Fraudulent Detection in Credit Card System Using SVM & Decision Tree

[1]Vijayshree B. Nipane, [2]Poonam S. Kalinge, [3]Dipali Vidhate, [4]Kunal War, [5]Bhagyashree P. Deshpande

SSBT's College of Engineering & Technology
Bambhori, Jalgaon - 425 001 (MS)

Abstract—With growing advancement in the electronic commerce field, fraud is spreading all over the world, causing major financial losses. In current scenario, Major cause of financial losses is credit card fraud; it not only affects trades person but also individual clients. Decision tree, Genetic algorithm, Meta learning strategy, neural network, HMM are the presented methods use to detect credit card frauds. In contemplate system for fraudulent detection, artificial intelligence concept of Support Vector Machine (SVM) & decision tree is being used to solve the problem. Thus by implementation of this hybrid approach, financial losses can be reduced to greater extend.

Keywords-SVM; Decision Tree; Model File; Prediction File; Libsvmtools; LIBSVM; fraudster;

I. INTRODUCTION

Credit card fraud can be defined as a wide-ranging term for theft & fraud committed using or involving a payment card, such as a credit card or debit card, are fraudulent source of funds in various kind of transaction. The purpose may be to obtain goodies without paying, or to obtain unauthorized funds from an account or to avail some kind of service. Prevention & detection of fraud in systems are two important aspects that are to be considered so as to avoid frauds & losses due to fraudulent activities.

Dem& of online shopping is increasing day by day. Credit card provides great ease for online transaction. According to ACNielsen who conducted study in 2015, analyzed that about one-tenth of the world's population is shopping online.

India's credit card base may have topped 20 million in the just concluded 2013-14 financial year -its highest level in the past five years -according to World line India, which provides services for critical online or electrically operated actions in the country. As the number of credit card holders increases world-wide, the opportunities for fraudster are also increases.

Now a day the customers prefer the most popular payment mode with credit card for the convenient way of paying bills, online shopping is easiest way & handy. At the same time the fraud transaction risks using credit card is a crucial problem which must be avoided. So there are many artificial intelligence techniques available to avoid these risks effectively. In existing research we modelled the sequence of operations in credit card transaction processing using a support vector machine & decision tree algorithm & shown how it can be used for the detection of frauds. To elude computational complexity & to provide better accuracy in fraud detection in proposed work. Support vector machine is a method used in pattern recognition & classification. It is a classifier to predict or to classify patterns into two categories which may be fraudulent or non fraudulent. A decision tree is a tree structure which attempts to separate the given transaction records into mutually exclusive subgroups., thus helps us identifying the behaviour.

II. PROBLEM DEFINITION

Fraud Identification in credit card system is most concerning issues in online transaction. With growing popularity on online shopping provided by various web services, various kind of fraudulent activities are being observed. With emergence of information technology & improvement in data communication, fraud is expanding causing huge financial losses. Complete elimination of banking fraud is not possible; however we can limit its occurring to certain level & prevent them from happening by artificial intelligence technique. Approach of artificial intelligence in credit card detection is newly used. Thus probability of effective results is likely to be more. In this work, fraudulent detection for credit card system using decision tree induction algorithm & support vector machine algorithm is proposed. SVM classifier is mainly used for differentiating unauthorized user from all kind of users, & decision tree is used for fraud handling based on behaviour of user. Decision tree technique used is built on classification or predication models based on recursive partitioning of data.

III. RELATED WORK

To understand & advancement of credit card fraud detection techniques, it is essential to examine their history. Credit card frauds are increasing day by day irrespective of the various techniques developed for its detection. Fraudsters are so talented that they generate new ways for committing fraudulent transactions on each day, which demands constant innovation for its detection techniques.

R. Dhanpal & P. Gayathiri [1] has focused on the information gain based. This method estimates the best split of purity measures, entropy & information gain ratio to test the best classifier attribute.

R. D. Patel & D. K. Singh [3] presented the system to generate fraud transactions generated with the given sample dataset. If genetic algorithm is applied in bank for credit card fraud detection, the chance of fraud transactions predicted soon after credit card transactions is in process, & anti-fraud strategies are adopted to prevent banks from great losses before the transaction & reduce risks.

Y. Sahin & E. Duman, in [2], demonstrates the advantages of applying the data mining techniques including Decision Trees & Support Vector Machine (SVM) to the credit card fraud detection problem for reducing the banks risk. The results show that the classifiers & other Decision Tree approaches outperform SVM approaches in solving the problem under investigation.

J. Pun & Y. Lawryshyn, in [4], described Meta-Learning aims to filter the legitimate transactions from the fraudulent & by quickly & accurately identifying the fraudulent transactions, fraud losses can be reduced. Meta-learning techniques introduced by Stolfo & Chan. There are two techniques of combining algorithms that were introduced by Stolfo & Chan, the arbiter & the combiner strategies. Chan & Stolfo [4] found the combiner strategy which performs more effectively than the arbiter strategy. In the combiner strategy, the correct classifications & attributes of credit card transaction instances are used to train multiple base classifiers.

R. Patidar & L. Sharma, in [5], presented fraud detection using Neural Network is totally based on the human brain working principal. Neural Network method has made a computer capable to think. As human brain learns through past experience & use that knowledge or experience to take the decision in daily life problem. The same technique is applied with the credit card fraud detection technology.

A. Srivastava & A. Kundu, in [6], presented a HMM is a double embedded stochastic process with two hierarchy levels. It is complicated stochastic processes as compared with traditional Markov Model. A Hidden Markov Model has a finite set of states monitored by a set of transition probabilities. In a particular state, observation or an output generated according to an associated probability distribution. It is only the output & not the state that is visible to an external observer.

G. Singh et al., in [7], presented Support Vector Machines have developed from Statistical Learning Theory of AI domain. It has been widely applied to fields such as character & text recognition, handwriting digit, & more recently to satellite image classification. SVMs & other nonparametric classifiers have a reputation for being effective & reliable. SVMs function by non-linearly projecting the training transaction dataset in the input space to a feature space of higher dimension by use of a kernel function is used. The results are stored in a separable dataset that can be separated with linear classifier. The process enables the classification of transaction datasets which are usually non- linearly separable in the input set. The functions used to project the data from input set to feature sets are called kernels (kernel machines), examples of which include Gaussian, polynomial & quadratic functions. Each function has alone parameters which have to be checked prior to classification & it also usually calculate through a cross validation process.

IV. SYSTEM OVERVIEW

We proposed fraudulent detection system which provides three level of security. First & foremost is the Login & password, unless & until authenticated user enters valid Login ID & password he cannot enter into system. Second level of security is provided by SVM, which monitors user behaviour & decision tree, which helps in determining whether user behaviour is normal or abnormal. Third level provides questionnaires to user which has to submit by user without a single wrong answer. These are the question which is to be inserted by user at the time of registration. Even though using SVM presents promising ways to fraud detection, the number of attributes provided to it is huge. This generally leads to large processing times. Some attributes might even have the probability of creating a negative impact on the result, which is undesirable. Hence identification of attributes plays a crucial role in determining the accuracy of the final result. Analyzing the spending behaviour pattern of the customer is a promising way to detect the credit card frauds.

Behaviour based fraud detection model means that the data use in the model are from the transactional behaviour of cardholder directly or derived from them. Fraud detection based on the analysis of existing spending behaviour of user or cardholder is a promising way to find the credit card frauds. Based on the spending pattern the customer's normal activities such as transaction amount, billing address etc are learned & noticed. Some of the anomalous behaviours include change in billing address or shipping address, maximum amount of purchase, large transaction done far away from current residential place etc.

V. SYSTEM ARCHITECTURE

The architecture is a system that unifies its components or elements into logical & functional blocks. The architecture shows the structure of system & modules included in system.

The Proposed System comprises of three tire architecture:

1) Support Vector Machine:- SVM is a binary classification, hence the transactions are labelled either as fraudulent, or legitimate. This helps us to identify abnormal behaviour of user i.e. Fraud User. It uses regression technique.

2) Decision tree algorithm: - The decision tree is a structure that includes root node, leaf node & branch. Each internal node denotes a test on attribute, the conditional results of test denotes each branch & the class label holds by each leaf node. The root node is the topmost node in the tree.

3) Questionnaires:- Comprises of all questions, of which few may be user defined & some system defined.

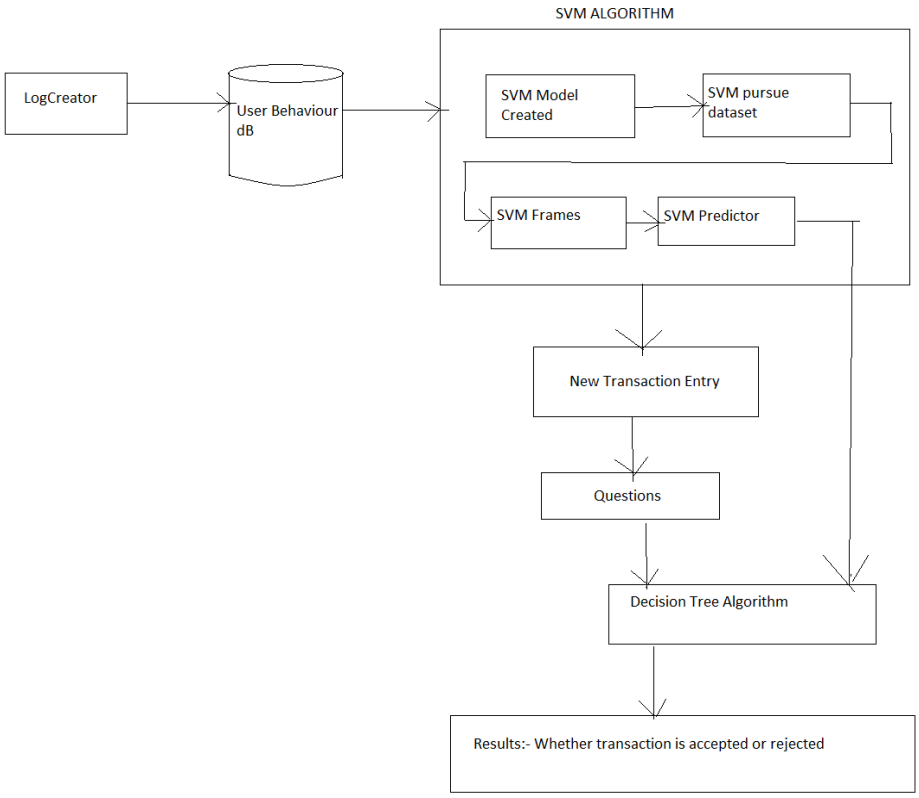

Fig:1-Architecture of Proposed System

The front end comprises of all the login access of user, which mainly consist of user module, login module & login details database module. User module contains the entire interface GUI used for interaction. Login consists of access providing facility to user. & login detail database module consist of all log information about user & also all transaction details.
Backend consist of truncation module, account details module, verification module & security module. Account details consist of behavioural data of user. Transaction contains details of current transaction to be carried out & which is in process. Verification module works in collaboration with security. This mainly consists of SVM, Decision tree & questionnaires.

VI. IMPLEMENTATION DETAILS

Support Vector Machines:- (SVMs) are a popular machine learning method for classification, regression, & other learning tasks. LIBSVM is a library for Support Vector Machines (SVMs). A typical use of LIBSVM involves two steps: first, training a data set to obtain a model & second, using the model to predict information of a testing data set. For SVC & SVR, LIBSVM can also output probability estimates. Many extensions of LIBSVM are available at libsvmtools.
A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples.

1. Set up the training data for model creation
2. Set up SVM's parameters
3. SVM Trainer
4. SVM Predictor

Decision Tree:- A decision tree is a tree in which each branch node represents a choice between a number of alternatives, & each leaf node represents a decision. The ID3 algorithm is used by training on a dataset to produce a decision tree which is stored in memory. At runtime, this decision tree is used to classify new unseen test cases by working down the decision tree using the values of this test case to arrive at a terminal node that tells you what class this test case belongs to.

1. Calculate the entropy of every attribute using the dataset.
2. Split the dataset into subsets using the attribute for which entropy is minimum (or, equivalently, information gain is maximum)
3. Make a decision tree node containing that attribute
4. Recurse on subsets using remaining attributes.

VII. RESULT

A) RESULTS & ANALYSIS OF DATASET CREATOR

Table:-1 refers to the Log Creator file, which includes the modules as read ip address of the device that is been used for transaction processing, read users for transaction processing, read page urls, read dates. The log creator is used for generating a large data set so we require an array list of users, ip's, pages, dates etc. The log creator has to take these modules as an input &

create a large data which has to be feed to the SVM to create a log file. This log file generated by log creator is used as an input by SVM. Large Dataset created by Log file creator is shown in Table 2 below:

Table 1-Log File Creator

Read ip	Read users	Read urls	Read dates	Create log for SVM
192.0.0.1	Poonam	Ora.html http/1.0	5/6/12	652345
192.0.0.2	Riya	sql.html http/1.0	1/9/15	871236

Table 2- Large Dataset created by Log file creator

TID	CN	TAMT	TD	TT	OS	Brow	TIP	TLOC
1	23456	1000	5/6/12	Shopping	MAC	UC	192.0.0.1	Pune
2	12366	2000	1/9/15	Transfer	IOS	Opera	192.0.0.2	Mumbai
3	76542	3000	2/4/14	Booking	Linux	UC	192.0.0.3	Delhi

B) RESULTS & ANALYSIS FOR SUPPORT VECTOR MACHINE

SVMs are a popular machine learning method for classification, regression, &other learning tasks. LIBSVM is a library for Support Vector Machines (SVMs). A typical use of LIBSVM involves two steps: first, training a data set to obtain a model & second, using the model to predict information of a testing data set.

For SVC & SVR, LIBSVM can also output probability estimates. Many extensions of LIBSVM are available at libsvmtools. The main functions carried out by SVM are as follows:-

1. Set up the training data for model creation
2. Set up SVM's parameters for the dataset that is created so as to send them for SVM training.
3. SVM Trainer, which trains each & every individual data from the large dataset.
4. Once the dataset is trained completely the SVM Predictor does prediction of that trained data.

The training of the dataset that is created before is done as follows:-

optimization finished, #iter = 1
nu = 1.0
obj = -1.0000326009430864, rho = 0.0
nSV = 2, nBSV = 2
*
optimization finished, #iter = 1
nu = 1.0
obj = -1.0006755256908946, rho = 0.0
nSV = 2, nBSV = 2
Total nSV = 700

C) RESULTS & ANALYSIS FOR QUESTIONNAIRES CARRIED OUT BY DECISION TREE ALGORITHM

The ID3 algorithm is used by training on a dataset to produce a decision tree which is stored in memory. At runtime, this decision tree is used to classify new unseen test cases by working down the decision tree using the values of this test case to arrive at a terminal node that tells you what class this test case belongs to. The main function of SVM module is to detect the normal or abnormal behaviour of the user, & if abnormal then inform to the decision tree. The ID3 algorithm is used for this purpose & it generates user based questions, which has to be answered by the user. If the answers matches with the dataset then decision tree gives an output that transaction is accepted else transaction is rejected.

1. Calculate the entropy of every attribute using the dataset.
2. Split the dataset into subsets using the attribute for which entropy is minimum (or, equivalently, information gain is maximum)
3. Make a decision tree node containing that attribute
4. Recurse on subsets using remaining attributes.

VIII. DICUSSION

Analysis on result showed that as the transaction records increases prediction value given by SVM gradually decreases. Because large dataset with variable behaviour of user is being trained by SVM. Thus for most transaction questionnaires are issued. Thus greater security is assured. Decision tree simply uses the prediction value given by SVM & transaction detail for Decision making process i.e. whether transaction is to be accepted or rejected. This approach provides two levels Security to user i.e., SVM & Decision Tree Algorithm. A system with this security difficult to user attempt successful transaction but increases to maximum security level. Change in behaviour not exactly a fraudster; it may be valid user so give one more chance for verification. This hybrid approach is not easily identified by the fraudster, thus may provide security to authenticate user for long time.

IX. CONCLUSION

This research provides a complete composition of structures for efficient fraud detection. It initially starts with the usage of clustering & outlier detection techniques. These techniques are considered to be the basis for finding data that does not belong to the current data pattern. These are further made accurate by the usage of SVM & behaviour bases SVM. SVM, being a binary classifier helps in providing the user with a result of whether the current transaction is legitimate or fraudulent. But the false positive rates in SVM are high, hence an additional methodology of collective animal behaviour is incorporated to provide more accurate results.

Detecting the fraudulent process is the most important functionality that a bank could offer its customers. But this could prove to be a serious downside if the transaction detected as fraudulent by the system proves to be a legitimate one. This could lead to reduction in goodwill of the company. Hence reduction in false positives is a compulsory attribute that should be adopted by every system. In this system, the accuracy rate reaches to 59%.

REFERENCES

[1] R. Dhanpal & P. Gayathiri, "**Credit card fraud detection using decision tree for tracing email & ip,**" *International Journal of Computer Science Issues*, vol. 9, no. 2, 2012.

[2] R. D. Patel & D. K. Singh, "**Credit card fraud detection & prevention of fraud using genetic algorithm,**" *International Journal of Soft Computing & Engineering (IJSCE), vol. 2,no. 6, 2013.*

[3] Y. Sahin & E. Duman, "**Detecting credit card fraud by decision trees & support vector machines,**" *Proceeding of the International MultiConfrence of Engineers & Computer Scientist, vol. I, 2011.*

[4] J. Pun & Y. Lawryshyn, "**Improving credit card fraud detection using a meta-classification strategy**," *International Journal of Computer Applications, vol. 56, no. 10, 2012.*

[5] R. Patidar & L. Sharma, **"Credit card fraud detection using neural network,"** *International Journal of Soft Computing & Engineering (IJSCE), vol. 1, 2011.*

[6] A. Srivastava & A. Kundu, "**Credit card fraud detection using hidden markov model,**" IEEE *Transactions on Dependable & Secure Computing, vol. 5, no. 1, 2008.*

[7] G. Singh, R. Gupta, A. Rastogi, M. Ch&el, & A. Riyaz**, "A machine learning approach for detection of fraud based on svm,"** *International Journal of Scientific Engineering & Technology, vol. 1, no. 3, 2012.*