International Journal on Applications in Engineering and Technology
Volume 3: Issue3 : December 2017, pp 7 -16 www.aetsjournal.com *ISSN (Online) : 2455 -0523*

A novel approach for credit card Fraud recognition using tool and genetic algorithm in Distributed data mining

Sridharan.K, Dr. Komarasamy.G

Abstract— Due to enormous growth of E-Commerce, credit card usage for online purchases has dramatically increased and it caused an outburst in the credit card fraud. Credit card becomes the mode of payment for both online as well as regular purchase. Credit card fraud is a serious and growing problem. In real life, fraudulent transactions are spread with genuine transactions and simple pattern matching techniques are not often enough to detect frauds accurately. Implementation of effective fraud detection systems has thus become necessary for all credit card issuing banks to minimize their losses. Various techniques like Data mining, Fuzzy logic, Machine learning etc., has evolved in detecting various credit card fraudulent transactions. In this paper we evaluate two advanced data mining approaches, support vector machines and random forests, together with the well-known logistic regression, as part of an attempt to better detect credit card fraud.

Keywords -- Electronic Commerce, Credit card fraud, Logistic regression, Random forests, Genetic Algorithm.

I. INTRODUCTION

In today's increasingly electronic society and with the rapid advances of electronic commerce on the Internet, the use of credit cards for purchases has become convenient and necessary Data mining approaches for credit card fraud detection are relatively few, possibly due to the lack of available data for research. Credit card transactions have become the major standard for Internet and Web based e-commerce. Intuitively, when banks lose money because of credit card fraud, cardholders pay for all of that loss through higher interest rates, higher fees, and reduced benefits. The credit card fraud-detection domain presents a number of challenging issues for data mining:

Sridharan.K, *Research scholar, Anna University, (Email :* Srijik77@gmail.com)

Dr. Komarasamy.G , *Assistant Professor (senior Grade), Department of CSE, Bannari Amman Institute of Technology, India.(Email :* gkomarasamy@gmail.com)

There are billions of credit card transactions processed per day. Mining huge amount of data requires highly effective techniques that scale.

The data are skewed—many more transactions are legal than fraudulent. Typical accuracy-based mining techniques can generate highly accurate fraud detectors by simply predicting that all transactions are legitimate, although this is equivalent to not detecting fraud at all.

Each transaction record has a different dollar amount and thus has a variable potential loss, rather than a fixed misclassification cost per error type, as is commonly assumed in cost-based mining techniques Support vector machines and random forests are sophisticated data mining techniques which have been noted in recent years to show superior performance across different applications. It examines aggregation over different time periods on two real-life datasets and finds that aggregation can be advantageous, with aggregation period length being an important factor. Aggregation was found to be especially effective with random forests. Random forests were noted to show better performance in relation to the other techniques, though logistic regression and support vector machines also performed Single decision tree models, though popular in data mining application for their simplicity and ease of use, can have instability.

Random forests combine the random subspace method with bagging to build an ensemble of decision trees. They are simple to use, with two easily set parameters, and with excellent reported performance noted as the ensemble method of choice for decision trees. They are also computationally efficient and robust to noise. Various studies have found random forests to perform favorably in comparison with support vector machine and other current techniques and reliability issues. logistic regression. It is well-understood, easy to use, and remains one of the most commonly used for data-mining in practice. It thus provides a useful baseline for comparing performance of newer methods supervised learning methods for fraud detection face two challenges.

The first is of unbalanced class sizes of legitimate and fraudulent transactions, with legitimate transactions far outnumbering fraudulent ones. For model development, some form of sampling among the two classes is typically used to obtain training data with reasonable class distributions. Various sampling approaches have been proposed in the literature, with random oversampling of minority class cases and random under sampling of majority class cases being the simplest and most common in use.

The second problem in developing supervised models for fraud can arise from potentially undetected fraud transactions, leading to mislabeled cases in the data to be used for building the model. For the purpose of this study, fraudulent transactions are those specifically identified by the institutional auditors as those that caused an unlawful transfer of funds from the bank sponsoring the credit cards .

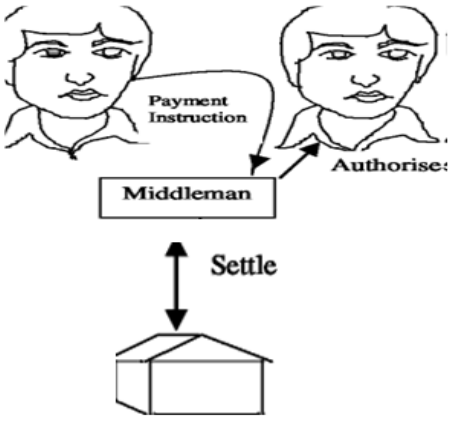

Fig 1 General Model of Internet Transaction

To deal with credit card fraud, credit card fraud prevention and credit card detection techniques are employed. Prevention approaches include fluorescent fibers, multitone drawings, watermarks, laminated metal strips and holographs on banknotes etc. while detection methods comes into picture when fraud prevention fails. In committing fraud, the range of fraudsters highly varies, some may be masters in doing so and some may be newcomers. For dealing with the masters the detection techniques must be updated constantly as fraudsters are quite prepared enough to penetrate the present detection methods. While for newcomers, the existing methods may work well. So a balanced approach is expected for the purpose of detection of frauds.

II. CREDIT CARD FRAUD

Credit card fraud can be defined as "Unauthorized account activity by a person for which the account was not intended. Operationally, this is an event for which action can be taken to stop the abuse in progress and incorporate risk management practices to protect against similar actions in the future". In simple terms, Credit Card Fraud is defined as when an individual uses another individual s credit card for personal reasons while the owner of the card and the card issuer are not aware of the fact that the card is being used. And the persons using the card has not at all having the connection with the cardholder or the issuer and has no intention of making the repayments for the Purchase they done.

Credit card fraud is a wide-ranging term for theft and fraud committed using a credit card or any similar payment mechanism as a fraudulent source of funds in a transaction. The purpose may be to obtain goods without paying, or to obtain unauthorized funds from an account. Credit card fraud is also an adjunct to identity theft. According to the Federal Trade Commission, while identity theft had been holding steady for the last few years, it saw a 21 percent increase in 2008. However, credit card fraud, that crime which most people associate with ID theft, decreased as a percentage of all ID theft complaints for the sixth year in a row.

The cost of card fraud in 2006 was 7 cents per 100 dollars worth of transactions Due to the high volume of transactions this translates to billions of dollars. In 2006, fraud in the United Kingdom alone was estimated at £535 million, or US$750–830 million at prevailing 2006 exchange rates .

III. TYPES OF FRAUD

There are many ways in which fraudsters execute a credit card fraud. As technology changes, so does the technology of fraudsters, and thus the way in which they go about carrying out fraudulent activities. Frauds can be broadly classified into three categories, i.e., traditional card related frauds, merchant related frauds and Internet frauds. The different types of methods for committing credit card frauds are described below.

A. Merchant Related Frauds

Merchant related frauds are initiated either by owners of the merchant establishment or their employees.

The types of frauds initiated by merchants are described below:

1) Merchant Collusion:

This type of fraud occurs when merchant owners or their employees conspire to commit fraud using the cardholder accounts or by using the personal

information. They pass on the information about cardholders to fraudsters.

2) Triangulation:

Triangulation is a type of fraud which is done and operates from a web site. The products or goods are offered at heavily discounted rates and are also shipped before payment. The customer while browse the site and if he likes the product he place the online information such as name, address and valid credit card details to the site.

When the fraudsters receive these details, they order goods from a legitimate site using stolen credit card details. The fraudsters then by using the credit card information purchase the products.

Bankruptcy fraud is the use of credit report from credit bureau as a source of information regarding the applicants' public records as well as a possible implementation of a bankruptcy model. Bankruptcy fraud is one of the most difficult types of fraud to predict. However, some methods or techniques may help in its prevention. Bankruptcy fraud means using a credit card while being insolvent. In other words, purchasers use credit cards knowing that they are not able to pay for their purchases. The bank will send them an order to pay. However, the customers will be recognized as being in a state of personal bankruptcy and not able to recover their debts.

B. Internet Related Frauds

The internet is the base for the fraudsters to make the frauds in the simply and the easiest way. fraudsters have recently begun to operate on a truly transnational level. With the expansion of trans-border, economic and political spaces, the internet has become a new worlds market, capturing consumers from most countries around the world. The below described are most commonly used techniques in Internet fraud:

1) Site Cloning:

Site cloning is where fraudsters close an entire site or just the pages from which the customer made a purchase. Customers have no reason to believe they are not dealing with the company that they wished to purchase goods or services from because the pages that they are viewing are identical to those of the real site. The cloned site will receive these details and send the customer a receipt of the transaction through the email just as the real company would do. The consumer suspects nothing, while the fraudsters have all the details they need to commit credit card fraud.

2) False Merchant Sites:

Some sites often offer a cheap service for the customers. That site requests the customer to fill his complete details such as name and address to access the webpage where the customer gets his required products.

Many of these sites claim to be free, but require a valid credit card number to verify an individual s age. These kinds of sites in this way collect as many as credit card details. The sites themselves never charge individuals for the services they provide. The sites are usually part of a larger criminal network that either uses the details it collects to raise revenues or sells valid credit card details to small fraudsters.

3) Credit Card Generators:

These are the computer programs that generate valid credit card numbers and expiry dates. These generators work by generating lists of credit card account numbers from a single account number. The software works by using the mathematical Luhn algorithm that card issuers use to generate other valid card number combinations. This makes the user to allow to illegally generating as many numbers as he desires, in the form of any of the credit card formats.

Credit card fraud is essentially of two types: application and behavioral fraud. Application fraud is where fraudsters obtaining new cards from issuing companies using false information or other people's information. Behavioral fraud can be of four types: mail theft, stolen/lost card, counterfeit card and 'card holder not present' fraud. Mail theft fraud occurs when fraudsters intercept credit cards in mail before they reach cardholders or pilfer personal information from bank and credit card statements. Stolen/lost card fraud happens when fraudsters get hold of credit cards through theft of purse/wallet or gain access to lost cards. However, with the increase in usage of online transactions, there has been a significant rise in counterfeit card and 'card holder not present' fraud. In both of these two types of fraud, credit card details are obtained without the knowledge of card holders and then either counterfeit cards are made or the information is used to conduct 'card holder not present' transactions, i.e. through mail, phone, or the Internet. Card holders information is obtained through a variety of ways, such as employees stealing information through unauthorized 'swipers', 'phishing' scams, or through intrusion into company computer networks. In the case of 'card holder not present' fraud, credit cards details are used remotely to conduct fraudulent transactions focus on theft fraud and counterfeit fraud, which are related to each other.

Theft fraud means using a card that is not yours. The perpetrator will steal the card of someone else and use it as many times as possible before the card is blocked. The sooner the owner will react and contact the bank, the faster the bank will take measures to stop the thief. Similarly, counterfeit fraud occurs when the credit card is used remotely; only the credit card details are needed. At one point, one will copy your card number and codes and use it via certain web-sites, where no signature or physical cards are required. Recently, Pago, one of the leading international acquiring & payment service providers, reveals in its Pago Report (2005) that credit card fraud is a growing threat to businesses selling goods or services through the internet. On-line merchants are at risk because they have to offer their clients payment by credit card. In cases where fraudsters use stolen or manipulated credit card data the merchant loses money because of so-called "charge-backs".

Application fraud is when someone applies for a credit card with false information. To detect application fraud, the solution is to implement a fraud system that allows identifying suspicious applications. To detect application fraud, two different situations have to be distinguished: when applications come from a same individual with the same details, the so-called duplicates, and when applications come from different individuals with similar details, the so called identity fraudsters.

The evolution of credit card fraud over the years is chronicled In the 1970s, stolen cards and forgery were the most prevalent type of credit card fraud, where physical cards were stolen and used. Later, mail-order/phone-order became common in the '80s and '90s. Online fraud has transferred more recently to the Internet, which provides the anonymity, reach, and speed to commit fraud across the world. It is no longer the case of a lone perpetrator taking advantage of technology, but of well-developed organized perpetrator communities constantly evolving their techniques.

Boltan and Hand note a dearth of published literature on credit card fraud detection, which makes exchange of ideas difficult and holds back potential innovation in fraud detection. On one hand academicians have difficulty in getting credit card transactions datasets, thereby impeding research, while on the other hand, not much of the detection techniques get discussed in public lest fraudsters gain Knowledge and evade detection. Credit card transaction databases usually have a mix of numerical and categorical attributes. Transaction amount is the typical numerical attribute, and categorical attributes are those like merchant code, merchant name, date of transaction etc. Some of these categorical variables can, depending on the dataset, have hundreds and thousands of categories. This mix of few numerical and large categorical attributes have spawned the use of a variety of statistical, machine learning, and data mining tools .We faced the challenge of making intelligent use of numerical and categorical attributes in this study. Several new attributes were created by aggregating information in card holders' transactions over specific time periods. The quicker a fraud gets detected, the greater the avoidable loss. However, most fraud detection techniques need history of card holders' behavior for estimating models.

In most banks, to be eligible for a credit card,applicants need to complete an application form.This application form is mandatory except for social fields. The information required includes

1. Identification information, location information,
2. Contact information, confidential information and
3. Additional information. Recurrent information

available would be for identification purposes, such as the full name and the date of birth. The applicant would inform the bank about his/her location details: the address, the postal code, the city and the country. The bank would also ask for contact details, such as e-mail address, land-line and mobile phone numbers. Confidential information will be the password. In addition, the gender will be given. All those characteristics may be used while searching for duplicates.

4) Credit Card Data And Cost Models

Chase Bank and First Union Bank, members of the Financial Services Technology Consortium (FSTC), provided us with real credit card data for this study. The two data sets contain credit card transactions labeled as fraudulent or legitimate. Each bank supplied 500,000 records spanning one year with 20% fraud and 80% nonfraud distribution for Chase Bank and 15% versus 85% for First Union Bank. In practice, fraudulent transactions are much less frequent than the 15% to 20% observed in the data given to us. These data might have been cases where the banks have difficulty in determining legitimacy correctly. In some of our experiments, we deliberately create more skewed distributions to evaluate the effectiveness of our techniques under more extreme conditions. Bank personnel developed the schemata of the databases over years of experience and continuous analysis to capture important information for fraud detection.

We cannot reveal the details of the schema beyond what we have described elsewhere.2 The records of one schema have a fixed length of 137 bytes each and about 30 attributes, including the binary class label (fraudulent/legitimate transaction). Some fields are numeric and the rest categorical. Because account identification is not present in the data, we cannot group transactions into accounts. Therefore,

instead of learning behavior models of individual

customer accounts, we build overall models that try to differentiate legitimate transactions from fraudulent ones. Our models are customer-independent and can serve as a second line of defense, the first being customer-dependent models.

Most machine-learning literature concentrates on model accuracy (either training error or generalization error on hold-out test data computed as overall accuracy, true-positive or false-positive rates, or return-on-cost analysis). This domain provides a considerably different metric to evaluate the learned models' performance—models are evaluated and rated by a cost model. Due to the different dollar amount of each credit card transaction and other factors, the cost of failing to detect a fraud varies with each transaction. Hence, the cost model for this domain relies on the sum and average of loss caused by fraud.

IV. DATA-MINING TECHNIQUES

We investigated the performance of three Techniques in predicting fraud: Logistic Regression (LR), Support Vector Machines (SVM), and Random Forest (RF). In the paragraphs below, we briefly describe the three techniques employed in this study.

1) Logistic Regression

Qualitative response models are appropriate when dependent variable is categorical In this study, our dependent variable fraud is binary, and logistic regression is a widely used technique in such problems. Binary choice models have been used in studying fraud. For example, binary choice models in the case of insurance frauds to predict the likelihood of a claim being fraudulent. In case of insurance fraud, investigators use the estimated probabilities to flag individuals that are more likely to submit a fraudulent claim. Prior work in related areas has estimated log it models of fraudulent claims in insurance, food stamp programs, and so forth. It has been argued that identifying fraudulent claims is similar in nature to several other problems in real life including medical and epidemiological problems

2) Support Vector Machines

Support vector machines (SVMs) are statistical learning techniques that have been found to be very successful in a variety of classification tasks. Several unique features of these algorithms make them especially suitable for binary classification problems like fraud detection. SVMs are linear classifiers that work in a high-dimensional feature space that is a non-linear mapping of the input space of the problem at hand. An advantage of working in a high-dimensional feature space is that, in many problems the non-linear classification task in the original input space becomes a linear classification task in the high- dimensional feature space.

SVMs work in the high dimensional feature space without incorporating any additional Computational complexity. The simplicity of a linear classifier and the capability to work in a feature-rich space make SVMs attractive for fraud detection tasks where highly unbalanced nature of the data (fraud and non-fraud cases) make extraction of meaningful features critical to the detection of fraudulent transactions is difficult to achieve.

Applications of SVMs include informatics, machine vision, text categorization, and time series analysis. The strength of SVMs comes from two important properties they possess — kernel representation and margin optimization. In SVMs, mapping to a high-dimensional feature space and learning the classification task in that space without any additional computational complexity are achieved by the use of a kernel function. A kernel function can represent the dot product of projections of two data points in a high-dimensional feature space. The high-dimensional space used depends on the selection of a specific kernel function. The classification function used in SVMs can be written in terms of the dot products of the input data points. Thus, using a kernel function, the classification function can be expressed in terms of dot products of projections of input data points in a high-dimensional feature space. With kernel functions, no explicit mapping of data points to the higher-dimensional space happens while they give the SVMs the advantage of learning the classification task in that higher

Dimensional space classification function is arrived at.

SVMs minimize the risk of over fitting the training data by determining the classification function (a hyper-plane) with maximal margin of separation between the two classes. This property provides SVMs very powerful generalization capability in classification. In SVMs, the classification function is a hyper-plane separating the

different classes of data.hw; xi + b = 0 ð1Þ The notation hw; xi represents the dot product of the coefficient vector w and the vector variable x.

The solution to a classification problem is then specified by the coefficient vector w. It can be shown that w is a linear combination of data points xi · i=1,2,···,m i.e., w=Σiai xi, ai≥0. The data point's xi with non-zero αi is called the support vectors.

A kernel function k can be defined as kðx1; x2Þ = hΦðx1Þ; Φðx2Þi where Φ:X→H is a mapping of points in the input space X into a higher-dimensional space H. As can be seen, the kernel function implicitly maps the input data points into a higher-dimensional space and return the dot product without actually performing the mapping or computing the dot product. There are several kernel functions suggested for SVMs. Some of the widely used kernel functions include,

Linear function, kðx1; x2Þ = hx1; x2i; Gaussian radial basis function (R B F), kðx1; x2Þ = e−σtx1−x2t2 and polynomial function, kðx1; x2Þ = hx1; x2id. The selection of a specific kernel function for an application depends on the nature of the classification task and the input data set. As can be inferred, the performance of SVMs is greatly depended on the specific kernel function used.

The classification function has a dual representation as follows, where yi are the classification labels of the input data points.

Σ
i
aiγihxi; xi + b = 0

Using a kernel function k, the dual classification function above in The high-dimensional space H can be written as

Σ
i
aiγi kðx1; xÞ + b = 0

As mentioned earlier, in SVMs, the best classification function is the hyper-plane that has the maximum margin separating the classes. The problem of finding the maximal margin hyper-plane can be formulated as a quadratic programming problem. With the dual representation of the classification function above in the high-dimensional space H, the coefficients ai of the best classification function are found by solving the following (dual) quadratic programming problem. Maximize

WðαÞ =Σmi
= 1αi−1
2
Σmi
;j = 1αiαjγiγjkðxi; xjÞ
subject to
0≤αi≤ c
m
ði = 1; ···;mÞ
Σmi
= 1αiγi = 0

The parameter C in the above formulation is called the cost parameter of the classification problem. The cost parameter represents the penalty value used in SVMs for misclassifying an input data point. A high value of C will result in a complex classification function with minimum misclassification of input data whereas a low value of C produces a classification function that is simpler.

The parameter C in the above formulation is called the cost parameter of the classification problem. The cost parameter represents the penalty value used in SVMs for misclassifying an input data point. A high value of C will result in a complex classification function with minimum misclassification of input data whereas a low value of C produces a classification function that is simpler. Thus, setting an appropriate value for C is critical to the performance of SVMs. The solution of the above quadratic programming problem is a computationally intensive task, which can be a limiting factor in using SVM with very large data

3) Random forests

The popularity of decision tree models in data mining arises from their ease of use, flexibility in terms of handling various data attribute types, and interpretability. Single tree models, however, can be unstable and overly sensitive to specific training data. Ensemble methods seek to address this problem by developing a set of models and aggregating their predictions in determining the class label for a data point. A random forest model is an ensemble of classification (or regression) trees. Ensembles perform well when individual members are is similar, and random forests obtain variation among individual trees using two sources for randomness: first, each tree is built on separate bootstrapped samples of the training data; secondly, only a randomly selected subset of data attributes is considered at each node in building the individual trees. Random forests thus combine the concepts of bagging, where individual models in an ensemble is developed through sampling with replacement from the training data, and the random subspace method, where each tree in an ensemble is built from a random subset of attributes.

Given a training data set of N cases described by B attributes, each tree in the ensemble is developed as follows:

- Obtain a bootstrap sample of N cases
- At each node, randomly select a subset of b<B attributes.

Determine the best split at the node from this reduced set of b Attributes - Grow the full tree without pruning Random forests are computationally efficient since each tree is built independently of the others. With large number of trees in the ensemble, they are also noted to be robust to over fitting and noise in the data. The number of attributes, b, used at a node and total number of trees T in the ensemble are user-defined parameters. The error rate for a random forest has been noted to depend on the correlation between trees and the strength of each tree in the ensemble, with lower correlation and higher strength giving lower error. Lower values of b correspond to lower correlation, but also lead to lower strength of individual trees. An optimal value for b can be experimentally determined. Following] and

as found to be a generally good setting for b in , we set b=√B.

Attribute selection at a node is based on the Gini index, though other selection measures may also be used. Predictions for new cases are obtained by aggregating the outputs from individual trees in the ensemble. For classification, majority voting can be used to determine the predicted class for a presented case.

Random forests have been popular in application in recent years. They are easy to use, with only two adjustable parameters, the number of trees (T) in the ensemble and the attribute subset size (b), with robust performance noted for typical parameter values.

They have been found to perform favorably in comparison with support vector machine and other current techniques. Other studies comparing the performance of different learning algorithms over multiple datasets have found random forest to show good overall performance

Random forests have been applied in recent years across varied domains from predicting customer churn, image classification, to various bio-medical problems. While many papers note their excellent classification performance in comparison with other techniques including SVM, a recent study finds SVM to outperform random forests for gene expression micro-array data classification. The application of random forests to fraud detection is relatively new, with few reported studies.

A recent paper finds random forests to show superior performance in credit card fraud detection. Random forests have also been successfully applied to network intrusion detection, a problem that bears similarities to fraud detection.

4)Neuralnetworks:

Neural networks are also often recommended for fraud detection. Dorronsoro et al. (1997) developed a technically accessible online fraud detection system, based on a neural classifier. However, the main constraint is that data need to be clustered by type of account. Similar concepts are: Card watch (Aleskerov et al., 1997); Back-propagation of error signals (Maes et al.,2002); FDS (Ghosh & Reilly, 1994); SOM (Quah & Sriganesh, 2008; Zaslavsky & Strizkak, 2006);improving detection efficiency "mis-detections"(Kim & Kim, 2002). Data mining tools, such as 'Clementine' allow the use of neural network technologies, which have been used in credit card fraud (Bayesian networks are also one technique to detect fraud, and have been applied to detect fraud in the Telecommunications industry and also in the credit card industry.

Results from this technique are optimistic. However, the time constraint is one main disadvantage of such a technique, especially compared with neural networks (Maes et al., 2002).

Furthermore, expert systems have also been used in Credit card fraud using a rule-based expert system.

However, no matter the statistical techniques chosen, the fraud detection system will need to fulfill some conditions. As the number of fraudulent transactions is much less than the total number of transactions, the system will have to handle skewed distributions of the data. Otherwise, the data need to be split into training samples, where the distribution is less skewed (Chan et al., 1997). The system has to be accurate with actual performing classifiers and to be capable of handling noise in the data; a suggested solution is to clean the data (Fawcett & Provost, 1997). The system should also be able to handle overlaps; fraudulent transactions may be similar to normal transactions. As fraudsters reinvent new techniques constantly, the system needs to be adaptive and evaluated regularly. A cost profit analysis is also a must in fraud detection to avoid spending time on uneconomic cases.

5) Credit Card Fraud Detection System Using Genetic Algorithm

Genetic algorithms are evolutionary algorithms which aim at obtaining better solutions as time progresses. Since their first introduction by Holland,

they have been successfully applied to many problem domains from astronomy to sports, from optimization to computer science, etc. They have also been used in data mining mainly for variable selection and are mostly coupled with other data mining algorithms. In this study, we try to solve our classification problem by using only a genetic algorithm solution.

A novel credit card fraud detection system using genetic algorithm is proposed. Genetic algorithms are evolutionary algorithms which aim at obtaining better solutions as time progresses. When a card is copied or stolen or lost and captured by fraudsters it is usually used until its available limit is depleted. Thus, rather than the number of correctly classified transactions, a solution which minimizes the total available limit on cards subject to fraud is more prominent. It aims in minimizing the false alerts using genetic algorithm where a set of interval valued parameters are optimized.

V. CREDIT CARD FRAUD DETECTION USING NOVEL TOOL AND GENETIC ALGORITHM

1) Fraud Detection Using Genetic Algorithm

We collected real time data sets for analysis of fraudulent activities and obtained the following results.

Genetic algorithms have been successfully applied to a wide range of optimization problems including design, scheduling, routing, and control, etc. Data mining is also one of the important application fields of genetic algorithms. In data mining, GA can be used to either optimize parameters for other kinds of data mining algorithms or discover knowledge by itself. In this latter task the rules that GA found are usually more general because of its global search nature. In contrast, most other data mining methods are based on the rule induction paradigm, where the algorithm usually performs a kind of local search. The advantage of GA becomes more obvious when the search space of a task is large.

Genetic algorithms to determine optimal weights of the attributes, followed by *k*-nearest neighbor algorithm to classify the general practitioner data. They claim significantly better results than without feature weights and when compared to CBR. Genetic algorithm is one of the commonly used approaches on data mining. In this paper, we put forward a genetic algorithm approach for classification problems. Binary coding is adopted in which an individual in a population consists of a fixed number of rules that stand for a solution candidate. The evaluation function considers four important factors which are error rate, entropy measure, rule consistency and hole ratio, respectively.

2) Credit Card Fraud Detection Using CRCFRDET

A. Curbing 80% Online Fraud Through Device Identity

One of the fraud screening processes we use is Device Identification. It involves gathering information of visiting devices, probing its operating system, and querying the browser for its time zone including gathering data of the HTTP header information and the screen resolution settings of the device. Credit Card Fraud Detection Tool**(CRCFRDET**) maintains a data of 10 million devices used for online fraud.

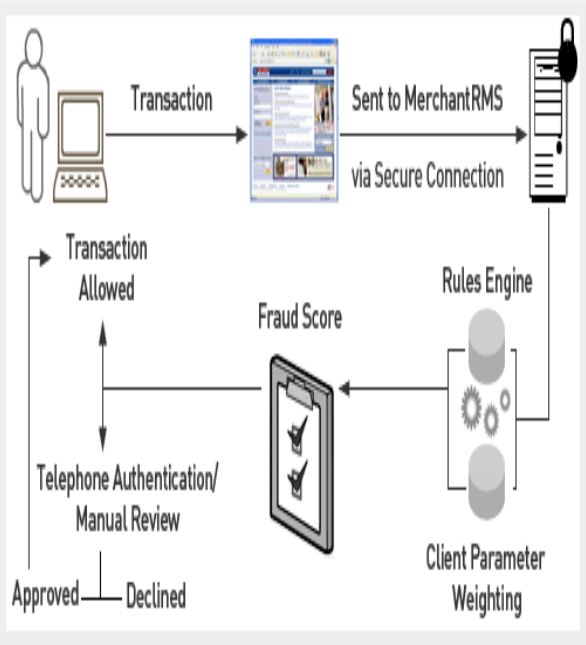

Fig 1 Credit Card Fraud Detection Tool

B. Working Of Credit Card Fraud Detection Tool

- The customer initiates a transaction on the Merchants website and enters the required information (Name, Address, Telephone, email, etc) for order processing.
- **CRCFRDET** system receives the information and captures additional technical information via a secure connection
- **CRCFRDET** system further evaluates this information based on the Merchant's customized rules
- After evaluating **CRCFRDET** sends the fraud score back to the Merchant's e-commerce system.
- Based on the fraud score the transaction is either: Allowed – Sent to Merchant's Bank Payment Processor Denied – Sent to the Merchant's Website

and Transaction is denied Reviewed – Transaction is flagged for Manual Review.

- Optionally Merchant can also trigger a telephone verification for high-fraud-score orders before manual review to verify the end-users phone .

CardID	Auth	Cur.BB	CU	Avg.BB	OD	CCAge	CUT	Loc	LocT	ODT	AmtT
11111,	111,	20000,	13,	60000,	4,	125,	0,	3,	0,	0,	0
11112,	112,	25000,	40,	55000,	20,	264,	6,	4,	2,	0,	9000
11113,	113,	15000,	21,	45000,	3,	111,	2,	10,	2,	1,	15000
11114,	114,	100000,	90,	60000,	29,	350,	1,	11,	14,	0,	8500
11115,	115,	15000,	85,	61000,	17,	211,	3,	3,	7,	0,	12000
11116,	116,	72000,	51,	60000,	19,	321,	5,	9,	0,	1,	12000
11117,	117,	54000,	51,	75000,	9,	275,	6,	9,	0,	1,	7000
11118,	118,	72000,	46,	40000,	12,	271,	1,	7,	2,	0,	19000

VI. RESULTS

Critical Values of each transaction of given DataSet

Acc.,No	Fraud Occurrence	critical Value
11111.0	0.0	0.0
11112.0	0.0	0.0
11113.0	1.0	0.14285715
11114.0	0.0	0.0
11115.0	3.0	2.1289055
11116.0	0.0	0.0
11117.0	0.0	0.0
11118.0	0.0	0.0
11119.0	0.0	0.0
11120.0	4.0	5.171741
11121.0	0.0	0.0
11122.0	1.0	0.18181819
11123.0	0.0	0.0
11124.0	1.0	0.63265306
11125.0	4.0	4.289769
11126.0	1.0	0.16666667
11127.0	1.0	0.1764706
11128.0	0.0	0.0

Value of Critic, Monitor and Ordinary Frauds

3.669799 1.5985714 1.3809904

Fraud Detected used Genetic Algorithm:

Credit Card with ID 11120.0 is detected as fraud with 4.0 occurrences and its critical value is 5.171741

Critical Fraud Detected:

Credit Card with ID 11125.0 is detected as fraud with 4.0 occurrences and its critical value is 4.289769

Credit Card with ID 11130.0 is detected as fraud with 3.0 occurrences and its critical value is 4.8449016

able Fraud Detected:

Credit Card with ID 11115.0 is detected as fraud with 3.0 occurrences and its critical value is 2.1289055

VII. CONCLUSION

The novel method proves accurate in deducting fraudulent transaction and minimizing the number of false alert. Genetic algorithm is a novel one in this literature in terms of application domain. 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. And a series of anti-fraud strategies can be adopted to prevent banks from great losses and reduce risks.

From an ethical perspective, it can be argued that banks and credit card companies should attempt to detect all fraudulent cases. Yet, the unprofessional fraudster is unlikely to operate on the scale of the professional fraudster and so the costs to the bank of their detection may be uneconomic.

The bank would then be faced with an ethical dilemma. Should they try to detect such fraudulent cases or should they act in shareholder interests and avoid uneconomic costs

The objective of the study was taken differently than the typical classification problems in that we had a variable misclassification cost. As the standard data mining algorithms does not fit well with this situation we decided to use multi population genetic algorithm to obtain an optimized parameter.

REFERENCES

[1] M. Hamdi Ozcelik, Ekrem Duman, Mine Isik, Tugba Cevik, Improving a credit card fraud detection system using genetic algorithm, International conference on Networking and information technology 2010.

[2] Wen-Fang YU, Na Wang, Research on Credit Card Fraud Detection Model Based on Distance Sum, IEEE International Joint Conference on Artificial Intelligence 2009.
[3] Clifton phua, vincent lee1, kate smith & ross gayler, A Comprehensive Survey of Data Mining-based Fraud Detection Research, 2005.
[4] C. Paasch, Credit Card Fraud Detection Using Artificial Neural Networks Tuned by Genetic Algorithms, Hong Kong University of Science and Technology (HKUST), Hong Kong, Doctoral Dissertation, 2007
[5] http://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/tcw2/report.html
[6] D.WHITLEY,―Genetic Algorithm And Neural Network,2003.
[7] Elio Lozano, Edgar Acu˜na, Parallel algorithms for distance-based and density-based outliers , 2006.
[8] Credit card fraud detection using hidden markov model – Abinav Srivastava,Amlan Kundu,Shamik Sural,Arun K.majumdar.
[9] Wang Xi. Some Ideas about Credit Card Fraud Prediction China Trial. Apr. 2008, pp. 74-75.
[10] A. Chiu, C. Tsai, ―A Web Services-Based Collaborative Scheme for Credit Card Fraud Detection,‖ Proceedings of the IEEE International Conference on e-Technology, e-Commerce and e-Service, pp.177-181, 200.
[11] Wang Xi. Some Ideas about Credit Card Fraud Prediction China Trial. Apr. 2008, pp. 74-75.
[12] A. Chiu, C. Tsai, ―A Web Services-Based Collaborative Scheme for Credit Card Fraud Detection,‖ Proceedings of the IEEE International Conference on e-Technology, e-Commerce and e-Service, pp.177-181, 200.
[13] Jitendra Dara,Laxman Gundemoni, "Credit Card Security And E-Payment."2006