving credit card fraud increasing, this would disturb the parties financial order seriously. It cause losses to bank and cardholder, and affect development of banks. How to strengthen the ability of identifying and preventing credit card fraud has become the focus of banks risk management[2]. Traditional detection methods discern frauds mainly depending on the support of the database system and clients education level, whose disadvantages lie in bad in-time, inaccurate and hysteretic nature then Based on discriminate analysis and based on regression analysis had been presented. The two analyses identify frauds through giving credit grade to cardholder and credit card transaction and are used widely [3], but the shortcoming of big amount of data still exist. In recent years, Data Mining becomes increasingly important and has widely applied in process industry, which make people began to concern credit card fraud detection model based on Data Mining. Relative to the whole deals, credit card fraud transaction belongs to the fewness of abnormity data. In this paper, the HMM based method of detection outliers is used for set up a detection model, which could mine fraud transactions as outliers [4]; thereby provide decision support to prevent frauds and to control risks. Many outlier detection algorithms such as base on statistics [5] and distance [6, 7] are gain good application. It is based on the character of item set those above - mentioned algorithms to check outlier in data mining, and they are not suitable for the outlier checking in MODM and comprehensive evaluating. Therefore, this paper puts forward a detection model to check credit card fraud based on HMM which account the similar coefficient sum between objects to check outliers hidden in data by using the outlier mining arithmetic based on similar coefficient sum. Compared with other abnormal detection technologies, this model needn't the process of training, thus it overcome the problem of high false alarm rate. And experiments have shown that this model is feasible and veracity.

2. Related Work

Detecting credit card fraud is a difficult task when using normal procedures, so the development of the credit card fraud detection model has become of significance, whether in the academic or business community recently. These models are mostly statistics-driven or artificial intelligent-based, which have the theoretical advantages in not imposing arbitrary assumptions on the input variables. Ghosh, Reilly (1994) used a neural network based fraud detection system to train on a large sample of credit card account transactions which come from a credit card issuer. The network detected significantly more fraud accounts with significantly fewer false positives over rule-based fraud detection procedures. Hanagandi, Dhar and Buescher (1996) used historical information on credit card transactions to generate a fraud score model. The report described a fraud-nonfraud classification methodology using a radial basis function network with a density based clustering approach. The methodology tested on a fraud detection problem and the preliminary results obtained were satisfactory. Hansen, McDonald, Messier, and Bell (1996) used a powerful generalized qualitative response model to predict management fraud based on a set of data developed by an

international public accounting firm. The model included the probit and logit techniques. The results indicated a good predictive capability for both symmetric and asymmetric cost assumptions. Dorronsoro, Ginel, Sgnchez and Cruz(1997) built an online system for fraud detection of credit card operations based on a neural classifier. To ensure proper model construction, a nonlinear version of Fisher's discriminant analysis has been used. The system is fully operational and currently handles more than 12 million operations per year with very satisfactory results.

3. HMM Background

An HMM is a double embedded stochastic process with two hierarchy levels. It can be used to model much more complicated stochastic processes as compared to a traditional Markov model. An HMM has a finite set of states governed by a set of transition probabilities. In a particular state, an outcome or observation can be generated according to an associated probability distribution. It is only the outcome and not the state that is visible to an external observer [13].

HMM-based applications are common in various areas such as speech recognition, bioinformatics, and genomics. In recent years, Joshi and Phoba [14] have investigated the capabilities of HMM in anomaly detection. They classify TCP network traffic as an attack or normal using HMM. Cho and Park [15] suggest an HMM-based intrusion detection system that improves the modeling time and performance by considering only the privilege transition flows based on the domain knowledge of attacks. Ourston et al. [16] have proposed the application of HMM in detecting multistage network attacks. Hoang et al. [17] present a new method to process sequences of system calls for anomaly detection using HMM.

The key idea is to build a multilayer model of program behaviors based on both HMMs and enumerating methods for anomaly detection. Lane [18] has used HMM to model human behavior. Once human behavior is correctly modeled, any detected deviation is a cause for concern since an attacker is not expected to have a behavior similar to the genuine user. Hence, an alarm is raised in case of any deviation. An HMM can be characterized by the following [18]: The diagrams (figure 1 & 2) below shows the general architecture of an instantiated HMM.

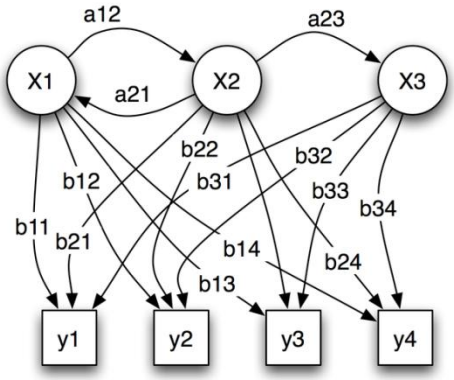

Figure 1. Architecture of an HMM

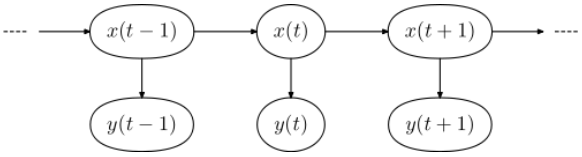

Figure 2. States & values representation in an HMM

Each oval shape represents a random variable that can adopt any of a number of values. The random variable $x(t)$ is the hidden state at time t (with the model from the above diagram, $x(t) \in \{ x_1, x_2, x_3 \}$). The random variable $y(t)$ is the observation at time t (with $y(t) \in \{ y_1, y_2, y_3, y_4 \}$). The arrows in the diagram (often called a trellis diagram) denote conditional dependencies. From the diagram, it is clear that the conditional probability distribution of the hidden variable $x(t)$ at time t, given the values of the hidden variable x at all times, depends *only* on the value of the hidden variable $x(t-1)$, the values at time $t-2$ and before have no influence. This is called the Markov property. Similarly, the value of the observed variable $y(t)$ only depends on the value of the hidden variable $x(t)$ (both at time t). In the standard type of hidden Markov model considered here, the state space of the hidden variables is discrete, while the observations themselves can either be discrete (typically generated from a categorical distribution) or continuous (typically from a Gaussian distribution). The parameters of a hidden Markov model are of two types, *transition probabilities* and *emission probabilities* (also known as *output probabilities*). The transition probabilities control the way the hidden state at time t is chosen given the hidden state at time $t-1$. The hidden state space is assumed to consist of one of N possible values, modeled as a categorical distribution. This means that

for each of the N possible states that a hidden variable at time t can be in, there is a transition probability from this state to each of the N possible states of the hidden variable at time $t + 1$, for a total of N^2 transition probabilities. (Note, however, that the set of transition probabilities for transitions from any given state must sum to 1, meaning that any one transition probability can be determined once the others are known, leaving a total of $N(N - 1)$ transition parameters.) In addition, for each of the N possible states, there is a set of emission probabilities governing the distribution of the observed variable at a particular time given the state of the hidden variable at that time. The size of this set depends on the nature of the observed variable. For example, if the observed variable is discrete with M possible values, governed by a categorical distribution, there will be $M - 1$ separate parameters, for a total of $N(M - 1)$ emission parameters over all hidden states. On the other hand, if the observed variable is an M-dimensional vector distributed according to an arbitrary multivariate Gaussian distribution, there will be M parameters
controlling,the means and $M(M+1)/2$ parameters controlling the covariance matrix, for a total

$$N(M + \frac{M(M+1)}{2}) = NM(M+3)/2 = O(NM^2)$$

of emission parameters. (In such a case, unless the value of M is small, it may be more practical to restrict the nature of the covariances between individual elements of the observation vector, e.g. by assuming that the elements are independent of each other, or less restrictively, are independent of all but a fixed number of adjacent elements.)

4. Proposed Algorithm

Here we detail the proposed algorithm for classification of Fraud Transactions.

Step 1: Generate the synthetic data according to given Probability. Use to separate distribution for Genuine and Fraud transactions.

Step 2: Read the generated data.

Step 3: Re-categorize the data into five groups as transaction month, date, day, amount of transaction & difference between successive transaction amounts.

Step 4: Make each transaction data as vector of five fields.

Step 5: Make two separate groups of data named True & False transaction group (if false transaction data is not available add randomly generate data in this group).

Step 6: Train HMM.

Step 7: Save the trained matrix.

Step 8: Read the current Transaction.

Step 9: Repeat the process from **step3** for current transaction data only.

Step 10: Place the saved Matrix & currently generated vector in classifier.

Step 11: Take the generated decision from the classifier.

5. Implementation

Since there is no real data is available because of privacy maintained by banks. Hence for testing of implementation of our algorithm we generated the data of true & false Transaction using different mean & variance & then mixed them with different probability. We used the MATLAB for the implementation of the algorithm because of its rich sets of mathematical functions and also supporting the inbuilt functions for HMM.

6. Simulation Results

The results are simulated for five different Fraud probabilities from 0.3 to 0.5 & changing the training data size from 30 to 100, then according to output of the program, following tables is drawn

Total DATA	Fraud Prob.	TPR	TNR	FPR	FNR	Accu--racy
30	0.30	0.90	0.72	0.15	0.18	0.83
30	0.40	0.61	0.59	0.38	0.41	0.60
30	0.50	0.26	0.77	0.33	0.50	0.56
60	0.30	0.98	0.22	0.38	0.03	0.72
60	0.40	0.77	0.61	0.32	0.26	0.70
60	0.50	0.70	0.75	0.29	0.24	0.73
100	0.30	0.89	0.27	0.39	0.20	0.67
100	0.40	0.65	0.43	0.51	0.38	0.54
100	0.50	0.81	0.48	0.38	0.24	0.67

This shows the maximum accuracy up to 83%, & maximum training time 1.2 seconds & maximum matching time of 0.17 seconds in P4 system with 2GB of RAM.

7. Conclusion

Referring to results we can say that proposed algorithm can be used for automatic Fraud transaction classification with excellent accuracy & negligible delay. We can enhance this model for dynamic improvements in training of HMM.

REFERENCES

[1] Wang Xi. Some Ideas about Credit Card Fraud Prediction China Trial. Apr. 2008, pp. 74-75.
[2] Chen Lei. Fraud and Prevention of International Credit Card. China Credit Card. Jun. 2004, pp. 43-47. vol. 294, Dec. 2001, [3] 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
[4] Han J W, Kamber M. Data Mining: Concepts and Techniques. Beijing: Higher Education Pr. and Morgan Kaufmann Publishers, 2007.
[5] Barnett V, Lewis T. Outliers in Statistical Data New York: John Wiley & Sons, 1994.
[6] Knorr E, Ng R. A Unified Notion of Outliers:Properties and Computation In proc. 1997 Int. Conf. Knowledge Discovery and Data
Mining(KDD 97), Newport Beach, CA, 1997, pp. 219-222.
[7] Arning A, Agrawal R, Raghavwn P. A Linear Method for Deviation Detection in Large Database. In Proc. 1996 Int. Conf. Data Mining and Knowledge Discovery(KDD07), Portland, OR, Aug. 1996, pp. 164-169.
[8] E. Aleskerov, B. Freisleben, and B. Rao, "CARDWATCH: A Neural Network Based Database Mining System for Credit Card Fraud Detection," Proc. IEEE/IAFE: Computational Intelligence for Financial Eng., pp. 220-226, 1997.
[9] M.J. Kim and T.S. Kim, "A Neural Classifier with Fraud Density Map for Effective Credit Card Fraud Detection," Proc. Int'l Conf. Intelligent Data Eng. and Automated Learning, pp. 378-383, 2002.
[10] W. Fan, A.L. Prodromidis, and S.J. Stolfo, "Distributed Data Mining in Credit Card Fraud Detection," IEEE Intelligent Systems, vol. 14, no. 6, pp. 67-74, 1999.
[11] R. Brause, T. Langsdorf, and M. Hepp, "Neural Data Mining for Credit Card Fraud Detection," Proc. IEEE Int'l Conf. Tools with Artificial Intelligence, pp. 103-106, 1999.
[12] C. Chiu and C. Tsai, "A Web Services-Based Collaborative Scheme for Credit Card Fraud Detection," Proc. IEEE Int'l Conf. e-Technology, e-Commerce and e-Service, pp. 177-181, 2004.
[13] C. Phua, V. Lee, K. Smith, and R. Gayler, "A Comprehensive Survey of Data Mining-Based Fraud Detection Research," http:// w.bsys.monash.edu.au/people/cphua/, Mar. 2007.
[14] S. Stolfo and A.L. Prodromidis, "Agent-Based Distributed Learn-ing Applied to Fraud Detection," Technical Report CUCS-014-99,Columbia Univ., 1999.
[15] C. Phua, D. Alahakoon, and V. Lee, "Minority Report in Fraud Detection: Classification of Skewed Data," ACM SIGKDD Explora-tions Newsletter, vol. 6, no. 1, pp. 50-59, 2004.
[16] V. Vatsa, S. Sural, and A.K. Majumdar, "A Game-theoretic Approach to Credit Card Fraud Detection," Proc. First Int'l Conf. Information Systems Security, pp. 263-276, 2005.
[17] S. Axelsson, "The Base-Rate Fallacy and the Difficulty of Intrusion Detection," ACM Trans. Information and System Security, vol. 3, no. 3, pp. 186-205, 2000.
[18] L.R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," Proc. IEEE, vol. 77, no. 2, pp. 257-286, 1989.