Investigation of Data Mining Techniques in Fraud Detection: Credit Card

R. Gayathri
Assistant Professor School of IT and Science
Dr. G.R.D. College of Science
Coimbatore

A. Malathi
Assistant Professor
PG and Research Department of Computer Science
Government Arts College, Coimbatore

ABSTRACT

In recent times the more secure data transfer takes place almost by means of internet. Apart from the corporate companies, publics also started using the network media. At the same time the risk also increases in secure data transfer. One of the major issue among them is credit card fraud detection systems which has a significant percentage of transactions labeled as fraudulent are in fact legitimate. Thus this may delay the fraudulent transaction detection. Due to ever increasing volumes of data needed to be analyzed using data mining methods and techniques which are being used more and more. The aim of this study is to analyze the five most frequently used classification techniques in fraudulent detection. Neural Network, Decision Tree, Naïve Bayes, K-nn and Support Vector Machine are taken in to consideration. This paper discusses on each techniques and their limitations. Still they suffer from the problem of false detection rate highly.

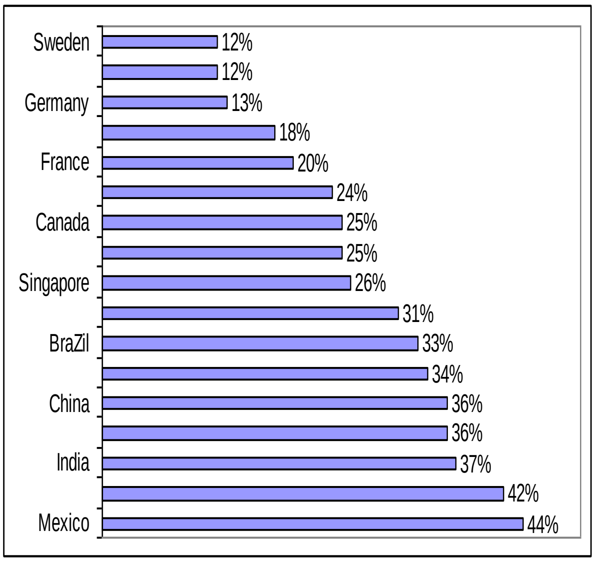
Keywords: Neural Networks, Decision Tree, Support Vector Machines, Credit Card Fraud Detection.

1. INTRODUCTION

The term fraud here refers to the mistreatment of a revenue organization's system without essentially leading to direct permissible consequences. In a aggressive environment, fraud can become a business serious problem if it is very widespread and if the deterrence procedures are not guaranteed. Fraud detection, being part of the overall fraud control, mechanizes and helps to diminish the manual parts of a transmission process. This area has turn into one of the most well-known data mining applications.

Timely information on fraudulent activities information is a main goal and a good strategy for banks and industries as well. Credit card fraud detection is the process of identifying those transactions that are fraudulent and partitioned these database into two classes of legitimate (genuine) and fraudulent transactions. Credit card frauds can be broadly classified into three categories such as internet, traditional card and, merchant related frauds [3, 5, 6, 7 & 8]. Further section of this paper discusses about some of the existing works involved in fraudulent detection.

Percentage of respondents who have experienced card fraud (N=5,114)



The above figure shows that according to the Aite Group, ACI Worldwide study of 5,223 consumers in 17 countries [9]. The percentage of respondents who have experienced both credit and debit card fraudulent in 17 different countries is shown. India holds the third place with 37% and the highest rate of fraudulent is in mexico with 44% and the lowest rate of fraudulent is in Sweden with 12%.

2. NEURAL NETWORK BASED CREDIT CARD FRAUDULENT DETECTION

Even though there exist a several fraud detection technology based on data mining or knowledge discovery, it is not possible to detect fraud while the transaction is in progress. This is due to less chance of fraudulent occurrence during its transaction. It has been seen that Credit card fraud detection has two highly peculiar characteristics.

Neural network based fraud detection is based totally on the principle of human brain. This technology has made a computer capable of think. From the past experience human brain will get trained and use its knowledge or experience in making the decision in daily life problem the same technique is applied with the credit card fraud detection technology. When a particular person uses their credit card, there is a standard pattern of its usage, which is made by the way consumer uses its credit card.

Limitations

Problem with neural networks is that a number of parameter [10] has to be set before any training can begin. Conversely, there are six clear rules how to situate these parameters. Up till now these parameters establishes the triumph of the training. In general, group of neurons will form a neural networks and each one has a number of inputs which are mapped to its related output. Networks differ in the way their neurons are interconnected, in the way the output of a neuron determined out of its inputs and in their temporal behavior.

The topology placed a major role on a network performance but, there is a lack of methods exists to determine the optimal topology for a given problem due to its high complexity of large networks. The preference of the necessary parameters like network topology, learning rate, initial weights are often previously determines the accomplishment of the training process.

3. DECISION TREE IN FRAUDULENT DETECTION

The proposal of a similarity tree using decision tree logic has been developed. A similarity tree is defined recursively nodes are labeled with attribute names, edges are labeled with values of attributes that assure some condition and 'leaves' that have an strength aspect which is defined as the ratio of the number of transactions that satisfy these condition(s) over the totality number of justifiable transaction in the performances[11]. The benefit of the method that is optional is that it is easy to implement, to recognize and to display. However, a shortcoming of this system is the necessities to check each operation one by one. Even so, resemblance trees have given proven results [12] also worked on decision trees and in particular on an inductive decision tree in order to launch an intrusion detection system, for another type of fraud.

As a substitute of classifying the given transaction is either legal or fraud. In [13] they discover the place of the customer through IP address. IP address traces the transaction location of the customer/merchants.

Limitations

Decision-tree learners can create over-complex trees that do not generalize well from the training data. The reliable information in the decision tree depends on providing the required internal and external information properly. Large changes can be made in a tree with even small changes incorporated in the input data. Variable change, without including the duplicate information, or sequence alteration midway can lead to major changes and might possibly require redrawing the tree.
Another fundamental flaw of the decision tree analysis is that the decisions contained in the decision tree are based on expectations, and thus these expectations lead to many errors in the decision tree. Although it follows a natural way by tracing relationships between events, contingencies which arise from a decision may not be possible and thus it in turn leads to bad decisions.

4. NAÏVE BAYES CLASSIFIER IN FRAUDULENT DETECTION

The Naïve Bayesian classifier is a influential probabilistic method that make use of class in sequence from training cases to forecast the class of prospect instances. This algorithm was first introduced by John and Langley [14] and is better in its speed of learning while preserving exact predictive power. Examinations on real-world data have frequently shown that the Naïve Bayesian classifiers perform comparably to more classy induction algorithms. Clark & Niblett (1989) showed that Bayesian classifiers attain similar accurateness levels compared to rule induction methods such as CN2 and ID3 algorithms in medical domains. John & Langley [14] show that by using kernel density estimation instead of a Gaussian distribution, the Naïve Bayesian classifier achieves equally as well and in some cases better than the decision tree algorithm C4.5. Still, this method goes by the name "Naïve" because it naively assumes independence of the attributes given the class. Classification is then completed by applying Bayes rule to work out the probability of the correct class given the particular attributes of the credit card transaction as in [14],

$$P(Fraud \mid Evidences) = \frac{P(Evidences \mid Fraud) * P(fraud)}{P(Evidences)}$$

Where $P(Fraud \mid Evidences)$ is the posterior probability; the probability of the hypothesis (the transaction being fraudulent) after considering the effect of the evidences (the attribute values based on training examples). *P(fraud)* is the a-priori probability; the probability of the hypothesis given only past experiences while ignoring any of the attribute values. $P(Evidences \mid Fraud)$ is called the likelihood.

5. K-NEAREST NEIGHBORS IN FRAUDULENT DETECTION

The k-Nearest Neighbour (kNN) method is a straightforward algorithm that provisions all available instances and classifies new cases based on a similarity measure. The kNN algorithm is an pattern of an instance-based learner. In a sense, all of the other learning methods are "instance-based," as well, for the reason that they start with a set of instances as the initial training in sequence. However, for instance-based learners the instances themselves are used to represent what is learned, rather than using the instances to infer a rule set or decision tree. The nearest-neighbour classification method is performed when the existing instances are compared with every new instance using a distance metric. If any existing instances are closer then that will get assign with the new one. Sometimes more than one nearest neighbor is used, and the majority class of the closest k neighbors or the distance weighted average, if the class is numeric is assigned to the new instance.

The concept of the instance-based nearest-neighbor algorithm was first introduced by [16]. Generally, the standard Euclidean distance is used when computing the distance between several numerical attributes. However, this assumes that the attributes are normalized and are of equal importance i.e., one of the major troubles in learning is to decide which are the vital features. For cases when nominal attributes are there, such as contrasting the attribute values of the types of credit cards.

Some attributes are more important than others, and this is usually reflected in the distance metric by some kind of attribute weighting. Deriving suitable attribute weights from the training set is a key problem in instance-based learning. In this technique the instances do not really "describe" the patterns in data. Though, the cases combine with the distance metric to carve out boundaries in instance space that distinguish one class from another, and this is a kind of explicit representation of knowledge.

Limitations

Assumption of class conditional independence usually does not hold. Dependencies among the attributes cannot be modeled by Naive Bayesian Classifier. If the sample size increases significantly it cannot be handled efficiently.

The traditional KNN classification has three limitations.

1. High calculation complexity: To find out the k nearest neighbor samples, all the similarities between the training samples must be calculated. When the number of training samples is less, the KNN classifier is no longer optimal, but if the training set contains a huge

number of samples, the KNN classifier needs more time to calculate the similarities.

2. Dependency on the training set: The classifier is generated only with the training samples and it does not use any additional data. This makes the algorithm to depend on the training set excessively; it needs recalculation even if there is a small change on training set;

3. No weight difference between samples: All the training samples are treated equally; there is no difference between the samples with small number of data and huge number of data. So it doesn't match the actual phenomenon where the samples have uneven distribution commonly.

6. SUPPORT VECTOR MACHINES IN FRAUDULENT DETECTION

The Support Vector Machines (SVM) algorithm was first introduced by [17]. This algorithm finds a special kind of linear model, the maximum margin hyper plane, and it classifies all training instances correctly by separating them into correct classes through a hyperplane (a linear model). The maximum margin hyperplanes the one that gives the greatest separation between the classes – it comes no closer to any of the classes than it has to. The instances that are closest to the maximum margin hyperplane – the ones with minimum distance to it – are called support vectors. There is always at least one support vector for each class, and often there are more [18].

Limitations

The biggest limitation of SVM lies in the choice of the kernel (the best choice of kernel for a given problem is still a research problem).

- A second limitation is speed and size (mostly in training - for large training sets, it typically selects a small number of support vectors, there by minimizing the computational requirements during testing).

- The optimal design for multiclass SVM classifiers is also a research area.

7. CRITIQUE OF EXISTING APPROACHES

In most cases of real time fraud detection, data mining is the best choice which is more need on the realistic issues of requirements, constraints, and commitment towards drop of fraud than the technical issues balanced by the data.

- There has not been any efficient empirical evaluation of commercial data mining tools for fraud detection.
- Lack of knowledge in handling incomplete dataset
- Though there is a tremendous growth in the internet transaction there is a lack of strong security to the high end.
- The identification of fraudulent in earlier is more significant in terms of cost analysis.

Comparative Analysis

A software was developed to analyze 335 references that were extracted from computer science databases. This program extracts information such as the name of authors, publication year, keywords, URL, etc. This analysis gives a general picture of the most repeated, and arguably the most popular keywords and topics in the pool of extracted references. The same program is used to find the distribution of papers per year (see figure 2). This figure shows that there has been a significant growth in the number of publications related to market manipulation in securities market during the past few years. It should be mentioned, that the pool of papers that we have, was extracted in september 2013. This expects the number of publications in 2013, to continue the increasing trend.

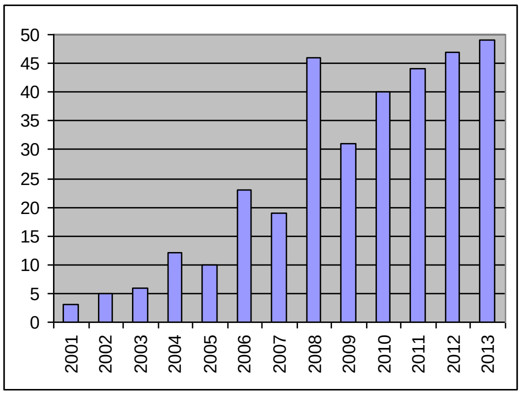


Figure 2. Distribution of 335 papers related to data mining for fraud detection in securities market per publication year

8. DISCUSSION

The primary objective of this paper is to define existing challenges in this domain for the different types of large data sets and streams. It categorizes, compares, and summarizes relevant data mining-based fraud detection methods and techniques in published academic and industrial research.

The second objective is to highlight promising new directions from related adversarial data mining applications such as epidemic or outbreak detection, intrusion detection, detection of spam, and terrorist. Knowledge and experience from these adversarial domains can be interchangeable and will help prevent repetitions of common mistakes and "reinventions of the wheel.

9. CONCLUSION & FUTURE WORK

In this survey paper we have explored and analyzed the credit card fraudulent detection. This paper takes into the account four different classification techniques which were most frequently used in the fraudulent detection using data mining based classification methods. Neural Networks, Decision Tree, K-NN, Naïve Bayes and support vector machine. Still they suffer from uncertainty in real world dataset which are not properly handled by these existing approaches. So our future work aims at developing a complete set of pattern recognition technique which overcomes the problem of missing values, handling voluminous data precisely and handling the incomplete dataset.

10. REFERENCES

[1] Elkan, C. (2001). Magical Thinking in Data Mining: Lessons from CoIL Challenge 2000. Proc. of SIGKDD01, 426-431.

[2] Lavrac, N., Motoda, H., Fawcett, T., Holte, R., Langley, P. & Adriaans, P. (2004). Introduction: Lessons Learned from Data Mining Applications and Collaborative Problem Solving. Machine Learning 57(1-2): 13-34.

[3] Brause, R., Langsdorf, T. and Hepp, M. (1999). Neural data mining for credit card fraud detection Proceedings 11th IEEE International Conference on Tools with Artificial Intelligence. tao guo, gui-yang li, neural data mining for

credit card fraud detection 978-1-4244-2096-4/08 ©2008 ieee, 3630, july 2008.

[4] Mirjana Pejic-Bach, Profiling intelligent systems applications in fraud detection and prevention: survey of research articles, 2010 International Conference on Intelligent Systems, Modeling and Simulation.

[5] Prabin Kumar Panigrahi, A Framework for Discovering Internal Financial Fraud using Analytics, International Conference on Communication Systems and Network Technologies 2011.

[6] Sahin, Y., Duman, E.: An overview of business domains where fraud can take place, and a survey of various fraud detection techniques. In: Proceedings of the 1st International Symposium on Computing in Science and Engineering, Aydin, Turkey (2010).

[7] V. Filippov L. Mukhanov B. Shchukin Credit Card Fraud Detection System.

[8] Y. Sahin, E. Duman "Detecting Credit Card Fraud by ANN and Logistic Regression" 2011.

[9] http://www.forbes.com/sites/halahtouryalai/2012/10/22/countries-with-the-most-card-fraud-u-s-and-mexico/.

[10] Raghavendra Patidar, Lokesh Sharma, Credit Card Fraud Detection Using Neural Network, NCAI2011, 13-14 May 2011, Jaipur, India, International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-1, Issue-NCAI2011, June 2011.

[11] Kokkinaki, A. (1997). On Atypical Database Transactions: Identification of Probable Frauds using Machine Learning for User Profiling. Proc. of IEEE Knowledge and Data Engineering Exchange Workshop, 107-113.

[12] Fan, W., Miller, M., Stolfo, S., Lee, W. & P Chan. 2001. Using Artificial Anomalies to Detect Unknown and Known Network Intrusions, Proc. of ICDM01; 123-248.

[13] Dr R.Dhanapal , Gayathiri.P, Credit Card Fraud Detection Using Decision Tree For Tracing Email And Ip, IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 5, No 2, September 2012.

[14] John, George H, and Pat Langley. "Estimating Continuous Distributions in Bayesian Classifiers.".

[15] Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence. SanMateo: Morgan Kaufmann Publishers, 1995. 338-345.

[16] Aha, David W., Dennis Kibler, and Marc K. Albert. "Instance-based learning algorithms." Machine Learning, 1991: 37-66.

[17] Cortes, Corinna, and Vladimir Vapnik. "Support-vector networks." Machine Learning, 1995: 273-297.

[18] Witten, Ian, and Eibe Frank. Data Mining: Practical Machine Learning Tools and Techniques. San Fransico: Elsevier, 2005.