Cost-Analysis of Credit Card Fraud Detection

Khyati Chaudhary, Bhashna Mallick
1,2Dept. of Computer Science, MTU, GCET, Greater Noida, UP, India

Abstract
With the growth in Information Technology and advancement in communication channels, fraud is spreading all over the world, results in enormous financial losses. Due to a rapid improvement in the Electronic Commerce, use of credit cards has been increasing in dramatic pace. Since credit card becomes the most popular and common mode of payment, credit card frauds are becoming increasingly epidemic in recent years. It caused an explosion in the credit card fraud as credit card usage is increasing day by day. With the increase of usage of credit card as the most popular mode of payment for both online as well as regular purchase, cases of fraud link with it are also rising drastically. In actual, fraudulent transactions has been scattered with the genuine transactions. As simple pattern matching techniques are not often sufficient to detect those frauds accurately and precisely. Implementation of efficient Fraud Detection Systems (FDS) has thereby become compulsory for all credit card issuing banks to minimize their losses. In this paper, we survey several novel approaches for Credit Card Fraud Detection (CCFD), which combines indications from current as well as past behavior of card holders. Receiver Operating Characteristic (ROC) analysis technology is also introduced which ensures the accuracy and effectiveness of fraud detection.

Keywords
Electronic Commerce, Credit Card Fraud, Artificial Neural Networks, Logistic Regression, Fraud Prevention Techniques

I. INTRODUCTION
Fraud is a million dollar business and is increasing every year. "Fraud refers to obtaining services/goods and money by some illegal way" [1]. Fraud deals with evidence which involve criminal motives that mostly are difficult to identify. The global economic crime survey of 2009 suggests that close to 30% of company worldwide reported fallen victim to fraud in the last year. Most of the online merchants who accept credit card payments sooner or later must have to deal with the so-called carders who steal credit card information to pay for orders in online stores. This type of illegal activity is called credit card fraud. Data mining is a field that is increasing its importance and availability of application day by day. A sub-domain of data mining, the anomaly detection is also rising in importance over the past few years. However a long time ago, the last 5 or 10 years, the uses of anomaly detection are also increasing, therefore making it a important technique to discover fraud, network intrusions, medicine side effects and many other useful anomalies within a wide set of data. Credit Card Fraud is one of increasing rampant for the online as well as offline users. Due to the dramatic increase of fraud which results in loss of billions of dollars worldwide each year; several modern techniques in detecting fraud are continually evolved and applied to many business fields. Fraud detection process involves monitoring the behavior of populations of users in order to estimate, detect, or avoid undesirable behavior. Many modern techniques based on Artificial Intelligence, Neural Network, Data mining, Fuzzy logic, Machine learning, Sequence Alignment, Genetic Programming etc., has been evolved in detecting various credit card fraudulent transactions. Every year business losses billions of dollars from credit card theft [2]. This paper presents a brief survey of various techniques used in credit card fraud detection mechanisms and evaluates each methodology based on certain design criteria. Fraud detection involves detect illegitimate usage of a communication network. Fraud is growing noticeably with the expansion of modern technology and the universal superhighways of communication, resulting in the loss of billions of dollars worldwide each year. The prevention of credit card fraud is an important application for prediction techniques.

Online banking and Electronic Commerce have been experiencing drastic growth over the past few years and show tremendous promise of growth even in the future. It has made it easier for fraudsters to indulge in new and different ways of committing credit card fraud over the Internet. Banks, organization and institutions try to optimize their detection systems in order to detect fraud and investigate suspect online behavior and transactions as soon as possible. Credit card fraud costs the industry about a billion dollars a year or 7 cents out of every $100 spent on plastic. Credit card fraud, a aggregated term for theft and fraud committed on any payment mechanism as a fraudulent resource of funds in a transaction(s). Currently, detecting credit card fraud is a quite difficult task when using normal process, so the development of the credit card fraud detection models, methods has become of importance whether in the academic or business organizations. However, role of fraud has been changed all of a sudden during the last few decades along with improvement of technologies. Credit Card is a plastic card issued to certain number of users as one of the mode of payment [3]. It allows the cardholders to purchasing goods and services based on the cardholder’s promise. In simple words, Credit Card Fraud (CCF) is defined as, “when an individual uses another individuals’ credit card for personal use while the owner of the card as well as the card issuer are not aware of the thing that the card is being used”. Number of systems/models, processes and preventive measures would help to prevent credit card fraud and reduce financial risks. Banks and credit card companies have aggregate huge amount of credit card account transactions. Secure credit services of banks and development of E-business, a reliable fraud detection system is essential to support safe credit card usage. Fraud detection based on analyzing existing purchase data of cardholder (present spending behavior) is a promising way for reducing the rate of credit card frauds. Fraud Detection Systems (FDS) come into scenario when the fraudsters exceed the fraud prevention systems and start fraudulent transactions illegally. With the developments in the Information Technology (IT) and improvements in the communication channels, fraud is spreading all over the world with results of large amount of fraudulent loss [4]. Credit card frauds could be processed in different ways such as simple theft, Never Received Issue (NRI), counterfeit cards, application fraud and Electronic fraud (where the card holder is not present). Anderson (2007) has identified and described the different types of fraud. Credit Card Fraud Detection (CCFD) is dreadfully typical, but also common problem for solution.

A. Credit Card Fraud
Credit card fraud is a term for theft and fraud committed using a credit card as a fraudulent source of funds in a transaction. The purpose might be to obtain goods without paying or to obtain unauthorized funds from an account [5]. According to

www.ijcst.com

International Journal of Computer Science and Technology

515
the Federal Trade Commission (FTC), in 2008 while identity theft had been holding steady for the past few years it saw a 21 percent increase. Surprisingly, United States (US), with its high number Credit Card transactions has a minimum fraud rate. Proper authorized users are permitted to do credit card transaction by using the several parameters such as signatures, credit card number, card holder’s address, expiry date and so on. Credit card fraud detection is quite secure and confidential. There are several popularly used fraud detection methods as ANNs (Artificial Neural Network(s)), rule-induction techniques, Support Vector Machines (SVM), Decision Trees, Logical Regression (LR), and meta-heuristics such as k-means Genetic Algorithms (GA), clustering, and nearest neighbor algorithms [6].

B. Cost Based Models for Credit Card Fraud Detection
Mostly, machine learning literature focuses on Model Accuracy (either training/supervised error or generalization error on hold out test data, True Positive/False Positive Rates (TPR/FPR(s)), or ROC analysis). Cost Model domain provides a different metric to evaluate performance of learned models. In the context of financial transactions, cost is measured in dollars. Because of the different dollar amounts of each credit card transaction along with other factors, the cost of failing to detect a fraud varies with each transaction. Therefore, the cost model for this domain is based on the sum and average of loss caused by fraud. Here, we demonstrate for a set of transactions T, a fixed overhead amount, and a fraud detector (or classifier) C:

\[
\text{THEN, Cumulative Cost (T; C; overhead) = } \sum_{t \in S} \text{Cost(C (t; overhead)) + Average Cost}(S; C; \text{overhead}) = \text{Cumulative Cost}(S; C; \text{overhead}) \text{ n where Cost(t; overhead) is the cost associated with transaction t and n is the total number of transactions in a test set T.}
\]

Here, the cost of a transaction is not only its “transaction amount”, but is also a function of an overhead amount. If the amount of a transaction is smaller than the overhead, it is not of any importance to investigate the transaction even if it is suspicious. For instance, if it takes 15 dollars to investigate a potential loss of one dollar, it is more economical not to investigate. Assuming a fixed overhead, we devised the cost model for each transaction t and classifier C. The overhead threshold, for good reason, is a closely guarded secret for important reasons and may vary over time. There are various cost enrolled with the detection of fraud as:

1. Damage Costs
The damage cost in cost model characterizes the amount of damage by an attack when anomaly detection is unavailable. The defined cost function per attack should be used to measure the cost of damage. This means, rather than simply measuring False Negative FN, as a rate of missed anomalies, rather we measure total loss based upon $DCost(s, a)$, which varies with the service(s) and the particular type of attack(a).

2. Challenge Costs
The challenge cost is the cost to act upon an alarm when there is indication of a potential intrusion. For Intrusion Detection System (IDS), one may consider suspending a suspicious connection and attempting to stop, by analyzing the service request (SR), if any system resources have been blocked from other legitimate users. As a first cut, these costs can be estimated by the amount of CPU

and disk resources needed to challenge a suspicious connection. In simple, instead of estimating the challenge cost for each intrusive connection, we can determine “average” the challenge costs to a single challenge cost per potential intrusive connection, i.e., overhead.

3. Operational Costs
The major issue in operational costs for IDS (Intrusion Detection System) is the amount of resources for extracting and test features from raw traffic data. Some of the features are costlier than others to gather and at the times, costlier features are more informative for detecting intrusions overhead, therefore clearly ignoring these attacks saves cost. Hence, for a true positive (TP), if overhead > $DCost(s; a)$, the intrusion is not challenged and the loss is $DCost(s; a)$. But if overhead < $DCost(s; a)$, the intrusion is challenged and the loss is limited to overhead. FP cost. When IDS falsely allege an event of being attack and the attack type is regarded as high cost, a challenge will ensue. Naturally, when evaluating IDS we should concern with measuring this loss. For this, we define the loss is just overhead for False Positive (FP), True Negative (TN) cost. IDS correctly decide that a connection is normal and truly not an attack. Therefore, as far we have only considers costs that depend on the outcome of IDS, we now put together the operational cost, OpCost. In this point of OpCost which measures the cost of computing values of characteristics in the IDS. We indicate OpCost(c) as the operational cost for a connection, c. Now, we can describe the cost-model for IDS. When evaluating an IDS over some test set S of well-labeled connections, c $\in$ S.

We can define the cumulative cost (S) for a detector as follows:

\[
\text{CumulativeCost}(S) = X c \in S \text{Cost(c) + OpCost(c)}
\]

where Cost(c) is defined analogous to the credit card case. Hereby, s is the service requested by connection c and “a” is the attack type detected by the IDS for the connection. Thereby, in realistic contexts, in order to minimize CumulativeCost(S), we need to investigate the trade off between OpCost(c) and Cost(c) in given Equation 1.

4. Shipping Cost
Mostly, it is relevant in a Card-Not-Present (CNP) scenario. Since the shipping cost is usually grouped in the value of the order, then the merchant will also need to absorb the cost of shipping for goods sold in a various fraudulent transaction. Further, fraudsters typically request high-priority shipping for their requested orders to enable rapid completion of the fraud, resulting in high shipping costs.

5. Card Association Fees
Visa and MasterCard have taken fairly strict programs that penalize merchants generating excessive charge backs. Generally, if a merchant exceeds established chargeback rates for any three-month period, the merchant could be penalized with a fee for every chargeback.

6. Merchant Bank Fees
In addition to the penalties charged by card associations, the merchant has to pay an additional processing fee to the acquiring bank for every chargeback claimed.
7. Administrative Cost

Every transaction which generates a chargeback requires important administrative costs for the merchant. On average process, each chargeback requires one to two hours to process.

C. Credit Card Fraud Prevention Technologies

While fraudsters are using complex methods to gain access to credit card information and commit fraud, new technologies are available to help merchants to detect as well as prevent fraudulent transactions [7]. Each technique provides some incremental value in terms of detection ability. Fraud detection technologies enable merchants, organizations and banks to perform highly automated screenings of incoming transactions and flagging all suspicious transactions.

D. Fraud Prevention Techniques

A. Review Manually

This method consists of reviewing every transaction manually having signs of fraudulent activity. This can be very expensive as well as time consuming. Furthermore, sometimes Manual Review (MR) is unable to detect some of the patterns of fraud such as use of a single credit card multiple times on multiple locations (physical or web sites) in a short span.

B. Address Verification System

This technique is applicable in Card-Not-Present (CNP). Address Verification System (AVS) matches the record with the card issuers by first few digits of the street address and the ZIP code information given for delivering/billing the purchase to the corresponding information.

C. Card Verification Methods

Card Verification Methods (CVM) consists of a 3- or 4-digit numeric code printed on the card but is not available in the magnetic stripe. The merchant can request the cardholder to provide this numeric code in case of card-not-present (CNP) transaction and submit it with authorization. The purpose of CVM is to ensure that the person submitting the transaction is an intended of the actual card, since the code cannot be copied from receipts or skimmed from magnetic stripe. Further, fraudsters who have temporary possession of a card could, read and copy the CVM code.

D. Negative and Positive Lists

A Negative List (NL) is a database used to identify high-risk transactions based on specific data fields. A merchant can build negative lists in a similar way based on billing names, street addresses, E-Mails and Internet Protocols (IPs) that have resulted in fraud or attempted fraud. A acquirer/merchant could create and submit it with authorization. The purpose of CVM is to ensure that the person submitting the transaction is an intended of the actual card, since the code cannot be copied from receipts or skimmed from magnetic stripe. Further, fraudsters who have temporary possession of a card could, read and copy the CVM code.

E. Payer Authentication

Payer authentication (PA) is an emerging technology that promises to bring in a new level of security to Business-to-Consumer (B2C) Internet commerce. In 2002, the first implementation of this type of service is Visa Payer Authentication Service (VPAS) program, launched worldwide by Visa. This program is based on a Personal Identification Number (PIN) associated with the card and a secure direct authentication channel between the consumer and the issuing bank. The PIN is issued by the bank when the cardholder enrolls for the card with the program and will be used to authorize online transactions exclusively.

F. Lockout Mechanisms

Automatic card number generators represent one of the new technological tools utilized by fraudsters frequently. The traits of frauds initialized by a card number generator are the following:

- Multiple transactions with similar card number (e.g. same Bank Identification Number (BIN))
- A large number of declines acquire banks/merchant sites can put them in-place prevention mechanisms specifically designed to detect number generator attacks.

II. Related Work

Mostly work for fraud detection or anomaly detection was classification based. Some of the algorithms are utilized for fraud detection are NB (naïve Bayesian), proposed by Elkan et al, C4.5, by Quinlan and BP (Back-propagation). The researches of Elkan, Domingos and Witten NB algorithm is very efficient in many real world data sets and is extremely effective. However, when there are extra attributes and are not normally distributed, the predictive accuracy is proportionally reduced. Hereby, C4.5 can output not only accurate predictions but also explains the patterns, Rule Set and Decision Tree in it. Back Propagation (BP) neural networks, not only process very large number of instances but also have a high tolerance to noisy data [8]. The BP (Back Propagation) algorithm requires long training times and also has extensive testing of parameters. For solving problem of fraudsters, a fraud detection system (FDS) is proposed to detect fraudulent transactions in an online system. It profiles user behavior pattern dynamically. Currently, many fraud detection techniques involve screening of transactions for tracking of customer behavior and spending patterns are being deployed by both banks as well as merchant companies. Some of the most common techniques include Address Verification Systems (AVS), Rule-based systems, Card Verification Method (CVS), Personal Identification Number (PIN), and Biometrics. Neural Networks which are capable of being “trained” can derive patterns out of data and are “adaptive” to changing schemes of fraud. Analysis shows that review of 2.0% of transactions can result in reducing fraud losses. The primary key to minimize total costs is to categorize transactions as well as review only the potentially fraudulent cases. It should involve deployment of a step-by-step authentication, screening, filtering and review mechanism. A second level of screening can involve comparison with positive and negative lists of customers, IP address, geographical regions, etc. This would not only filter out transactions in every step but also a few transactions will require manual review. Such type of solution can reduce the overall processing time as well as total costs. Ghosh and Reilly (1994) have proposed credit card fraud detection with a neural network. In this, they have built a Fraud Detection System (FDS) which is trained on a large sample of labeled credit card account transactions. It is possible to achieve a reduction of from 20% to 40% in total fraud losses. Aleskerov and Freisleben (1997) present CARDWATCH, a database mining system used for credit card fraud detection (CCFD). The system uses neural network to train specific historical consumption data and generate relevant neural network model [9]. The model was adopted to detect fraudulent cases. Kim and Kim have identified skewed distribution of data as well as mixture of legitimate and fraudulent transactions as the
two main reasons for the complexity of credit card fraud detection. Sam and Karl (2002) suggest a credit card fraud detection system (CCFDS) using Bayesian and Neural Network techniques to learn models of fraudulent credit card transactions. Based on the observations, use of fraud density of real transaction data as well as confidence value which generate corresponding weighted fraud score to reduce the number of misdetections. The threshold could not be adjusted dynamically based on frequency of fraudulent transaction.

### III. Research Methodology

#### A. Data set

As it is very difficult to obtain available credit card data sets because of the security, privacy and also cost issues. Presently, mostly researchers generate realistic synthetic data using data generator to facilitate the development and testing of data mining tool. All data sets have been saved in database for future preprocessing. This part of function is done through Bayesian networks. It is necessary for Bayesian network to differentiate genuine as well as fraudulent transactions. The original set of transactions was sampled in such a way that all fraud transactions were included, while a sample of the good transactions was chosen so as to have ratio of roughly 100 good transactions for each fraudulent transaction in the training data set. Credit card payment-related training data attributes include: time (Time), location (Cy_pq), type of merchandise (Mer_type), transaction amount for consumption (P_m business code for merchandise (Bu_code), business type for merchandise (Bu_type). Actual target value 1 represents abnormal data and 0 represents normal.

#### IV. Methods

### A. Neural Networks

Neural networks architectures or topologies formed by organizing nodes into layers and by linking these layers of neurons with some modifiable weighted interconnections (Rumelhart, 1986). Recently, Neural Network (NN) researchers have associated methods from numerical and statistics analysis into their networks. A non-linear mapping relationships from the input space to output space. Neural Networks can learn from the given cases and summarize the internal principles of data even without knowing the potential data principles. As well, it can adapt its own behavior to the new environment with the results of formation of general capability of evolution from present situation to the new environment. From the concept of the pure theory, the nonlinear neural networks method is superior to those with statistical methods in the application for credit card fraud detection. Sometimes unusual, even though the common advantages of the neural networks as a possible result of usage of improper network structure and learning computing method. On the other side, there are many disadvantages for the neural networks, such as the efficiency of training, difficulty to confirm the structure excessive training, and so on. It can adapt its own behavior to the new environment with the results of formation of general capability of evolution from present situation to the new environment. In this system, we use multi-layer neural network model and back propagation (BP) algorithm runs on the network. Back propagation learns by iteratively processing a given data set of training tuples \( X = \{x_1, x_2, \ldots, x_n\} \), comparing the network’s prediction for each tuple with the actual known target value. Some modifications are made in the backwards direction, that is, from the output layer \( Y = \{y_1, \ldots, y_m\} \), through each hidden layer down to the first hidden layer.

### B. Logistic Regression

Statistical models were applied at data mining tasks include analysis of regression, Logistic Regression, multiple discriminant analysis, and Probit method, etc (Altman, Marco 1994; Flitman, 1997). Logistic Regression (LR) is very useful for situations in which we want to be able for prediction of the presence or absence of a characteristic or outcome based on some pre-defined values of a set of predictor variables. It is similar to a Linear Regression (LR) model but is also suited to models where the dependent variable is contradictions. Logistic Regression (LR) coefficients can be used for estimation odds ratios for each of the Independent Variables in the model and it is applicable to a broader range of research situations than discriminant analysis. Linear probability and some multivariate conditional probability models (Logit and Probit) were introduced to the business failure literature prediction and the contribution of these methods was in estimation of the odds of a firm’s failure with probability (Ohlson, 1980; Martin, 1997). Logistic regression is a well-established statistical method for predicting binomial or multinomial outcomes. Multinomial Logistic Regression algorithm (MLR) could produce models when the target field is a set field with two or more possible values.

### References


Khyati Chaudhary did B.Tech.in Computer Science & Engg. from Uttar Pradesh Technical University, Lucknow and currently pursuing M.TECH. from Mahamaya Technical University, Noida. Her research area is Data Mining.

Bhawna Mallick received the B.Tech. in Computer Technology from Nagpur University, Nagpur, India and M.Tech. in Information Technology from Punjab University, Patiala, India. She is currently, pursuing the Ph.D. degree and working as Head & Associate Professor, Department of Computer Science & Engineering at Galgotias College of Engineering & Technology, Greater Noida affiliated to UP Technical University, India. She has 14 years of academic and industry experience with organizations like Infosys Technologies Ltd, Chandigarh, India and NIIT Technologies Ltd, New Delhi, India. She is member of IEEE (Institute of Electrical and Electronics Engineers), USA. Her area of research is Data Mining focusing on Sequential Mining of Progressive Databases. She has several publications in International Journals and Conferences.