CREDIT CARD FRAUD DETECTION BASED ON ONTOLOGY GRAPH

Ali Ahmadian Ramaki[1], Reza Asgari[2] and Reza Ebrahimi Atani[3]

[1]Department of Computer Engineering, Guilan University, Rasht, Iran
ahmadianrali@msc.guilan.ac.ir
[2]Department of Computer Engineering, Guilan University, Rasht, Iran
rezaasgari@msc.guilan.ac.ir
[3]Department of Computer Engineering, Guilan University, Rasht, Iran
rebrahimi@guilan.ac.ir

ABSTRACT

Using graphs as to extracting and presenting data has a wide range of applications. Such applications may appear in detecting semantic and structural patters and exploiting graphs toward such applications have steadily been growing. In this paper we are going to display one of the most perilous abnormalities in credit cards industry on such concept basis. With advancing technology in field of banking, the rate of use of credit cards has remarkably been escalated. Correspondingly frauds frequency have increased in this area which to surmount such anomalies we model them by means of graphs. Of the prominent advantage of proposed approach is drop of system overload rate during running computations in order to detecting frauds and consequently acceleration of detection speed.

KEYWORDS

Credit Card, Fraud, Legal Transaction, Fraudulent Transaction, Graph Model

1. INTRODUCTION

Today one of the biggest threats to commercial institutes is fraud in credit cards. Understanding of fraud mechanism for fighting back its effects is subsequently a necessarily salient task. Fraudsters swindling by credit cards take advantage of disperse methods to perpetrate such illegal acts [1]. Not long ago banking researches have been performed toward morality in banking industry and proportionately fraudulent acts have become intricate. Fraud in fact connotes gaining a service, commodity and money by means of dishonest methods growing all over the world. Fraud as a crime is occasionally hard to detect.

Such fraud may occur through any type of credit production like personal loans, home loans or retailing. Moreover, methods to fraud have marvellously been expanded in parallel to technology promotion. A necessity hence, for different prominent businesses such as banking is to applying systems and processes to thwart fraud in their business field. Varied methods to fraud detection have been developed each of which has its own pros and cons [2]. As to significance of transactions performance in such field alongside with synchronization of system responses against every transaction performed by credit cards owners different approaches are practiced. Of the most important approaches we can list data mining and neural networks [3]. Their most remarkable disadvantages are high computational overload and time-consuming detection process

DOI : 10.5121/ijsptm.2012.1501

because of using all legal and illegal acts performed by a user through registered history of transactions in a system [4].

Detecting such abnormalities credit cards operations by exploiting a concept called ontology is a very efficient approach through which low computational overload and less storage for managing credit cards transactions data is required and data mining is utilized for abnormalities detection. In this approach by use of ontology graph for every user's transaction behaviour and then storage in the system, during abnormality detection only those transactions from registered history of transactions are selected to perform computation which are highly similar to entry transactions.
The rest of paper is organized as follows: in section 2 the main framework for graph model is expounded. In section 3 semantic relations between data credit cards owners' behaviour based on ontology graph is explained and then its stand in suggested model is examined. In section 4 fraud detection computation according to a specified threshold and detection of outlier is described. And finally in section 5 the conclusion is drawn.

2. Related Work

From the work of view for preventing credit card fraud, more research works were carried out with special emphasis on data mining and neural networks Ghosh and Reilly [5] have proposed credit card fraud detection with a neural network. They have built a detection system which is trained on a large sample of labelled credit card account transactions Aleskerov and Freisleben [6] present CARDWATCH, a database mining system used for credit card fraud detection. The system uses neural network to train specific historical consumption data and generate neural network model. The model was adopted to detect fraudulence. Sam and Karl [7] suggest a credit card fraud detection system using Bayesian and neural network techniques to learn models of fraudulent credit card transactions. Kim and Kim have identified skewed distribution of data and mix of legitimate and fraudulent transactions as the two main reasons for the complexity of credit card fraud detection [8].

Fan et al. suggest the application of distributed data mining in credit card fraud detection. Brause et al. [9] have developed an approach that involves advanced data mining techniques and neural network algorithms to obtain high fraud coverage. Phua et al. [10] have done an extensive survey of existing data-mining-based fraud detection systems and published a comprehensive report. Prodromidis and Stolfo [11] use an agent-based approach with distributed learning for detecting frauds in credit card transactions. It is based on artificial intelligence and combines inductive learning algorithms and meta-learning methods for achieving higher accuracy and Phua et al. [12] suggest the use of metaclassifier similar to in fraud detection problems.

3. Framework for Fraud Detection

In the current section we define the suggested framework of fraud detection system. Credit cards owners perform their favourite transactions according to their needs throughout a term of a credit card use for a specific account number. For every transaction performed by user, data required for register of performed activities as a transaction in a graph form would be stored. The framework suggested for fraud detection in transactions connected to a credit card is shown in Figure 1. First, we state some suppositions about a legal transaction performance by a credit card owner. User has a credit card for a specific account number. In other words, all transactions of the credit card owner are performed manually or by the credit card in question. Conditions considered are mostly carried out either by means of ATM stations or POS machines.

In the first place data of transactions performed by the person carrying it out (whether credit card owner or any other person) are stored based on ontology graph. Data applied for storage in such ontology graph are those which are influential in detection process. These data are listed in Table 1.

Generally data stored in graph data structure for every entry transaction falls into four categories. Data which are similar in all transactions, data in term of amount of transaction, data in connection with transaction location occurrence, data related to quantity of performed transactions by user in a specific time interval. These data division into different categories is shown in Table 2. Another task which Make Ontology unit fulfils is pre-processing operation on every entry transaction so that there would be no conflict between data stored in each graph during processing operation on transactions.

In the next stage there is a comparing unit called Compare Unit which its main task is receiving the ontology graph created in the previous stage and then matching it with current patterns in Pattern DB. In fact, the principle of the framework is this part that regarding data graph received from new entry transactions, the data graph stored in history select those users' behaviour model resembling the graph and according to output data of the matching process the chief mission of outlier is defined. In Pattern DB both legal and fraudulent acts graphs are stored.

Using such method provides the system with three significant benefits. The first one is that computational overload to fraud detection operation would drop due to getting only similar patterns involved despite of all present ones in patterns part. The next benefit is real-time detection of entry transactions in case of resembling those fraudulent transactions already stored. And the last one is that with respect to the similar patterns to a specific entry transaction being applied for detection process, fulfilling offline fraud detection is feasible as well.

The next unit is Calculation Unit which its chief task mission is performing main computations on entry transaction data and transactions stored in similar behaviour patterns stored in patterns bank and how a transaction is labelled as fraudulent by discovering outlier is brought in section 4.
After finishing the computations, if entry transaction is identified as an outlier data graph pattern of analysis unit does not store it. The other unit is Decision Unit which categorizes the entry transaction according to finished computations by Calculation Unit consequently entry transactions are either performed or prevented. Offline transactions tagged as fraudulent would be issued alert.

4. ONTOLOGY GRAPH TO PRESENT DATA

Ontology graph needed to present data is shown in Figure 2. Regarding offered framework for fraud detection in this field we already stated once a transaction is completed by a user, data required for fraud detection must be stored in ontology graph and data relationship should be modelled in graphs.

To generate the ontology in question we gain aid from rules stated in for producing ontology graphs (by adding necessary elements to fulfil needs of system to present data) [13].

Ontology is mostly presented by three agents [14]:

- Describing relations between classes
- Describing relationship between instances
- Describing relationships between classes, attributes and instances

Therefore ontology is defined as a definite set O according to (1). We also utilize set R from (2) to describe relationships.

$$O = \{o_1, o_2, \ldots, o_n\} \tag{1}$$

$$R = \{isA, synOf, partOf, atr, val\} \tag{2}$$

In (1) O is a set defining database ontology. is o_i an ontology employed in database. In (2) we have: "isA" to describe relationships including inheritance between concepts, "synOf" to describe a set of all current synonyms for a concept, "partOf" to describe relationships of type Aggregation (relationships in which a concept is a subpart of a global concept). Although by omitting global concept subparts would not be deleted. For instance, imagine connection a football player and a club. By deleting club, nature of players to it would not be deleted because they can be employed in another club. "Atr" and "Ins" are used to respectively represent presentation of attributes to a concept and having access to values of a concept.

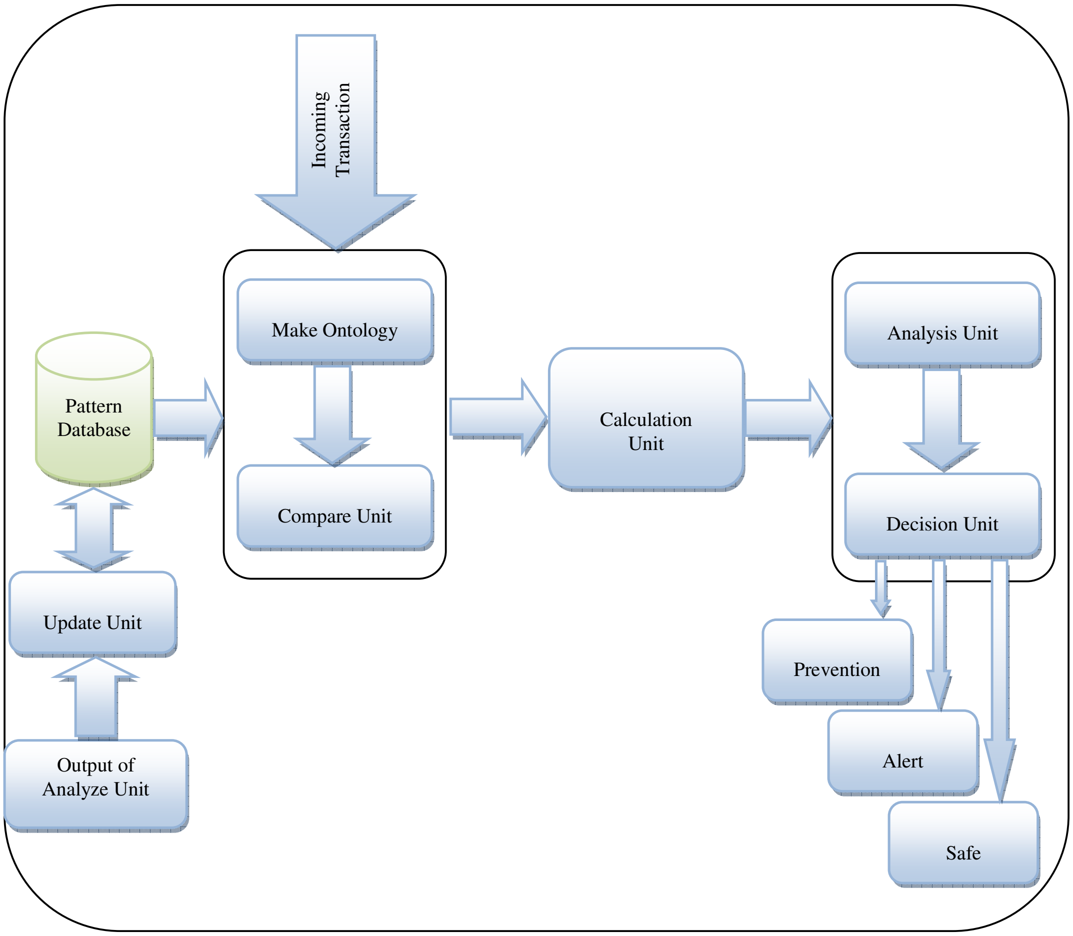

Figure 1. Proposal framework for credit card fraud detection based on ontology

Each $o_i \epsilon O$ and $r_i \epsilon R$ are considered respectively a concept and connection between two concepts. Ontology is defined as a graph G (V, E) [14] which V is a definite set of vertices and E is a definite set of edges. Every vertex is tagged with a concept and every edge is connection between two vertices.

Each tag for node $n\epsilon N$ is defined through function $N(n) = o_i \epsilon O$ mapping each node onto a string of o_i. Each edge $e\epsilon E$ is tagged through function $T(e)$ mapping each edge onto a string of R. Edges tagged as “atr” point to a head of link list including set of attributes of a concept values of concepts are approachable through tracing nodes of tag “val”. The other nodes point to current concepts in ontology. To illustrate the issue let’s take a look at Figure 2 which is part of an ontology of transactions of credit cards. Required data for fraud detection in credit cards are dropped in Figure 2. Ontology graph for user’s entry transaction is produced by such data and their patterns if needed, are then stored in patterns database part. On an accurate categorization required data are split into four groups according to their concept as in Table 1.

5. FRAUD DETECTION METHOD IN THE OFFERED FRAMEWORK

Method which Calculation Unit employs in here to detect fraud in credit cards transactions is founded on outlier detection. Outliers are applied in abnormality detection for determining detrimental behaviours. Outliers are abnormal behaviours which don not match any specific

Table I: Definition of variables

Attribute number	Attribute Name	Class of Attribute	Type of Attribute
1	ATM/POS Terminal Number	2	Integer
2	Account Number	1	Integer
3	Time of Transaction	1	Time
4	Date of Transaction	1	Date
5	Transaction Amount	3	Float
6	Credit Card Number	1	Integer
7	ATM Flag	1	Bit
8	The Number of Transactions of this Card in the Same Day	4	Integer
9	The Total Amount of Transactions of this Card in the Same Day	3	Float
10	The Average Transaction Amount of This Card in the Same Day	3	Float
11	The Number of Overdraft Transactions of this Card in the Same Day	4	Integer
12	The Number of Transactions of this Card in a Week	4	Integer
13	The Total Amount of Transactions of this Card in a Week	3	Float
14	The Average Transaction Amount of This Card in a Week	3	Float
15	The Number of Overdraft Transactions of this Card in a Week	4	Integer
16	Growth Ratio of Doing Transaction in Tow Consecutive Weeks	3	Float

Table II: Classification of variables

Class Name	Number of Class
All Transaction	1
Regional	2
Daily Amount	3
Daily Count	4

main behaviour patterns. The most salient advantage of this method in fraud detection is their being real-time and high accuracy [15, 16].

In this method, according to user's transactions we first choose pattern behaviour for him so that by deviating from the pattern it would be marked as outlier the main goal of detection of such frauds.

This approach is a subfield of data mining known as outlier mining focusing on discovering a small bunch of instances among an enormous dataset is disagreeing from behaviour or data mode perspective.

5.1. Preliminary Definitions

Outlier: Considering $T = \{t_1, t_2, \dots, t_n\}$ U is a data instance if P parts of dataset represented by S is distant from U where $S \in T$ and $U \in T$ then U is outlier [17]. Discovering such outliers is established on distance between data instances. We now define distance.

Distance: $DIS(p, d)$ is suggestive of distance between P parts of dataset S and outlier U where d is a radius with outlier U as centred and U is a distance- oriented outlier. To do so variance rate must be taken into account [17].

Variance Rate: Considering dataset $T = \{t_1, t_2, \dots, t_n\}$ wherein $c\%$ of data is outlier and c is called variance rate and $c = k/n$ where k stands for quantity of outliers and n represents current dots in dataset$T$.

5.2. Fraud Detection Algorithm

Outlier mining algorithm is described in terms of sum of distances as follows:

Assuming $X = \{x_1, x_2, \dots, x_n\}$ is the instance supposed to be detected. Every instance in database holds m characteristics denoted by $X_i = \{x_{1n}, x_{2n}, \dots, x_{in}\}$ and $i = \{1,2, \dots, n\}$. Data matrix hence is shown as below such that one of them is entry transaction and the rest are similar transactions extracted patterns already stored.

$$X = \begin{bmatrix} a_{11} & \cdots & a_{1m} \\ a_{21} & & a_{2m} \\ a_{31} & \ddots & a_{3m} \\ \vdots & & \vdots \\ a_{n1} & \cdots & a_{nm} \end{bmatrix} \tag{3}$$

We now determine a set of outliers of these n instances. As for variance rates of every instance in$X$, d_{ij} denoting the distance between two data instances is gauged through which matrix R is formed:

$$\mathrm{R} = \begin{bmatrix} d_{11} & \cdots & d_{1m} \\ d_{21} & & d_{2m} \\ d_{31} & \ddots & d_{3m} \\ \vdots & & \vdots \\ d_{n1} & \cdots & d_{nm} \end{bmatrix} \tag{4}$$

We now approach defining the distance function which is considered to the Euclidean form:

$$DIS(U,F) = \sqrt{\sum_{i=1}^{n}(o_i - f_i)} \tag{5}$$

$$P_i = \sum_{j=1}^{n} d_{ij} \tag{6}$$

Where p_i is the sum of each row in matrix R. The largest p_i is taken for outlier set:

$$\lambda_i = \frac{P_i - P_{min}}{P_{min}} \times 100\ \% \tag{7}$$

Where λ represents threshold, instances with $\lambda_i > \lambda$ is considered as outlier set. The model discriminates outliers in terms of distance measure and adjustments to threshold. Extent of such threshold is varied on the basis of various conditions and different applications as entries. Previous to distance, various characteristics and attributes of transactions of entry samples are supposed to become heterogeneous. Having current data in dataset heterogeneous for distance measure is one of main steps practiced by Make Ontology unit.

According to (7) in which X is dataset to outlier mining. X_i symbolizes ith instance –total is n. X_{ij} symbolizes the amount of *j*th attribute of instance *i* – total is *m.*

$$X = \begin{Bmatrix} \mathrm{X}_i \mid \mathrm{X}_i = (\mathrm{x}_{i1}, \mathrm{x}_{i2}, \mathrm{x}_{i3}, \dots, \mathrm{x}_{ij}, \dots, \mathrm{x}_{im}) \\ i = 1, 2, 3, \dots, n; j = 1, 2, 3, \dots, m \end{Bmatrix} \tag{8}$$

$\bar{X}_j$, R_j and S_j respectively represent [absolute deviation], *j*th attribute and [standard deviation] described as below:

$$\bar{X}_{j=} \frac{1}{n} \sum_{i=1}^{n} x_{ij}\ ; R_j = \frac{1}{n} \sum_{i=1}^{n} \left| X_{ij} - \bar{X}_j \right| \tag{9}$$

$$R_j = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} \left(X_{ij} - \bar{X}_j \right)^2} \tag{10}$$

Heterogeneous data are as below wherein [standard deviation] $S_j$is picked for data heterogeneity.

$$X_{ij}^{*} = \frac{X_{ij} - \bar{X}_j}{S_j} \tag{11}$$

Thus fraud detection process by outlier mining based on distance sum has three pillars:

1. Receiving a bunch of credit cards transactions data as entry; every transaction has *m* attributes after which data values are heterogeneous and form the final value of the sample.

2. d_{ij} Measurement representing distance between two records of credit cards transactions from dataset after which distance matrix is formed and according to (6) their sum is calculated then.

3. Outlier's threshold is measured by (7) and parameter λ denoting threshold is set. Every instance wherein $\lambda_i > \lambda$ are taken for set of outliers.

6. RESULTS AND ANALYSIS

In this section, the mechanism of the proposed framework for credit card fraud detection is evaluated. At first it should be told that for doing required experiments for inspecting described algorithm performance, the program of this algorithm is written in java on a general platform. Also required dataset based on its properties are listed in Table 1. By Using MATLAB software, we created a set of experimental data. It should be noted that the data values for this dataset are based on a standard dataset that is used by researchers I this field.

6.1. Mechanism

When the fraud detection system runs, each incoming transaction will be accepted by the system, its ontology graph is made and each of the previously stored patterns which are similar to it is selected for running algorithm on them. This is the assessment phase that the decision is based on whether the received Pattern is fraudulent or not. After the detection of similar patterns of input transactions, the matrix X is declared and then threshold parameter is initialized with performing algorithm steps, the distance between input transaction and patterns recorded previously in database is calculated. If this distance for a transaction is larger than the threshold, input transaction will be fraudulent. Another advantage of this method is detecting fraudulent transactions online and offline. It means that when an input transaction is received, its similar patterns are extracted and outlier mining is performed to avoid happening fraudulent transaction. If a fraudulent transaction that was diagnosed is the input transaction, the detecting process is online and otherwise the process is offline.

6.2. Data Set

In this study, 5,000 laboratory structured records have been produced automatically according to the specifications listed in Table 1. For evaluation of proposed method for detecting credit card frauds, different threshold is inspected for every fraudulent behaviour based on figure 3, can be seen that if the value of threshold is 12, the highest precision of detecting fraud occurs that this value is 89.4%. Also precision defines the ratio of real fraudulent transaction to estimated fraudulent transactions. With respect to this parameter, in various λ, False Positive (FP) is calculated that is shown in figure 4.

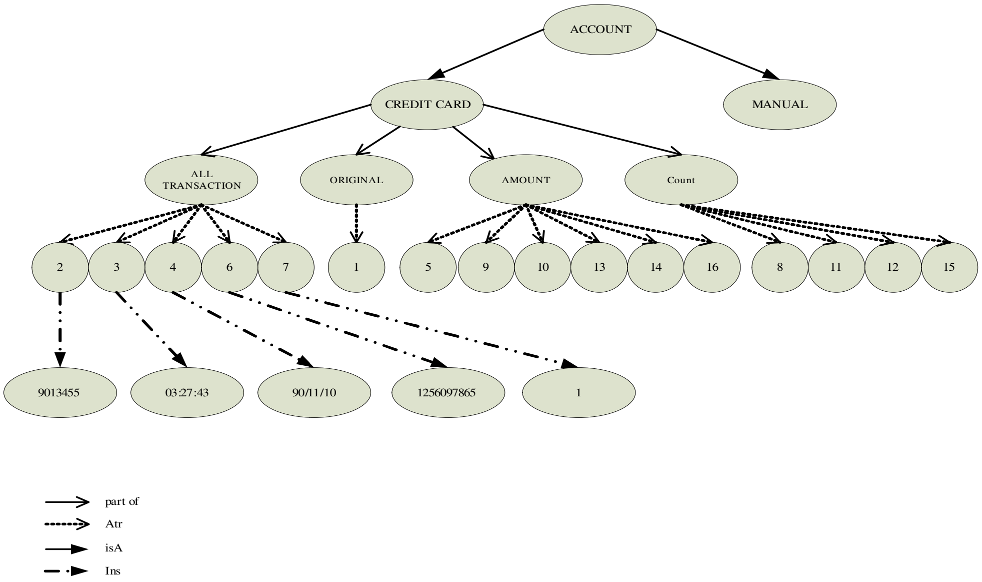

Figure 2. Relevant ontology for relationships between credit card transaction's data

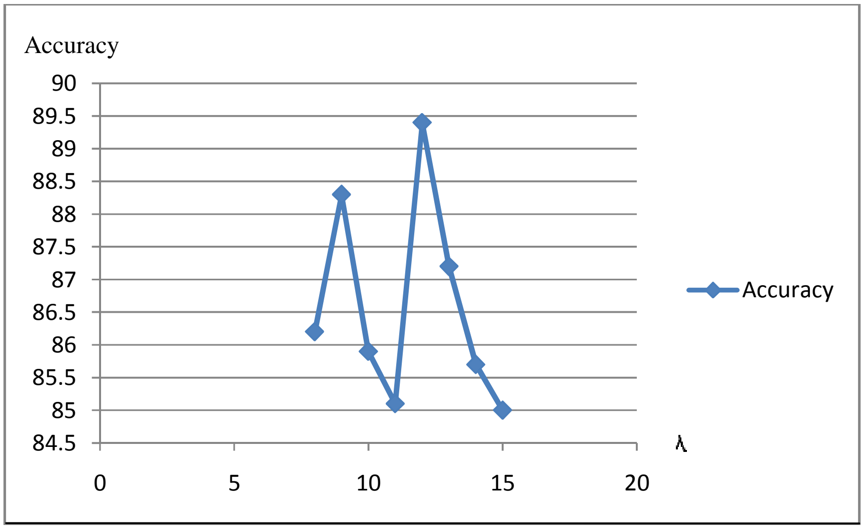

Figure 3. Correlation of threshold parameter and accuracy

Now, with consideration to this, when an input transaction received, the transaction is sent to a component that called Make Ontology unit in proposed framework (see figure 1) for evaluation that this component later than ontology graph construction of transaction, extracted other patterns that are similar to this transaction and exists before in pattern database, involves in fraudulent behaviour transaction detection. There are two main factors that are important in detection performance on this framework, the number of extracted similar patterns from pattern database and the effect of threshold parameter. Figure 5 shows the relationship between the number of extracted similar patterns to an input transaction and the precision of detection process whether online or offline is higher. As a conclusion, we explain that whatever the number of similar patterns to an input transaction is bigger and the value of threshold parameter is closed to 12, the total precision is higher. This solution can used banks, organizations and governmental canters for transaction management that prevent losses of this misuse or fraudulent behaviours.

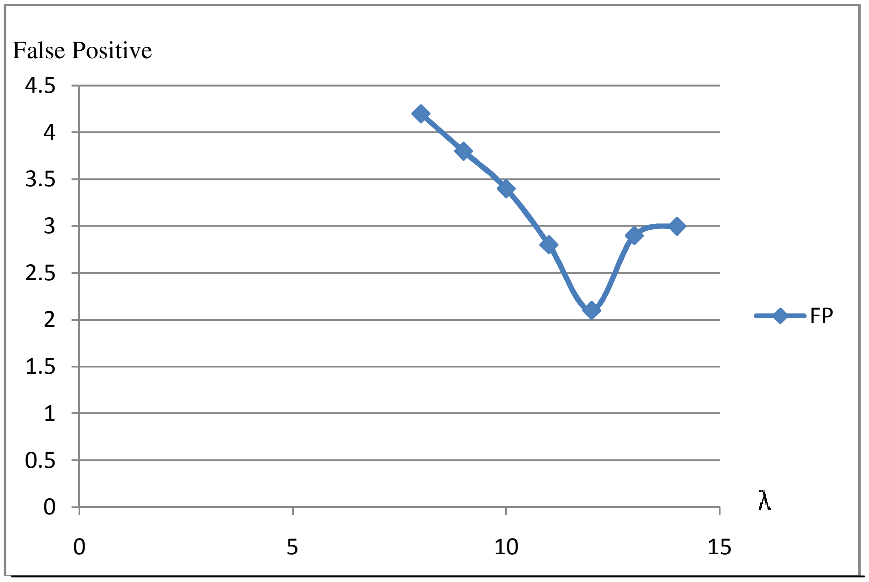

Figure 4. Correlation of threshold parameter and false positive

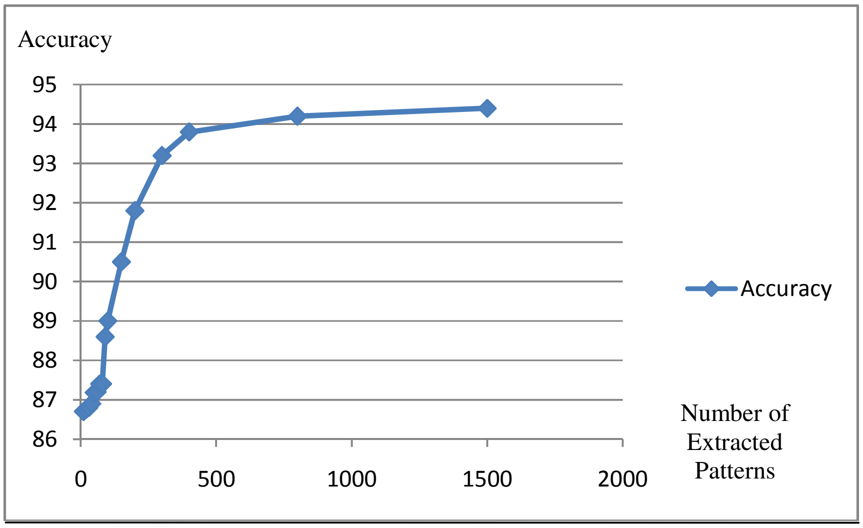

Figure 5. Correlation of the number of extracted patterns and accuracy

Finally this method proves accurate in predicting fraudulent transactions through outlier mining based on ontology concept emulation experiment of credit card transaction data set of one certain laboratory structured records. The experiment shows that outlier mining based on ontology 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.

8. CONCLUSIONS

In this paper, we intended to offer a model for fraud detection in credit cards on a semantic connection between data stored for every transaction fulfilled by a user basis and present it by ontology graph and store them then in patterns database. The main advantage from which detection process benefits are quickness of real-time detection and storage of only worthy behaviours models of credit card owner or someone else performing transactions. By this method in addition to real-time detection of fraud occurrences through running transactions, storage capacity of various data patterns is low because of no maintenance of similar patterns.

REFERENCES

[1] L. Delamaire, H. Abdou, and J. Pointon, “Credit Card Fraud Detection Techniques: A Review”, Banks and Banks Systems, 2009.

[2] A. Jay Harris, and D. Yen, “Biometric Authontication Assuring Access to Information”, Information Management & Computer Security, 2002.

[3] P. Chan, W. Fan, A. Prodromidis, and S Stolfo, “Distributed Data Mining in Credit Card Fraud etection”, IEEE Intelligent System, vol. 14, no. 6, pp. 67-74, 1999.

[4] T. Paul Bhatla, V. Prabhua, and A.Dua, “Understanding Credit Card Frauds”, 2003.

[5] S. Ghosh, and D.L. Reilly, “Credit Card Fraud Detection with a Neural-Network”, Proc. 27th HawaiiInternational Conference on System Sciences: Information Systems: Decision Support and Knowledge-Based Systems, vol. 3, pp. 621-630, 1994.

[6] 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 Engineering, pp. 220-226, 1997.

[7] S. Maes, K. Tuyls, B. Vanschoenwinkel, and B. Manderick, “Credit Card Fraud Detection Using Bayesian and Neural Networks”, First International NAISO Congress on Neuro Fuzzy Technologies, Havana, Cuba, 2002.

[8] M.J. Kim and T.S. Kim, “A Neural Classifier with Fraud Density Map for Effective Credit Card Fraud Detection,” Proc. International Conference on Intelligent Data Engineering and Automated Learning, Lecture Notes in Computer Science, Springer Verlag, no. 2412, pp. 378- 383, 2002.

[9] 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.

[10] C. Phua, V.Lee, K. Smith, and R. Gayler, “A Comprehensive Survey of Data Mining-Based Fraud Detection Research,”http://www.bsys.monash.edu.au/people/cphua/Mar.2007.

[11] S. Stolfo and A.L. Prodromidis, “Agent-Based Distributed Learning Applied to Fraud Detection,” Technical Report CUCS-014-99, Columbia Univ., 1999.

[12] C. Phua, D. Alahakoon, and V. Lee, “Minority Report in Fraud Detection: Classification of Skewed Data,” ACM SIGKDD Explorations Newsletter, vol. 6, no. 1, pp. 50-59, 2004.

[13] C.B. Necib, J. Freytag, “Ontology based query processing in database management systems”, Proceeding on the 6th international on ODBASE, pp. 37-99, 2003.

[14] A. Mazak, M. Lanzenberger and B. Schandl, “iweightings: Enhancing Structure-based Ontology Alignment by Enriching Models with Importance Weighting”, International Conference on Complex, Intelligent and Software Intensive Systems, pp. 992-997, 2010.

[15] K. Yamanishi, and J. Takeuchi, “A Unifying Framework for Detecting Outliers and Change Points from Non-Stationary Time Series Data”, In: SIGKDD 02 Edmonton, Alberta, Canada, 2002.

[16] F. Angiulli, and C. Pizzuti, ”Fast Outlier Detection in High Demensional Spaces”, Proceedings of the 6th European Conference on the Principles of Data Mining and Knowledge Discovery, 2006.

[17] W. Fang. YU, and N. Wang, “Research on Credit Card Fraud Detection Model Based on Distance Sum”, International Joint Conference on Artificial Intelligence, 2009.

Authors

Ali Ahmadian Ramaki
He was born in Iran on August 10, 1989. He recieved his BSc degree from University of Guilan, Iran in 2011. He is now MSc student at university of Guilan , Iran . His research interests in computer security, network security and intelligent intrusion detection.

Reza Asgari
Reza Asgari was born in Iran, Ghazvin. He received his BSc degree from university of Guilan, Iran in 2011. He is now MSc student at university of Guilan, Iran. His research interests in operating system and database security.

Reza Ebrahimi Atani
Reza Ebrahimi Atani received his BSc degree from university of Guilan, Rasht, Iran in 2002. He also received MSc and PhD degrees all from Iran University of Science and Technology, Tehran, Iran in 2004 and 2010 respectively. Currently, he is the faculty member and assistant professor at faculty of engineering, University of Guilan. His research interests in cryptography, computer security, network security, information hiding and VLSI design.