ISSN: 2278 – 909X
International Journal of Advanced Research in Electronics and Communication Engineering (IJARECE)
Volume 6, Issue 6, June 2017

A review: Credit card fraud detection using various machines learning algorithm

Deepika kaushik[1] (M.Tech scholar)
Dr.Indu kashyap[2] (Associate professor)
Simple Sharma[3] (Associate professor)
Department of Faculty engineering technology (CSE)
MANAV RACHNA INTERNATIONAL UNVERSITY (HARYANA)

ABSTRACT

In present time, with the great increase in credit card transactions, credit card fraud has increasing excessively in recent years. Fraud detection includes monitoring of the spending behavior of users/ customers in order to determination, detection, or avoidance of undesirable behavior. As credit card becomes the most prevailing mode of payment for both online as well as regular purchase, fraud relate with it are also accelerating. Fraud detection is concerned with not only capturing the fraudulent events, but also capturing of such activities as quickly as possible. The use of credit cards is common in modern day society. Fraud is a millions dollar business and it is rising every year. Fraud presents significant cost to our economy worldwide. Modern techniques based on Data mining, Machine learning, Sequence Alignment, Fuzzy Logic, Genetic Programming, Artificial Intelligence etc., has been introduced for detecting credit card fraudulent transactions. This paper survey on data mining techniques can be combined successfully to obtain a high fraud coverage combined with a low or high false alarm rate.

Keywords: Data mining, credit card, fraud detection, Decision tree, FP-growth, KNN etc.

I. INTRODUCTION

For some time, there has been a strong interest in the ethics of banking (Molyneaux, 2007; George, 1992), as well as the moral complexity of fraudulent behavior (Clarke, 1994). Fraud means obtaining services/goods and/or money by unethical means, and is a growing problem all over the world nowadays.

Fraud deals with cases involving criminal purposes that, mostly, are difficult to identify. Credit cards are one of the most famous targets of fraud but not the only one; fraud can occur with any type of credit products, such as personal loans, home loans, and retail. Furthermore, the face of fraud has changed dramatically during the last few decades as technologies have changed and developed. A critical task to help businesses and financial institutions including banks is to take steps to prevent fraud and to deal with it efficiently and effectively, when it does happen (Anderson, 2007). Anderson (2007) has identified and explained the different types of fraud, which are as many and varied as the financial institution's products and technologies [11], as shown in Figure 1.

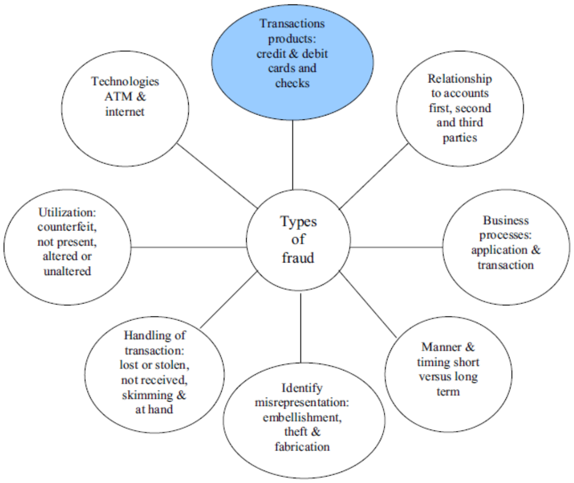
Transactions products: credit & debit cards and checks
Technologies ATM & internet
Relationship to accounts first, second and third parties
Utilization: counterfeit, not present, altered or unaltered
Types of fraud
Business processes: application & transaction
Handling of transaction: lost or stolen, not received, skimming & at hand
Identify misrepresentation: embellishment, theft & fabrication
Manner & timing short versus long term

Figure 1: Types of Fraud data

II. LITRATURE SURVEY

Credit card fraud detection has drawn a lot of research interest and a number of techniques, with special emphasis on neural networks, data mining and distributed data mining have been suggested.

Ghosh and Reilly [1] have proposed credit card fraud detection with a neural network. They have built a detection system, which is trained on a large sample of labeled credit card account transactions. These transactions contain example fraud cases due to lost cards, stolen cards, application fraud, counterfeit fraud, mail-order fraud, and no received issue (NRI) fraud.

Salvatore J. Stolfo, Wei Fan, and Wenke Lee [2] a distributed data mining system for credit card fraud detection is presented. In that system MADAM ID (Mining Audit Data for Automated Model for Intrusion Detection) is used to integrate cost based model with intrusion detection system to detect anomaly in credit card transactions a JAM project based model is used to provide better performance for fraud detection. In that technique cost model integrated with the data mining technique in distributed manner are used. Which provide an enhanced functionality to detect fraud in credit card transactions? But automated distribution of cost based trained data is not possible thus an automated distribution based technique is required to provide better performance to detect credit card frauds.

Stolfo et al. [3] suggest a credit card fraud detection system (FDS) using Meta learning techniques to learn models of fraudulent credit card transactions. Meta learning is a general strategy that provides a means for combining and integrating a number of separately built classifiers or models. A Meta classifier is thus trained on the correlation of the predictions of the base classifiers. The same group has also worked on a cost-based model for fraud and intrusion detection. They use Java agents for Meta learning (JAM), which is a distributed data mining system for credit card fraud detection. A number of important performance metrics like True Positive—False Positive (TP-FP) spread and accuracy have been defined by them.

Bell and Carcello [4] proposed a logistic regression model for estimating the likelihood of fraudulent financial reporting for an audit client. The model was conditioned on the presence or absence of several fraud-risk factors. The fraud risk factors identified in the final model included weak internal control system, rapid company growth, inadequate or inconsistent relative profitability, management that just want to achieve earnings projections anyhow while lying to the auditors or is overly evasive, company ownership status (public vs. private), and interaction term between a weak control environment and an aggressive management attitude towards financial reporting.

Kim and Kim[5] 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. Based on this observation, they use fraud density of real transaction data as a confidence value and generate the weighted fraud score to reduce the number of misdetections.

MubeenaSyeda, Yan-Qing Zbang and Yi Pan [6] a fast and efficient data mining technique for data mining and knowledge discovery of credit card fraud related data is presented. A parallel fuzzy neural network is used to train the dataset which contains data about credit card frauds in that technique multiple system works simultaneously to provide better performance to detect credit card frauds logs data of the various credit cards transactions are used to detect credit card frauds. But in that technique logs and updated logs are required to provide an enhanced credit card fraud detection mechanism which degrades the performance of the whole technique.

Philip K. Chan, Florida Institute of Technology Wei Fan, Andreas L. Prodromidis, and Salvatore J. Stolfo [7] a cost model based technique which uses data mining techniques in a distributed manner to provide an efficient mechanism to detect credit card frauds. In that technique dataset divided into various subsets and then data mining techniques are applied over these subsets to provide to generate classifiers for these subsets. In that way Meta classifiers are generated this provides an enhanced functionality to detect frauds in credit card transactions.

Chiu and Tsai [8] have proposed Web services and data mining techniques to establish a collaborative scheme for fraud detection in the banking industry. With this scheme, participating banks share knowledge about the fraud patterns in a heterogeneous and distributed environment. To establish a smooth channel of data exchange, Web services techniques such as XML, SOAP, and WSDL are used.

In 2002, Spathis et al. [9] proposed that statistical techniques like logistic regression may be suitable to develop a model to identify factors related to fraudulent financial statement. Nonparametric regression-based framework was used to run the falsified financial statement detection model. The proposed model was compared with discriminant analysis and logit regression methods for benchmarking.

Phua et al. [10] suggest the use of Meta classifier similar to in fraud detection problems. They consider naïve Bayesian, and Back Propagation neural networks as the base classifiers. A Meta classifier is used to determine which classifier should be considered based on skewness of data.Although they do not directly use credit card fraud detection as the target application, their approach is quite generic

III. TECHNIQUE USED IN CREDIT CARD FRAUD DETECTION

A. Decision tree algorithm

The decision tree is a structure that includes root node, leaf node & branch. Each internal node denotes a test on attribute, the outcome of test denotes each branch and the class label holds by each leaf node. The root node is the topmost node in the tree. The following decision tree is for concept to buy a computer, that denotes whether a customer at a company is likely to buy a computer or not. The test on the attribute represents each internal node. The all leaf node represents a class.

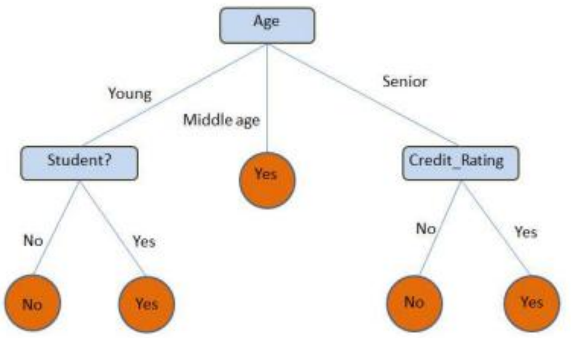

Figure 2: A Decision Tree

A. FP-Apriori

Frequent Pattern Mining (FPM) plays a key role to obtain associations and correlations among items in a large transactional dataset. A large number of algorithms have been proposed for frequent pattern generation. Almost all these algorithms are used for offline analytical task. With the increase in the demand of various real time business applications, online analysis is on demand. Fraudsters are coming up with new methods every day and between year 2012 and 2013, there has been nearly 15 percent increase of card frauds reported by cardwatch. Thus an incremental parallel frequent pattern mining techniques are essential to analyze the dynamically growing databases.Distributed frequent pattern mining algorithms to analyze the transactional data in the distributed database and to detect the online fraudulent transactions.

The popular algorithm Apriori1 forms the foundation for static frequent pattern mining. It generates candidate item sets iteratively, which makes the computational cost very high. Instead of using generate and test paradigm of Apriori, FP-tree approachesencode the dataset using a compact tree structure and directly extracts the frequent item sets from this structure. But it has to generate conditional pattern bases and sub-conditional pattern tree recursively. An interactive mining algorithm provides a lower and upper bound for its support for each set and generates frequent patterns in two database scans. Thus the user can interactively adjust the support and confidence at any time.

A dynamic algorithm facilitates incremental mining as well as interactive mining with one database scan. It keeps the entire transactions in the Tree for preparing frequent item sets; thus it requires more memory. It also provides incremental and interactive mining with less processing and I/O time but requires more memory to keep all combinations of items in the database.

B. K-Nearest Neighbor algorithm

The concept of nearest neighbor analysis has been used in several anomaly detection techniques. One of the best classifier algorithms that have been used in the credit card fraud detection is k-nearest neighbor algorithm that is a supervised learning algorithm where the result of new instance query is classified based on majority of K-Nearest Neighbor category. It was first introduced by Aha, Kibler, and Albert (1991) [12]. The performance of KNN algorithm is influenced by three main factors [Mohammed J. Islam]:

• The distance metric used to locate the nearest neighbors.

• The distance rule used to derive a classification from knearest neighbor.

• The number of neighbors used to classify the new sample. Among the various credit card fraud detection methods of supervised statistical pattern recognition, the K Nearest Neighbor rule achieves consistently high performance, without a priori assumptions about the distributions from which the training examples are drawn. K- Nearest neighbor based credit card fraud detection techniques require a distance or similar the measure defined between two data instances.

C. Neural Networks

Neural network is the need as a set of interconnected nodes designed to represent functioning of the human brain. Each node has a weighted connection to several other linked nodes in adjacent layers. Single node take input received from linked nodes and use the weights of the connected nodes together with easy function for computation of output values. Neural networks can be created for supervised and/or unsupervised learning. The user specifies the number of hidden layers along with the number of nodes within a specific hidden layer. On the other side, there are still many disadvantages for the neural networks, such as:

1. Difficulty to confirm the structure.
2. Excessive training
3. Efficiency of training and so on.

D. Naïve Bayes - Naïve Bayes is used as simple Probabilistic classifier based on Bayes conditional probability rule. Naïve Bayes follows strong (naive) statistical independence assumptions for the predictor variables. It is an effective classification tool that is easy to interpret and particularly suited when the dimensionality of the inputs is high. The efficiency of predicting financial fraud data was higher with no false positives with relatively low false negatives. Naïve Bayes methods are widely used in banking and financial fraud detection and claim fraud detection.

IV. CHALLENGES IN DATA MINING TO DETECT FRAUD:

• There are millions of transactions each day. To extract large amount of data from a database requires highly efficient techniques.

• The data or information is noisy.

• Data labels are not immediately available. Frauds or intrusions usually aware after they have already happened.

• It is hard to track user‟sbehaviours. All types of users (good users, business, and fraudsters) change their behaviors frequently

V. ADVANTAGE AND DISADVANTAGE OF VARIOUS MACHINE LEARNING ALGORITHM

In order to best comprise in fraud detection techniques we have been summarizes the advantages and disadvantages of the mentioned techniques, which demonstrated in Table1.

Techniques	Advantages	Disadvantages
Artificial Neural Network (ANN)	*Ability to learn from the past/lack of need to be reprogrammed/ Ability to extract rules and predict future activities based on the current situation/ High accuracy/ Portability/ high speed in detection/ the ability to generate code to be used in real-time systems/ the easiness to be built and operated/ Effectiveness in dealing with noisy data, in predicting patterns, in solving complex problems, and in processing new instances/Adaptability /Maintainability /knowledge discovery and data miming*	*Difficulty to confirm the structure/high processing time for large neural networks and excessive training/ poor explanation capability/ difficult to setup and operate/high expense/ non numerical data need to be converted and normalized/Sensitivity to data format.*
Artificial Immune System (AIS)	*High capability in pattern recognition/powerful in Learning and memory/Self-organization/ easy in integration with other systems/dynamically changing coverage/ self-Identity/ multilayered/ has diversity/ noise tolerance/ fault tolerance/ predator-prey dynamics/ Inexpensive / no need to training phase in DCA*	*Need high training time in NSA/ poor in handle missing data in ClonalG and NSA*
Genetic Algorithm	*Works well with noisy data/easy to integrate with other systems/ usually combined into other techniques to increase the performance of those techniques and optimize their parameters/ easy in build and operate/In expensive/fast in detection/ Adaptability/Maintainability/knowledge discovery and data miming*	*Requires extensive tool knowledge to set up and operate and difficult to understand.*
Hidden Markov Model (HMM)	*Fast in detection*	*Highly expensive/ low accuracy/not scalable to large size data sets*
Support Vector Machines (SVM)	*SVMs deliver a unique solution, since the optimality problem is convex/by choosing an appropriate generalization grade, SVMs can be robust, even when the training sample has some bias.*	*Poor in process largedataset/expensive/has low speed of detection/ medium accuracy/lack of transparency of results*
Bayesian Network	*High processing and detection speed/high accuracy*	*Excessive training need/ expensive*
Fuzzy Logic *Fuzzy Neural*	*Very fast in detection/good accuracy*	*Expensive*

Based System	*Network*		
	Fuzzy Darwinian System	*Very high accuracy/ Maintainability*	*Has very low speed in detection/ High expensive*
Expert System		*Easy to modify the KB/ easy to develop and build the system/ easy to manage complexity or missing information/high degree of accuracy/ explanation facilities/good performance/Rules from other techniques such as NN and DT can be extracted, modified, and stored in the KB.*	*Poor in handling missing information or unexpected data values/poor in process different data types /knowledge representation languages do not approach human flexibility/ poor in build and operate/ poor in integration*
Inductive logic programming (ILP)		*Powerful in process different data types/ powerful modeling language that can model complex relationships/powerful in handle missing data*	*Has low predictive accuracy/extremely sensitive to noise/ their performance deteriorates rapidly in the presence of spurious data.*
Case based reasoning (CBR)		*Useful in domain that has a large number of examples/ has the ability to work with incomplete or noisy data/effective/ flexible/ easy to update and maintain/ can be used in a hybrid approach.*	*May suffer from the problem of incomplete or noisy data.*
Decision tree (DT)		*High flexibility/good haleness/ explainable/easy to implement/easy to display and to understand*	*Requirements to check each condition one by one. In fraud detection condition is transaction.*

Table1. Advantages and disadvantages of fraud detection methods

V. CONCLUSION

Credit card fraud has become more and more widespread in recent years. Building an accurate, efficient and easy-handling credit card risk monitoring system is one of the chief tasks for the merchant banks for improving merchants risk management level in an automatic, scientific and effective way,. In this era of digital world, credit card is of extreme importance to financial organizations, institutions and companies. As credit card becomes the most accepted mode of payment for both online as well as regular purchase, cases of fraud associated with it are also increasing. For the purpose of reducing the bank's risk, various techniques have been employed. In this study, we characterize various fraud commitment and prevention methods as well. However model has been proposed for credit card fraud detection in catching the fraudulent transactions.

REFRERENCES

[1] R. J. Bolton and D. J. Hand"Unsupervised profiling methods for fraud detection" In conference of Credit Scoring and Credit Connol VII, EdinburghUK, Sept 5-7,2001

[2] Salvatore J. Stolfo, Wei Fan, Wenke Lee "Cost-based Modeling for Fraud and Intrusion Detection: Results from the JAM Project" IEEE, 2000.

[3] K. C. Cox, S. G. Eick, G. J. Wills, and R. J. Brachman. Visual data mining: Recognizing telephone calling fraud.J Data Mining and Knowledge Discover, 1(2):22>231, 1997.

[4] Bell, T., &Carcello, J. (2000). A decision aid for assessing the likelihood of fraudulent financial reporting.Auditing: A Journal of Practice & Theory, 9(1), 169– 178.

[5] X.D. Hoang, J. Hu, and P. Bertok, ―A Multi-Layer Model for Anomaly Intrusion Detection Using Program Sequences of System Calls,‖ Proc. 11th IEEE Int'l Conf. Networks, pp. 531-536, 2003

[6] MubeenaSyeda, Yan-Qing Zbang and Yi Pan, Parallel Granular Neural Networks for Fast Credit Card Fraud Detection,IEEE, 2002.

[7] Philip K. Chan, Florida Institute of Technology Wei Fan, Andreas L. Prodromidis, and Salvatore J. Stolfo, Columbia University "Distributed Data Mining in Credit Card Fraud Detection" November/December 1999.

[8] KhyatiChaudhary, JyotiYadav, BhawnaMallick, ―A review of Fraud Detection Techniques: Credit Card‖, International Journal of Computer Applications (0975 – 8887) Volume 45– No.1, May 2012.

[9] Spathis, C., Doumpos, M., &Zopounidis, C. (2002). Detecting falsified financial statements: a comparative study using multicriteria analysis and multivariate statistical techniques. The European Accounting Review, 11(3), 509–535.

[10] ―Credit card Fraud Detection with a neural network‖ by Ghosh and Reilly.IEEE‖ Proceedings of the Twenty Seventh Annual Hawaii International Conference on System Sciences,1994.

[11] Anderson, R. 2007. The Credit Scoring Toolkit: theory and practice for retail credit risk management and decisionautomation. New York: Oxford University Press.

[12] Joseph King-Fung Pun "Improving Credit Card Fraud Detection using a Meta-Learning Strategy". A thesis submitted in conformity with the requirements for the degree of Master of Applied Science Graduate Department of Chemical Engineering and Applied Chemistry University of Toronto. (2011).