Credit Card fraud detection using classification techniques

[1]Rutuja S. Gore, [2]Dr. Khan Rahat Afreen

[1]Student, [2]Associate Professor
Computer Science and Engineering
Deogiri institute of engineering and management studies Aurangabad

***Abstract*: Credit card fraud has increased considerably due to the growth of the latest technologies and global communications routes. Credit cards cost billions of dollars in consumer and financial companies per year. Con artists try to find new plans and procedures to continue illegal operations. Therefore, fraud detection systems are necessary for banks and financial institutions to reduce losses. The most common technique used to create fraud detection patterns Furthermore, detecting and preventing credit card fraud is one of the most important problems in the digital world that requires precise transaction analysis. One way to detect fraud is to investigate suspicious changes in user behavior. We have implemented a linear biased classifier using kernel method that allows us to classify data based upon selective attributes by performing vectorization of values and using probability threshold that is set to 90 % based upon which it classifies record as fraud or non fraud. We have compared our results with other techniques such as Bayesian networks, C45, J48 and decision tree along with Bay minimum risks. Our method to improve fraud detection in Credit Cards is the main objective of this work with improvement in current fraud detection process by improving fraudulent account prediction. In addition, there is a collection and discussion of the evaluation criteria used in literature. Therefore, the issue opened for detecting credit card fraud will be described as a guideline for new researchers.**

***Index Terms*: Training dataset, classifiers, Data mining.**

I. INTRODUCTION

Today, with the expansion of credit cards along with online transactions is a serious problem for financial institutions trying to prevent credit card fraud. There are many ways of fraud with all types of credit products, such as tax evasion, illegal trade in goods, obtaining a loan using false information, money transfers under the guidance of the head of a fictitious business operation, donations for fictitious charitable organizations, etc. [1] The use of effective fraud detection systems has become a necessity for all credit card issuing banks to reduce their waste. Many modern methods based on artificial intelligence, data mining, fuzzy logic, machine learning, sequencing, genetic programming, etc., are being developed to detect fraudulent credit card transactions. A clear understanding of all these methods will lead to the creation of effective credit card fraud detection systems. [1] This study solves this problem with data mining algorithms for detecting suspicious credit card transactions. Credit card fraud detection is a process to monitor the behavior of customer transactions throughout the period [2].

Types of credit card fraud:

1. The first type that is most common is application corruption. The person will forge the application to accept the credit card. That person will give false information about their financial status in order to accept credit cards.
2. The second type is considered identity. Suppose someone's identity is in the long-term form for credit card fraud. That person will forge a name with a temporary address.
3. The third category is financial fraud which occurs when a person wants to receive more credit than is currently available. They will apply for a credit card under their own name. But information about their financial status is false
4. The fourth is skimming technology. Reading magnetic card data is a small portable device that Has the sole purpose of collecting and storing information on any credit card [2]

Credit Card Fraud Detection Technique: From the literature survey, found a variety of fraud detection methods. Finally, we come to the conclusion that the detection of credit card fraud has several methods such as [3] 1. Bayesian network 2. Bayesian minimum risk 3. Genetic algorithm 4. Markov model Hidden. 5. Metaphysics. Digging data consists of various techniques and features that can be used to detect credit fraud. Bayesian Bay network is a graph that has a Directed Acyclic in which each node represents a random variable and is associated with the conditional probability of the node received from the parent. This model shows each variable in the domain that defined as a node in the graph and the dependency between these variables is the arc that connects the relevant node. That is, all the edges in the graphic model are guided and not around [4]. For the purpose of detecting network fraud, Bayesian must explain the behaviour of car insurance. First, the Bayesian network was created to simulate behaviour under the assumption that the driver was deceived (F) and another version under the assumption that the driver is a legitimate user (NF) network of fraud. User net 'is set using data from users that are not deceptive. During the operation of the Internet users are adapted to specific users according to the information that occurs. By inserting evidence in these networks and spreading through the network, the probability of measuring x is less than two of the above. Which means that it helps to determine which user behaviors observed at any level are based on fraudulent behavior or general fraud. These quantities are called p (x | NF) and p (x | F) by Place the opportunity for fraud P (F) and P (NF) = 1-P (F) in general and by using Bay rules will give you a chance to be cheated due to measurement x.

$$\frac{P(F|x)=P(F)p(x|F)}{p(x)} \quad (1)$$

Where the denominator p(x) can be calculated as

$$P(x) = P(F) * p(x|F) + P(NF)p(x|NF) \quad (2)$$

Probability of fraud, P (F | x) due to the behavior of users who observe X can be used as an alarm level. On the other hand, Bayesian networks allow the integration of expert knowledge that we have set up. Models [5] On the other hand, user models will receive new training in ways that are not maintained by usage data, so our Bayesian approach has both knowledge and learning from The person professional

Probability of fraud P (F | x) due to user behavior observed On the other side, Bayesian networks allow the integration of expert knowledge from which we have set up models. [5] In contrast, user models will be retrained in ways that are not maintained by Use data Therefore, our Bayesian approach includes both expert knowledge and learning.

Bayes minimum risk is how to detect credit card fraud. As defined in [12], Bayes's Minimum Risk Identifier is a decision-making model based on the number of inter-decisions that are made using a variety of probabilities and costs that come with such decisions. In the case of credit card fraud detection, there are two decisions; Prediction of a transaction is fraudulent, pf or legal. The risks associated with anticipating transactions as defined are fraud.

$$R(pf|x) = L(pf|yf)P(pf|x) + L(pf|yl)P(pl|x) \quad (3)$$

and when the transaction is predicted as legitimate it is

$$R(pl|x) = L(pl|yl)P(pl|x) + L(pl|yf)P(pf|x) \quad (4)$$

Yf and yl are real labels for fraud and systematic transactions, respectively. P (pl | x) is the approximate probability of a formal transaction defined by x. P (pf | x) is the probability of fraudulent transactions receiving x. Finally, L (a, b) is a loss function and real label b when

predicting a transaction. When calculating both risks, the transaction is organized. Fraud R (pf | x) pl R (pl | x) means the risks associated with the decision to reduce the risks associated with that type of law.

Genetic mechanisms are evolutionary mechanisms that aim to get better solutions over time. The following figure shows the flow of the gene process. Repeating this process until the number of predefined specimens has passed can be seen as the best solution for using genetic algorithms. For best performance, it is necessary to follow the steps and parameters that list the parameters and settings needed to create the fraudulent transaction.

If the data set, T = {t1, t2, t3 ... tn} D - is a data object, p P P P T of the data set labeled D ∈ T P and D is the object, D is considered a general computer object. This gives credit card fraud a better chance of generating credit card fraud with effective false alerts. In the fraud, the proposed system was discovered based on customer behavior. New classification problems have been suggested, which have different variable classification costs. [8] Therefore, genetic mechanisms occur when a set of parameters with the most relevant time values are present. C_Freq, C_Loc, C_OD, C_BB and C_Ds can increase real numbers by defining the current values of the parameters. The critical values are then compared to the data set parameters, and the number of alerts does not exceed the correct size

II. Related Work

Previous work, the most commonly used method for detecting fraud or being a type of conflict is NB (Innocent Bayesian) [10], C4.5, Quinlan [16] and PP (Propagation) proposed. By Elkan et al] Domingos [8], Elkan [9] [10] and Witten [8] NP. This algorithm is very useful in many real data sets and is very useful in learning relationships. Although the linear and nonlinear properties are nonlinear and non-distributed properties are straight from the simulation data. But the accuracy of the prediction will be reduced not only to the C4.5 output, but also to accurately predict the system, structure, decisions and rules set However, the problem of scalability and efficiency may reduce performance considerably when using C4.5 with large data. Artificial neural network boxes may be processed many times and resist noisy data. The PP method takes time in Longer training and detailed test and training parameters The most common disadvantage of these methods is that they rely on supervised training which requires test cases to improve human participation and parameters in the preparation of training cases.

In order to solve this problem, fraud detection systems have been proposed to detect fraudulent transactions in the online system, which show dynamic user behavior patterns. Since most online systems have non-stationery information, the system can adjust the detector to keep up with changes in user behavior.

Brin et al [4] introduces dynamic item counting techniques to reduce the number of database scans. Ozden et al [15] presents circular and interesting digging rules. Ng et al [14] introduces techniques for making Limited rule mining

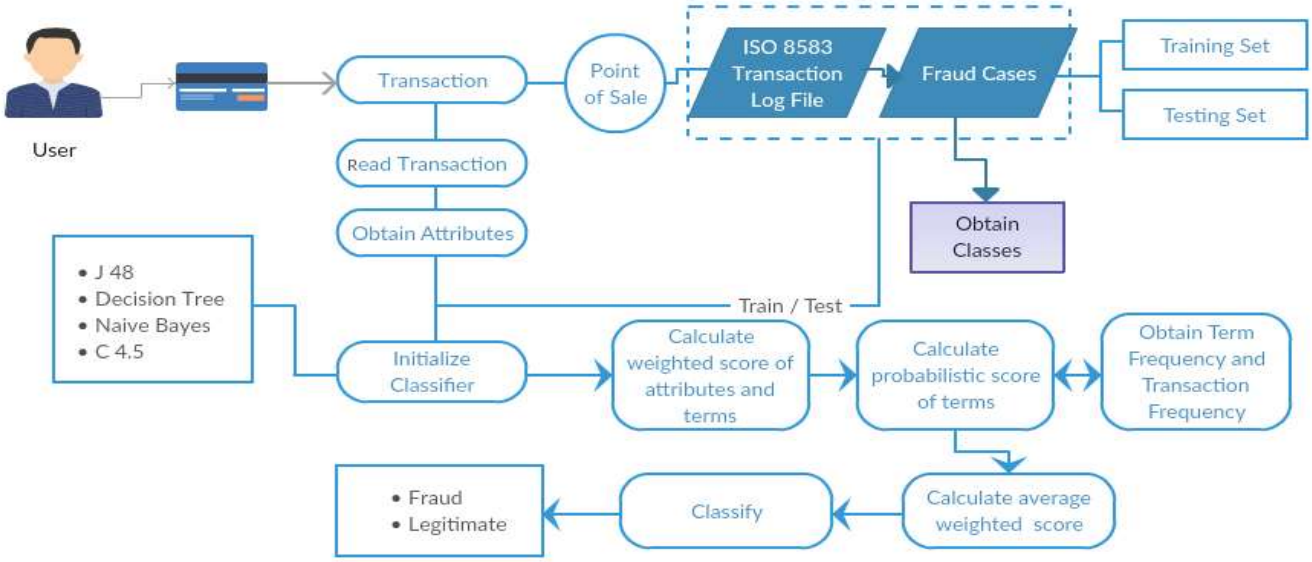
Cheung et al [6] offers techniques for adding unit updates to discover link rules in large databases. Some of the most popular law creators, RL, proposed by Clearwater [7], C4.5 by Quinlan [16], for example, are based on instructional learning.

FP tree structures (often patterned trees) and FPtree growth algorithms proposed by Han [12] are used to expose these hidden link rules from recent transactions for this user.

Online transaction systems including web applications and services to prepare OLTP (online transaction processing procedures), FDS (fraud detection system), database that stores transaction data and database replication. In order to reduce OLTP performance or FDS analysis is a backend process that affects the front end of the online system, it is reduced because it talks about a cool replica

database. That Fraud detection system consists of three main modules (1) Data engine acts as a interface between the replication database and FDS. It collects and formats the latest transactions of all individual customers online. (2) Modules The latest transaction rule engine to create a profile, which is a link rule stored in the FP tree for each user (3)). The rule check module monitors transactions. For all users New user-specific transactions will be compared to the FP-tree for that user to identify the fault. Due to the advanced electronic commerce technology, the use of credit cards has increased rapidly and has become a popular payment mode for online and offline purchases. The card is convenient to use. But the card has no risk There are many fraud detection systems. But detected fraud after suspicious transactions It is important that banks and commercial organizations develop effective fraud detection systems. [5] This article discusses various types of fraud in credit cards and how to find fraud in an economical manner. Presenting challenges that cardholders face, including card issuers, individuals who commit fraud, recent information about fraudsters, credit cards, and provide some protection techniques for cardholders to pursue fraudulent activities.
T.Abdul Razak, G.Najeeb Ahmed conducted a comparative analysis of credit card fraud techniques using data mining. In order to effectively improve the risk of shop management levels, creating a precise and easy credit card risk monitoring system is one of the important missions. The aim of this article is to discover the patterns of users who identify high level fraud cases. It is important to reduce access to resources and information without permission. In order to create a completely safe system with validation process, it is necessary to analyze the behavior before the transaction. Modern techniques based on decision trees, artificial neural networks, genetic algorithms, hidden Markov models are recommended for detecting fraudulent credit card transactions. Fraud detection based on customer behavior variables used in financial institutions Prodromidis and Stolfo [10]. Utilize distributed learning methods with distributed learning for fraud detection in credit card transactions. It is based on artificial intelligence and blending algorithms, inductive learning and metal earning methods for higher accuracy.
Phua et al. [11] proposes to use a similar Meta classifier in the fraud detection problem. They think naïve Bayesian and Back Propagation Neural Networks are the Xu Wei base classification and the SVM model development team is suitable for fraudulent online credit card transaction detection. Due to the risk of having to face security issues in online credit card transactions, this article focuses on the detection and prevention methods of credit card fraud. The principles of support for vector machine algorithms and model selection are discussed. And the final form of credit card fraud detection based on the support vector machine was created Such forms are applied to anti-fraud systems for online payments with credit cards. Advance data processing and find the best SVM model and compare with ID3 with BP. Hybrid models The performance of the SVM model is higher than the performance of the hybrid version of ID3 with BP.

learning package, we must first reduce the depth of the tree. To reduce the tree, we need to get function information. In precisely determining the information received, we need to calculate the entropy first.
When records must be categorized But there is a missing value for the field, separating its value in the agent field, can be used to separate The maximum number of surrogates is selected as Unlike the naive method, C4.5 uses the ratio of the entropy as an impurity measurement. C4.5 uses Chi-Square distribution for separation and integration operations. Therefore, two probabilities must be used, the first to represent the significant level for node separation and the second represents the priority level for node integration.



III. METHODOLOGY

This section deals with the work that this job offers, which is intended to improve the detection of credit card fraud. This proposed work consists of following steps:

Preprocessing

Data processing reduces the data complexity and offers better chances for subsequent analysis. The best features can be found by determining the dependency between any conditional feature and the decision feature. Features with higher dependency values are taken in the final subset of best features.

Classification using J48 classifier

In classification items are classified according to the item features with respect to the set of classes which are predefined. J48 classifier is a simple C4.5 decision tree for classification. A binary tree is created in this. The approach of decision tree is most useful in classification. With this method, a tree is built as the model of classification process. Once the tree is made, it is applied to every tuple of the database and classification for that tuple is obtained.

Using Naive Bayes Classifier

The Naïve Bayes machine learning classifier tries to predict a class which is known as outcome class based on probabilities, and also conditional probabilities of its occurrence from the training data. This kind of learning is very efficient, fast and high in accuracy for real-world scenarios, and also this learning type is known as supervised learning. The initial step for Naïve Bayes classification algorithm is the Bayes theorem for conditional probability, where 'x' is given data point and 'C' is a class:

$$P(C / x) = P(x/C)/P(x) \quad (5)$$

Even though it is often surpassed by other techniques such as boosted trees, random forests, Max Entropy, Support Vector Machines etc., Naive Bayes classifier is very effective as it is less costly in computing (both CPU and memory) And it requires a small amount of training data. In addition, the training time with Naive Bayes is much lower as opposed to the alternative methods.

Method C4.5 uses two parameters The first is to measure the impurity that is used to decide whether to separate, use which is chosen as the coefficient and the second is the maximum number of agents. The agent is the way to deal with the lost value. For each item in the tree, the decision tree method specifies the input field that is closest to the selected field. Those fields are representative for that separation.In order to find the best way to classify the

Key Index Parameters for Result Classification

In information retrieval with binary classification, precision (also called positive predictive value) is the fraction of retrieved instances that are relevant, while recall (also called sensitivity) is the fraction of the relevant instances that are retrieved. Precision and recall are therefore based on understanding and measuring relevance. In simple terms, high accuracy means that an algorithm returns significantly more relevant than irrelevant results, while a high recall means that an algorithm has yielded the most relevant results.

The most important category measurements for binary categories are

Precision	Recall	F Measure
$P = TP/(TP + FP)$	$R = TP/(TP + FN)$	$tp - tn/tp - tn + fp + fn$

Experimental Setup

We conducted experiment on pre collected data set from UCI respository and Kaggle Dataset webportal. We have conducted experiment on Core TM i3 - 3110 M CPU with speed efficiency of 2.40 GHZ, processing capability. The system contains 6 Gigabytes of usable memory for processing.

Kaggle Dataset:

The datasets contains transactions made by credit cards in September 2013 by European cardholders. This dataset presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transaction. It contains only numerical input variables which are the result of a PCA transformation. Unfortunately, due to confidentiality issues, we cannot provide the original features and more background information about the data. Features V1, V2, ... V28 are the principal components obtained with PCA, the only features which have not been transformed with PCA are 'Time' and 'Amount'. Feature 'Time' contains the seconds elapsed between each transaction and the first transaction in the dataset. The feature 'Amount' is the transaction Amount, this feature can be used for example-dependant cost-sensitive learning. Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise

UCI Repostitory Dataset:

Credit Card Dataset contains 20 attributes are as mentioned below:

1.Attribute	(qualitative) : Status of existing checking account
2.Attribute	(numerical) Duration in month
3.Attribute	(qualitative) Credit history
4.Attribute	(qualitative) Purpose
5.Attribute	(numerical) Credit amount
6.Attribute	(qualitative) Savings account/bonds
7.Attribute	(qualitative) Present employment since
8.Attribute	(numerical) Installment rate in percentage of disposable income
9.Attribute	(qualitative) Personal status and sex
10.Attribute	qualitative) Other debtors / guarantors
11.Attribute	(numerical) Present residence since
12.Attribute	(qualitative) Property
13.Attribute	(numerical) cc_age in months
14.Attribute	(qualitative) Other installment plans
15.Attribute	(qualitative) Housing
16.Attribute	(numerical) Number of existing credits at this bank
17.Attribute	(qualitative) Job
18.Attribute	(numerical) Number of people being liable to provide maintenance for
19.Attribute	(qualitative) Telephone
20.Attribute	(qualitative) foreign worker

The biased classifier uses Kill Threshold 1.0E-6 with convergence tolerance of 1.0E-4 with maximum iterations 10000. We have used Adaptive Over-Relaxation technique with the non-component-wise, that applies an over-relaxation technique whose goal is to accelerate convergence to the optimum weights, but which might possibly overshoot and thus increase the time required for convergence. It is not clear to us when this is a good option and when it is not; your mileage may vary.

We used Cache Exponentials to create a cache of computed exponentials that may reduce the overall amount of computation time if the cache becomes heavily used; it may also result in a small loss of accuracy because of sooner rounding.

We have implemented Cache Kernel Functions with cross-validation, a kernel is applied to the same two points many times across all folds of cross-validation, so this option creates a cache of the results of applying the kernel to all pairs of points. This requires normalization to be performed on the entire data set beforehand, so data that is left out in each fold will still be used in the normalization.

Confusion matrix:

	bad	good
bad	22	58
good	25	195

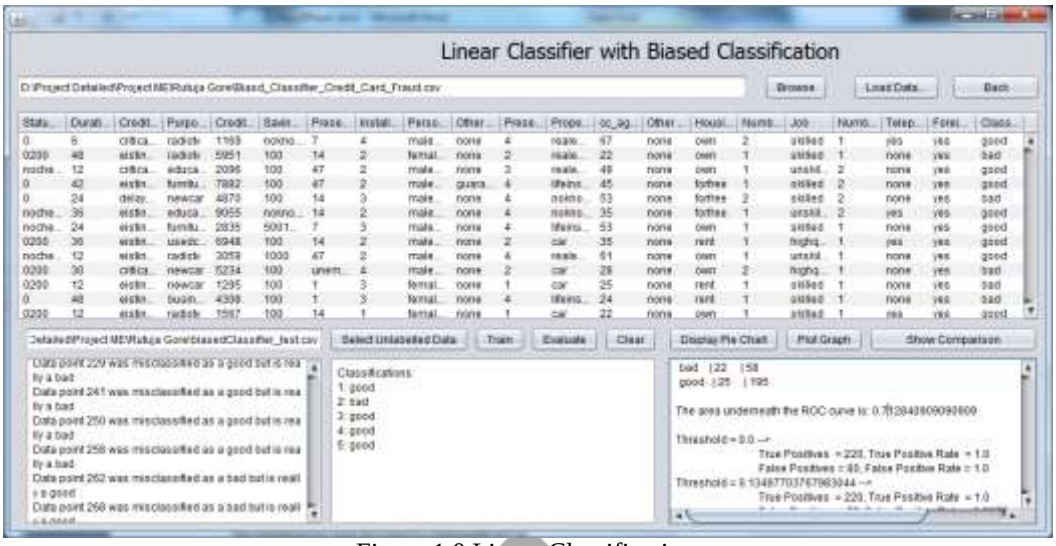


Figure 1.0 Linear Classification

ROC Curve Values:

	Values
Area underneath the ROC	0.712840909090909
Best threshold Rate	0.7521209304375942
Best threshold counts	0.4107564967369055
Threshold with respect to the ratio of positives to negatives is	0.7333333333333333

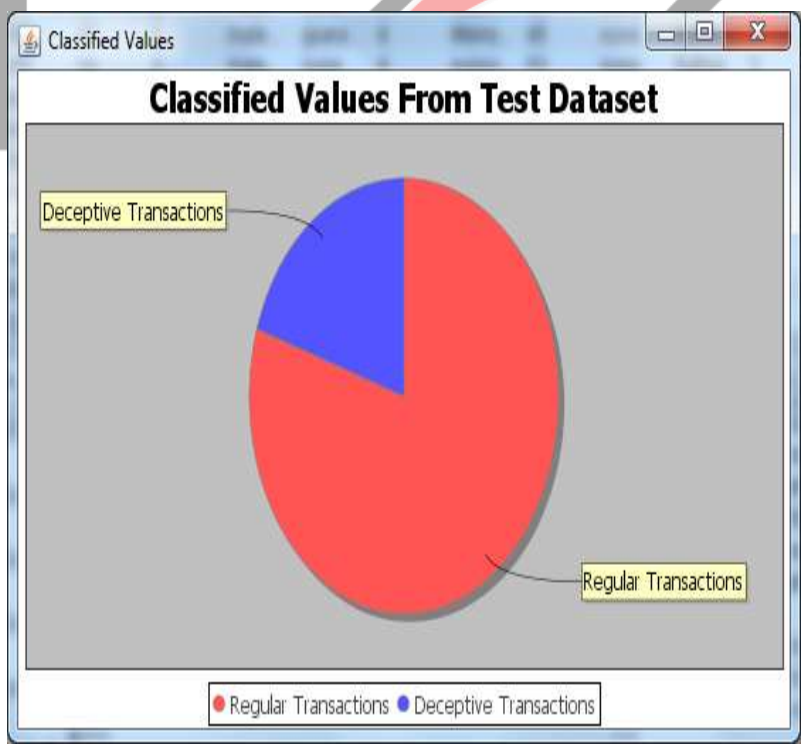


Figure 2.0 Pie Chart Representation of Regular transactions and Deceptive Transaction.

IV. Results and Discussion

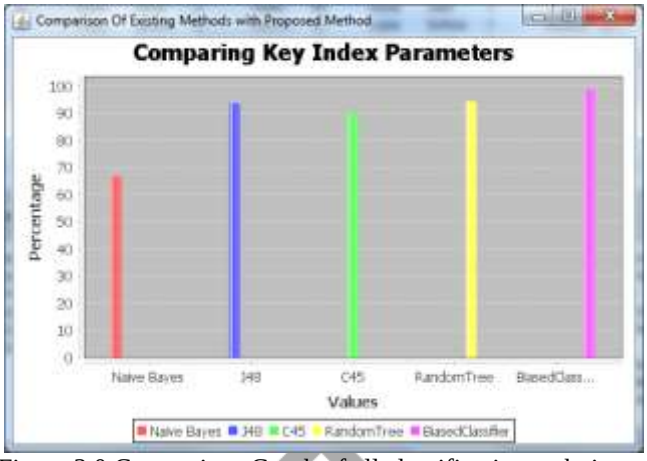


Figure 3.0 Comparison Graph of all classification techniques

As a shared information between features. The number of functions that are provided by the user The efficiency of the proposed technique was evaluated based on real-life microarray gene expression data to select non-duplicated genes and related genes. In addition, the efficiency of the proposed method is compared with well-known items and the results show that the proposed method performs schema for all data sets.

V. Conclusion and Future Scope

Under this framework, financial institutions can utilize credit card fraud detection models to compare the transaction information with the historical profile patterns to predict the probability of being fraudulent for a new transaction, and provide a scientific basis for the authorization mechanisms. We have proposed a method where relevance and redundancy are supported in parallel. To measure the relevance and redundancy of a characteristic or a fraudulent transaction in credit card was considered. Relevance is defined as the mutual information between a feature vector and class labels. Redundancy is described as mutual information among the characteristics. The number of resulting functions is provided by the user. The performance of the proposed technique is evaluated on the basis of some sets of credit cards transactions data to select fraud and non fraud entities.

Future Scope can be considered and furthermore, online resources that can track credit card transactions at real time using location and type of item bought with previous buying history of a user can be utilized and institutions can be more focused on more suspicious transactions to decrease the fraud levels

At the time of studying the above details, quite a long time passed due to the formation of train sets and test kits. Therefore, in order to ensure that the model is the latest model, we therefore do new training with the latest information released over the past five months Since deciding the actual status of a transaction that may be fraudulent, it takes some time. (Sometimes cardholders claim that the transaction is fraudulent. But after the investigation, it may become true by the family members of the cardholder) claiming to be outside the scope We identify which rules are not necessary and are well protected from fraud points and decide to eliminate 65 from

Rule out 7 online rules for screening and 8 offline rules for sending SMS. Offline rules Six rules are cut by cheating points and only the intersection decides to send SMSs. In other words, transactions that follow That rule should have a high fraud rating. However, one of the rules about Internet transactions is that it has expanded and transactions are small. But the higher fraud rate is sending SMS to rule out the rest of the online. Most are very specific.

Cases (such as information about copying card groups) and keeping them still However, we have added new online rules based on fraudulent scores.

There are ways to choose different properties in existing literature. But in most cases, we have seen that the basic purpose of this method is relevance or redundancy. In this article, we have proposed relevant and redundant methods supported in parallel. In measuring the relevance and redundancy of features or data-sharing genes is considered. Relevance is defined as a shared information between feature vectors and class labels. Redundancy is described

VI. Acknowledgment

I feel great pleasure express my sincere thanks to Guide Dr. Khan Rahat Afreen**,** for guiding me at every step in making of this project. She motivated me and boosted my confidence and I must admit that work would not have been accomplished without her guidance and encouragement.

REFERENCES

[1] Hobson, A. 2004. The Oxford Dictionary of Difficult Words. The Oxford University Press. New York.
[2] Bolton, R. J. and Hand, D. J. 2002. Statistical fraud detection: A review. Statistical Science 28(3):235-255.
[3] Kou, Y. et. al. 2004. Survey of fraud detection techniques. In Proceedings of the 2004 IEEE International Conference on Networking, Sensing and Control, Taipei, Taiwan, March 21-23
[4] Phua, C. et al. 2005. A comprehensive survey of data mining-based fraud detection research. Artificial Intelligence Review.
[5] 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, June 2010.
[6] Leonard, K. J. 1993. Detecting credit card fraud using expert systems. Computers and Industrial Engineering, 25
[7] Ghosh, S. and Reilly D. L. 1994. Credit card fraud detection with a neural network. In Proceedings of the 27th Hawaii International Conference on system Sciences, 3.
[8] Mena, J. 2003. Investigate Data Mining for Security and Criminal Detection. Amsterdam, Butterworth-Heinemann.
[9] Aleskerov, E., Freisleben, B. and Rao, B. 1997. CARDWATCH: A neural network based data mining system for credit card fraud detection. In Computational Intelligence for Financial Engineering. [10] Chen, R., Chiu, M., Huang, Y. and Chen, L. 2004. Detecting credit card fraud by using questionnaire-responded transaction model based on SVMs. In Proceedings of IDEAL2004.
[11] Brause, R., Langsdorf, T. and Hepp, M. 1999. Neural data mining for credit card fraud detection. In Proceedings of the 11th IEEE International Conference on Tools with Artificial Intelligence.
[12] Stolfo, S. J., Fan, D. W., Lee, W., Prodromidis, A. L. and Chan, P. K. 1997. Credit card fraud detection using meta-learning: Issues and initial results. In AAAI Workshop on AI Approaches to Fraud Detection and Risk Management. AAAI Press, Menlo Park, CA.
[13] Stolfo, S., Fan, W., Lee, W., Prodromidis, A. L. and Chan, P. 1999. Cost-based modeling for fraud and intrusion detection: Results from the JAM Project. In Proceedings of the DARPA Information Survivability Conference and Exposition. IEEE Computer Press, New York.
[14] Prodromidis, A. L., Chan P. and Stolfo S. J. 2000. Meta-learning in distributed data mining systems: issues and approaches. Advances of Distributed Data Mining, Editors Kargupta H. and Chan, P. AAAI Press.
[15] Chen, R.-C., Luo, S.-T., Liang, X. and Lee, V. C. S. 2005. Personalized approach based on SVM and ANN for detecting credit card fraud. In Proceedings of the IEEE International Conference on Neural Networks and Brain, Beijing, China
[16] Hand, D. J. and Blunt G. 2001. Prospecting gems in credit card data. IMA Journal of Management Mathematics, 12.
[17] Dahl, J. 2006. Card Fraud. In Credit Union Magazine.
[18] Schindeler, S. 2006. Fighting Card Fraud in the USA. In Credit Control, House of Words Ltd.
[19] Dorronsoro, J. R., Ginel, F., Sanchez, C. and Cruz, C. S. 1997. Neural fraud detection in credit card operations. IEEE Transactions on Neural Networks.
[20] Hanagandi, V., Dhar, A. and Buescher, K. 1996. Density-Based Clustering and Radial Basis Function Modeling to Generate Credit Card Fraud Scores. In Proceedings of the IEEE/IAFE 1996 Conference.