ISSN: 0976-3031

Available Online at http://www.recentscientific.com

CODEN: IJRSFP (USA)

International Journal of Recent Scientific Research
Vol. 8, Issue, 4, pp. 16308-16313, April, 2017

International Journal of Recent Scientific Research

DOI: 10.24327/IJRSR

Research Article

CREDIT CARD FRAUD DETECTION-A HYBRID APPROACH USING SIMPLE GENETIC AND APRIORI ALGORITHMS

Prarthana Adiga., Samvida V Bhat., Sanjana R Javagal., Lavanya B S and Chandrika J

Department of CS&E, Malnad College of Engineering, Hassan, India

DOI: http://dx.doi.org/10.24327/ijrsr.2017.0804.0125

ARTICLE INFO

Article History:

Received 15th January, 2017
Received in revised form 25th February, 2017
Accepted 23rd March, 2017
Published online 28th April, 2017

Key Words:

Credit Card, Fraud Detection, Genetic Algorithm, Transactions, Data Mining, Apriori Algorithm, Feature Selection

ABSTRACT

Drastic increase in E-commerce has lead to an evolution of credit cards where they act as the navigators to an evolving environment. However, this dramatic increase has also resulted in frauds in credit card transactions. Furthermore, over decades, the technologies have changed, developed and evolved dramatically to give entirely a new face to such frauds. In reality, fraudulent transactions are often amalgamated with genuine transactions. Numerous techniques have been designed to detect and prevent such frauds. Few such approaches include Hidden markov model, Neural network, Decision Tree and so forth. However, these techniques fail to decrease the number of false alerts where in cases of fraud are intimated to the users even when no frauds have actually occurred. In this paper, we present a technique to detect frauds using a combination of one of the novel approach called the Genetic Algorithm and Data mining. Genetic algorithm aims at providing an optimal solution to the problems of fraudulent transactions and also attempts to decrease the number of false alerts while data mining attempts to refine the result obtained from the genetic algorithm.

INTRODUCTION

Digitalization has led to a boost in online transactions. The various modes of payments include net banking, credit card, debit card transactions and so forth. The most popular mode of payment is credit cards. It allows the cardholder to perform online as well as offline transactions. Its security relies on the physical security of the card as well as the card number and other such important details. This increased use of credit cards using the internet has resulted in a substantial rise in fraudulent activities. Frauds are undesirable activities in an operational environment. There are various techniques that are adopted by the fraudsters to commit frauds. Fraud basically involves the physical theft of the card, or any compromise with the account details of the cardholder like card number, account number and other such important details that is necessary for a transaction to be carried out.

Fraud detection involves identifying of the use behavior so that the undesirable behavior is estimated and detected. Although stolen cards can be reported by the cardholder immediately, compromised accounts cannot be recognized easily and the fraudster can use it for days together without the cardholder’s notice. This will result in substantial loss for both the cardholder and the bank.

To address this problem, financial institutions employ various techniques tools like real-time credit card authorization, address verification systems (AVS), card verification codes, rule based detection, and so on. But fraudsters are adaptive, and given time, they devise several ways to circumvent such protection mechanisms. Despite the best efforts, the fraudsters have managed to adapt and circumvent these preventive mechanisms and thus the frauds continue to rise.

There are various methods proposed like fraud detection using hidden markov model, neural networks, decision tree, Fusion of Dempster Shafer and Bayesian learning, Hidden Markov Model, Artificial neural networks and Bayesian Learning approach, BLAST and SSAHA Hybridization, Fuzzy Darwinian System etc.

Most of the credit card fraud detection systems mentioned above are based on artificial intelligence, Meta learning and pattern matching. However, these algorithms give rise to one or the other disadvantage be it cost, performance, execution time, false alarm rate, etc. Hence, a combination of a novel approach with a conventional classification technique of data mining will

**Corresponding author: Prarthana Adiga*
Department of CS&E, Malnad College of Engineering, Hassan, India

provide an optimal and a better solution of detecting fraudulent transactions in a given dataset. This paper presents one such unique combination of genetic algorithm with a classification problem of data mining i.e Apriori algorithm. There are basically 2 ways in genetic algorithm to address this fraud detection problem. They are (i) using simple genetic algorithm and (ii) genetic K-means algorithm. However, when these two are compared, simple genetic algorithm is found to be more optimal than k-means algorithm.

This hybrid approach helps the user detect the fraudulent transactions from the given dataset which contains details of the card holder like account number, balance, authorization type and other such parameters. Out of these parameters, few of the significant parameters are selected using feature selection, and then are considered for the genetic algorithm process. As a preprocessing technique, feature selection is made where in important variables are selected using supervised attribute selection algorithms. These features are then used by the genetic operators to detect frauds.

Genetic algorithm is a search algorithm that follows survival of the fittest rule of natural selection. It uses three genetic operators, selection, crossover and mutation. This approach use fitness score and inherited good properties of parents to produce new generation due to which GA provide better results. Apriori algorithm is one of the classification algorithms. This is applied to get a refined dataset. When applied along with genetic algorithm, it provides an optimal solution to the defined problem. It uses bottom-up approach and works on the basis of hash tree and BFS (breadth first search). It is a process where in frequent items in the dataset are mined using association rule mining.

LITERATURE SURVEY

A Hidden Markov Model (HMM) despite of not requiring fraud signatures is still able to detect frauds by taking a cardholder's spending patterns into considerations. Despite, the Genetic Algorithm is of a better use in the domain of data mining prominently for variable selection and is also mostly coupled with some other conventional data mining algorithms and their amalgamation with other techniques has a very good performance. Genetic Algorithm has been used in credit card fraud detection for reducing the wrongly classified number of transactions viz. false alarms and is also easily accessible for a computer programming language implementation. Thus, it is evident that this helps in making it strong in credit card fraud detection. However, this method has high performance and is quite expensive.[1]

Availing the user of the card of both flexibility as well as a better security control for their e-cards can be made a higher priority. The cardholder may apply for some customized security measures such as specifying a limit on the value of each transaction, determining the expiry of their e-card, restricting an e-card to be used at a single merchant, limiting the total amount that can be spent using a single card, etc. Newer and better security measures in preventing the credit card frauds in online payments are desired by the stakeholders most importantly, the payment operators, authorities responsible for issuing the card, merchants and the cardholders themselves. Despite of its importance and increasing attraction from the stakeholders, the study shows that the effective implementation of such security measures is capital intensive because of which, the smaller organizations with weaker financial background may find it difficult to invest on this venture. Hence, they may have to look for alternative cost effective solutions for this problem.[2]

In this era of digitization, the popularity of electronic mode of payment system as well as transaction frauds are increasing simultaneously. A method that uses genetic algorithm proves to be beneficial in this regard where the test data is initially generated and the fraud is detected correspondingly. Evidently, genetic algorithm used to detect credit card frauds potentially has positive impacts on financial institutions, customers and also the merchants. A series of anti-fraud ideas can be incorporated to avoid the immense loses the financial institutions may have to incur and hence reducing the risks. This algorithm convincingly has higher benefits as compared to other techniques used.[3] The four components of fraud detection system are rule-based classifier, Dempster-Shafer adder, transaction history database and Bayesian learner. The classification of fraudulent transactions is done by the rule-based classifier. Next, the Dempster-Shafer's theory combines multiple such evidence. The transaction data includes both fraudulent as well as genuine transactions that are used to build the solution model. As soon as a transaction is found to be unusual, it can be determined if the corresponding transaction is nearer to being a fraudulent or a genuine one. Such systems superficially give good results in fraud detection but on the other hand, they increase false alarms too.[4]

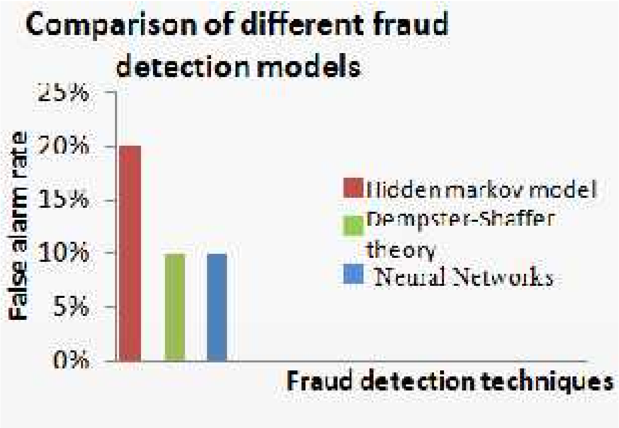
The neural network is trained in such a way that it matches each transaction pattern with the original data transaction pattern of the cardholder so that it distinguishes between the transactions into fraudulent or genuine. Thus, the transactions that are not suspicious are declared to be OK. The goal of using neural network technique is to determine the rank of the credit card by considering the fraud score that its transactions produces. The ranking thus produced is assigned a threshold and any transactions scoring below it is considered to be fraudulent and only good transactions will be ranked better. However, the drawback of this method is that the selected pattern recognition space may not effectively be able to segregate certain non-separable genuine and fraudulent distribution.[5]

Preserving privacy in the developing E-commerce technology is one of the important factors in this era. This paper concerns on providing security. It mainly uses data mining for providing privacy. It uses mainly five modules namely, Privacy-Preserving Data Analysis, Non-Cooperative Computation, Analyzing Data Analysis Tasks In the NCC Model, Security System. The privacy preserving data analysis protocols identify about the truthfulness of the user. The NCC model helps to learn the output of each function and identifies the correctness by giving first priority to every participating party. The NCC model makes use of SMC and DNCC to provide privacy to users. The security System mainly uses association rule mining which responds with the security code only if it is a valid credit card operation. The major disadvantage of this approach is the algorithm used can be applied only for the financial institution and the customer will not be initiated about this.[6]

Due to the modernization credit cards are large in use. So the fraud associated with the credit card is also increasing.

This paper concentrates in avoiding such credit card fraud using genetic algorithm. Genetic algorithm is the evolutionary algorithm which helps in detecting frauds. It makes use if the genetic operators like selection, crossover and mutation. The modules used are the user GUI, critical value identification and fraud detection process. The critical value is identified based on different parameters. The selection operator selects the transactions which are assumed to be fraudulent transactions based on the credit card parameters. It mainly makes use of the tournament and elitist selection to detect the fraud transactions. Thus this paper helps you n detecting the fraud transactions and also reduces the false alarms.[7]

This paper mainly concerns on analysing the credit card fraud detection techniques. The major frauds found are credit card frauds, telecommunication frauds, computer intrusions, Bankruptcy fraud, Theft fraud/counterfeit fraud, Application fraud and Behavioural fraud. Some of the techniques used for the fraud detection are Bayesian Networks, Decision tree, Hidden Markov Model, Support Vector Machine and Neural Network. The above mentioned techniques can detect the frauds but are inefficient. These techniques fail to provide accuracy. So combination of these algorithms if applied to the financial institutions helps to detect the frauds easily and effectively as soon as the credit card transactions are done. Applying such combination algorithms helps to avoid risks and provides security.[8]



Due to the rise in the E-commerce, the online transactions and the use of credit cards has rapidly increased. Electronic payment has become a boon but also has to lead to many security issues. Electronic payments include frauds associated with it. The most known credit card fraud is counterfeit fraud. The paper concentrates on providing mechanisms that will that will minimise such frauds. The detection of such frauds mainly proceeds in the following steps. The steps include application process, activation process, and transaction behaviour monitoring, detection and fraud prevention on internet. The algorithm proceeds by collecting the data, analysing and interpreting the data and applying some research approaches like quantitative research, positivist research, interpretive research and so on. Even though the research concentrates on security it is unable to reach the level of security that it is supposed to [9]

Payment card fraud causes huge amount of losses to the financial institutions. This paper provides a method to detect such payment frauds. It begins by proving the fraud statistics and defines the problem. It describes methods used by identity thieves to obtain personal and financial information. Later on provides a solution to detect the frauds. The categories of the fraud are credit card application fraud, account take over, lost or stolen card, card not received fraud, counterfeit fraud, ATM fraud, card not present transaction and so on. The paper detects and prevents the fraud by using anti-fraud techniques like card activation, card verification codes, consumer education, address verification and real time POS authorization. As the digitization is increasing rapidly the fraud associated will also increase so the solutions obtained to not able to avoid the frauds completely [10]

This paper provides an overview of the different types of feature selection techniques for attribute selection. Feature selection is the basic tool to pick required attributed for a model classification. Basically of three categories, supervised, semi-supervised and unsupervised, feature selection is at its best when supervised feature classification is used. The gene data is unlabelled, completely labeled or partially labeled. Semi-supervised and semi-unsupervised are extended versions of supervised and unsupervised feature selection that work on both labeled and unlabelled data. The main stages of feature selection include determining search direction, determining of search strategy, determining evaluation criteria, define stopping criteria and validating the result.[11]

The survey of the above research papers convey that algorithms used so far have ended up with some or the other limitations. Among the algorithms used genetic algorithm is found to provide a better solution to the problem. However, a combination of genetic algorithm with one of the conventional classification problem of data mining will provide an optimal solution.

Comparative Study

The below graph provides a detailed comparison of different methodologies that are applicable in detecting frauds in credit card transactions.

Hidden Markov Model gives very high false alarms and thus has the greatest disadvantage of all. Dempster-Shafer theory is highly expensive and has a low processing speed. Using Neural Networks, Frauds cannot be detected for first transaction.

However, with the implementation of the proposed system, the percent of transactions classified as fraudulent but in fact are legitimate can be reduced to as less as 5%

Simple genetic and apriori algorithms

Most of the online transactions and electronic payments today are carried out using credit cards and net banking facilities. Credit cards however, have become the most popular mode of payment because of the ease of use. In other words, Credit cards are the navigators to online transactions. But, one of the rising disadvantages along with the increased use of cards is the frauds associated with those transactions. The proposed project is to develop a credit card fraud detection system which detects the fraudulent transactions.

Genetic algorithm (or GA) is a search technique used in computing to find true or approximate solutions to optimization and search problems. It is a meta heuristic which is inspired by the natural selection process. It is a method for solving some of the constrained and unconstrained problems. It draws inspiration from the typical evolutionary biology. Accordingly, it uses approaches that resemble the typical selection, crossover and mutation which, in terms of genetic algorithm, are collectively known as the genetic operators.

Flow of Genetic Algorithm

Step 1: Initially the initial population is selected randomly from the sample space which has many populations.
Step 2: The fitness value is calculated for each chromosome in each population and is sorted out.
Step 3: In selection process two parent chromosomes are selected through tournament method.

The Crossover forms new offspring (Children) from the parent chromosomes using single point probability. Mutation mutates the new offspring using uniform probability measure. In elitism selection the best solution are passed to the further generation.

Selection

The selection operator is used to select the best individual from the population. Selection process is selection of genomes from the chromosomes which lead to the further breeding in the population.

Some of the selection types include tournament selection, Elitist selection and so on. Tournament Selection is a kind of selection where the indivuals is chosen randomly and select the best one from the randomly chosen genomes to become a parent. This process continues until the best parent is chosen from the population.

Elitist selection is one where the genomes are being selected from the limited number of chromosomes using the fitness function. Rank Selection is a probability proportional to the relative fitness and not to the absolute fitness. Among these we have found elitist selection seems to be the more optimal one.

Elitist selection in genetic algorithm

It is a selection procedure where a limited number of individuals are selected from the population using the fitness function. Using the fitness value the one having a higher fitness value will be allowed to produce the next set of population. The one which don’t have higher fitness function are not allowed to perform the next operations that are the crossover and mutation.

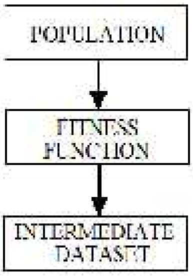


Fig 1 selection

The pseudo code of the selection operator is as follows

Selection()

```
    For each value of the selected parameters
        For i from 1 to k do
            If fitness(pop[i]) > fitness_val
                best[j] =  pop[i]
```

Crossover

Crossover is also a genetic operator which is succeeded by the selection. . It is is similar to biological crossover. It is a process where the child is obtained from the parent. The child takes chromosomes from the parent solutions to become an individual to proceed to the next operation. The operation is performed by selecting the crossover point randomly and the child take solutions from the parent using the crossover point.

Some of the types of crossover are single point crossover, two point crossover, uniform crossover and so on.

Single point crossover is crossover technique where a single crossover point is chosen randomly and the child take over the solution from that point. Two point crossover is a crossover technique where two crosspoints are chosen and the child swaps the genomes from the parent which yields to two children. Uniform crossover is a crossover technique which involves mixing up of the parent organisms to breed a child solution.

Among the crossover techniques the single point crossover provides a better solution

Single point crossover in Genetic algorithm

It is a crossover technique which involves selecting one crosspoint randomly from the population. That is the crossover point is chosen in both of the parent solutions. The parent solutions are being obtained from the selection. Using the randomly selected crosspoint and the parent solutions the crossover technique is applied. The data beyond that crossover point is intermixed which yields a single child.

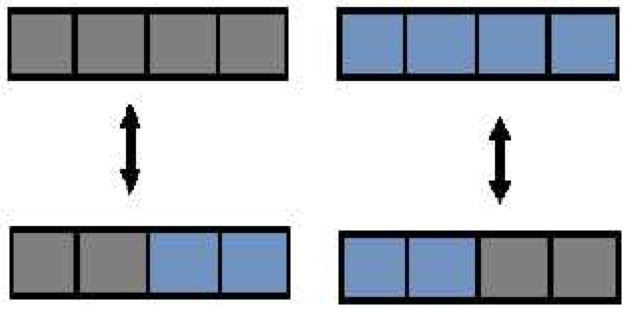


Fig 2 Crossover

The psuedocode of the crossover technique is as shown below

Crossover()

```
    For each intermediate transactions
    For crosspoint from 0 to Father.length do
    for i from 0 to father.length do
            if i < crosspoint
                child[i] =  father[i]
            else
                child[i] =  mother[i]
```

Mutation

It is also a genetic operator in the genetic algorithm. The mutation is mainly used to maintain heterogeneity in the population. It alters some of the genes because of which the entire solution. Hence forth using mutation in genetic algorithm we can obtain the better solutions from the selected dataset.

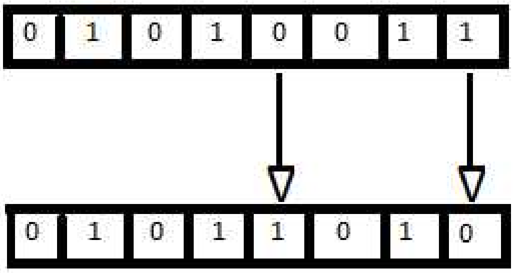


Fig 3 Mutation

The psuedocode of the crossover technique is as shown below

Mutation()

```
For each intermediate transaction data do
i <- random(0,1)
j <- random(0,N)
k<- random(0,3)
if k <- 1 then
            pop[l][j] <- pop[l][j]+I
else
            pop[l][j] <- pop[l][j]-I
```

Apriori Algorithm

Apriori algorithm is the classification algorithm used for mining the item sets from the population using association rules. Association rule helps in identifying the frequent items in the given dataset. The frequent item set is one whose threshold should be greater than the user defined threshold. The minimum threshold and the minimum confidence are the main parameters of the Apriori. It uses a “bottom-up approach”. The algorithm uses hash tree and breadth first search (BFS). Hash tree is a search technique which helps to search and keep track of the frequency count of items.BFS is used to avoid processing the same nodes many times.

The working of the apriori algorithm is as given below.

It initially counts the frequency of each item. It proceeds by assuming the minimum frequency count of item in the dataset as the threshold value. Later on, it determines the items which are its subsets and prunes such items whose threshold is less than the minimum threshold.

All the other items which satisfy the threshold condition are retained as such. The data items are first selected individually and the frequency count is being taken. This is then extended by coupling the transactions. These coupled transactions are again subjected to the threshold condition and the process continues iteratively.

In the proposed paper, the same apriori algorithm is applied to improve the accuracy. Initially the genetic algorithm detects the fraudulent transactions and those transactions are given as input to the apriori which helps in the further refinement of those transactions. The output of the apriori will be the more accurate fraudulent transactions.

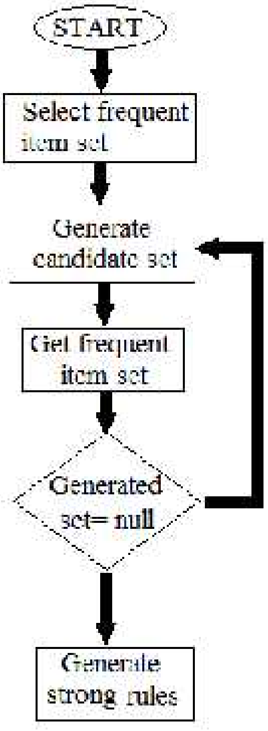


Fig 4 Apriori Algorithm

Modules

1. GUI(Graphical User Interface)
2. Feature selection using weka
3. Fraud detection process
4. Refining of the obtained dataset

Graphical User Interface

The user interface is designed using visual studio where in a form is provided to let the user to input various account details like cardId, authentication type, book balance and other such 12 parameters. Soon after the dataset is generated, the fraud detection process is carried out using Genetic and Apriori algorithm. All these are background processes and will not be visible to the user. Once the fraud is detected, the type of fraud, intensity and the cardID is displayed to the user.

Feature Selction Using Weka

Feature selection, or variable selection is a process of selecting required features that are to be used while constructing a model. Weka is a collection of machine learning algorithms used for data mining tasks. It contains tools for data-preprocessing, classification, regression, clustering, association rules etc. Once the database file is setup, it is utilized by weka where in useful attributes are selected. These attributes are selected using supervised attribute selection algorithms that are very flexible and lets vivid search and evaluation techniques to be combined. This lets the developer select important features to be considered while computing the critical values for the parameters while applying genetic operators and thus reducing the overhead.

Fraud Detection Process

This is the main part of the project wherein the frauds are detected, classified and displayed to the user. The fraud detection process is carried out as follows.

1. A set of data consisting of transaction details like cardId, authentication type, and other parameters is given as the input and stored in the dataset.
2. Critical values are computed using the parameters selected.
3. Generate critical values after limited number of generations.
4. Apply the genetic operators i.e. selection, crossover and mutation over the transaction dataset iteratively until optimal solution is obtained.

Refining of the Obtained Dataset

The Genetic algorithm will result in a set of data that is said to be the fraudulent dataset. In order to get a more refined dataset, the traditional Apriori algorithm is applied over the final dataset.

The pseudo code of the algorithm is as shown below

Apriori(pop[i][j],t)

```
d={frequent items}
For i from 1 to  dt do
      cs= i
for each item in pop[i][j] do
     cs=cs+1
if(i<t)
   i=i+1
return i, dt
where,
dt=frequent itemset
cs=candidates in itemset
t= threshold
```

Sample Dataset

The data set may contains the following parameters

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
11117,117,20000,43,40000,12,261,0,6,1,0,0
11118,118,23000,31,35000,9,259,4,7,4,0,19000
11119,119,12000,29,45000,7,183,1,10,2,0,16000
11120,120,35000,189,70000,30,269,5,4,10,1,11000
11121,121,77000,31,60000,7,311,2,8,2,0,11000
11122,122,50000,31,65000,9,208,0,2,11,0,0
11123,123,29000,51,55000,16,291,1,6,12,0,14000
11124,124,81000,62,70000,18,196,2,6,3,0,9000
11125,125,13000,83,55000,12,138,4,3,1,1,19000
11126,126,70000,32,50000,9,173,0,2,12,0,0
11127,127,54000,51,75000,9,275,6,9,0,1,7000
11128,128,72000,46,40000,12,271,1,7,2,0,19000
11129,129,14000,103,30000,22,318,1,11,4,1,22000
11130,130,20000,111,61000,29,201,6,5,11,0,14000[7]
```

CONCLUSION

In this paper, we present a hybrid approach of genetic algorithm and a classification problem of data mining to detect and distinguish fraudulent transactions from a given dataset. Genetic algorithm makes use of 3 operators selection, crossover and mutation to derive optimal solution i.e. the intensity of the frauds occurred in a set of transactions that contain around 12 parameters related to the credit cardholder's account. Few of the best features are selected using feature selection, are used to calculate the fitness values in genetic algorithm and then the result obtained from genetic algorithm is fed to the classification technique to get a more refined data.

The implementation and efficient ulitilization of the proposed system must help the financial institutions in a successful detection of the frauds occurred during the online transactions. This also must successfully reduce the numerous false alerts that get raised even when a genuine transaction is made.

References

1. S.Benson Edwin Raj, A. Annie Portia, "Analysis on Credit Card Fraud Detection Methods", IEEE International Conference on Computer, Communication and Electrical Technology, IEEE March 2011.
2. Jitendra Dara, Laxman Gundemoni, -Credit Card Security and E-Payment.‖ 2006.
3. M.Hamdi Ozcelik, Mine Isik, -Improving a credit card fraud detection system using Genetic algorithm‖, IEEE International Conference on Networking and Information Technology, IEEE 2010.
4. Amlan Kundu, Suvasini Panigrahi, Shamik Sural and Arun K. Majumdar, -Credit card fraud detection A fusion approach using Dempster–Shafer theory and Bayesian learning, Special Issue on Information Fusion in Computer Security, Vol. 10, Issue no 4, pp.354- 363, October 2009.
5. Credit card fraud detection using neural network, Raghavedra Patidar, Lokesh Sharma, ISSN: 2231-2307, Volume, IssueNCAI211, June-2011.
6. An efficient approach for identifying fraud transactions on E-commerce Data by D.Srikanth, M.Tech Scholar, N Zareena M.Tech, Assistant Professor, Department of Computer science.
7. K. RamaKalyani, D. UmaDevi. Fraud Detection of Credit Card Payment System by Genetic Algorithm. Department of Computer Science, Sri Mittapalli College of Engineering, Guntur, AP, INDIA
8. Analysis on credit card fraud detection techniques by Renu HCE Sonepat, Suman.HCE Sonepat
9. Credit card security and E-payment by Jitendra Dara and Laxman Gundemoni Lulea University of Technology
10. Payment Card Fraud: Challenges and Solutions by The University of Texas at Dallas
11. Supervised, Unsupervised and Semisupervised Feature Selection: A Review on Gene Selection Ang Jun Chin, Andri Mirzal, Habibollah Haron, Senior Member, IEEE, Haza Nuzly Abdull Hamed
12. https://en.wikipedia.org/wiki/Apriori_algorithm

****