IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
Vol. 08, Issue 11 (November. 2018), ||V (III) || PP 54-62

A Novel Machine Learning Approach to Detect Credit Card Fraud Using ECSVM

Anushree.B[1], Ramesh Kumar. B[2]
M.Phil Scholar, Department of Computer Science, Sree Narayana Guru College, Coimbatore, Tamil Nadu, India[1]
Assistant Professor, Department of Computer Science, Sree Narayana Guru College, Coimbatore, Tamil Nadu India[2]
Corresponding Author: Anushree.B

Abstract: Credit card fraud Detection is a serious and important problem in the current digital world for the banking customers and ecommerce websites. Customer's use the credit card in both ways one is direct transaction and other is digital transactions, direct transactions made through customer and billing department of the purchasing shop. But the digital transactions made through the website credit card user gives their details card number, expiry date, and ccv number in the website. Maximum credit card fraud transactions arise in the online transaction digital side only. To find out the solution for the credit card fraud detection while purchase transaction time in the proposed system use the adaptive boosting classification and majority voting techniques to discover the solution for credit card fraud detection. To develop a new algorithm ECSVM extended classifier support vector machine to detect the accuracy of the fraud transactions using the user credit card transaction dataset. This proposed system discovers the credit card dataset and calculate the accuracy of the financial transactions done by the customer. This system implements new decision tree based classification algorithm collects the details of the transaction such like as card number, card value, purchase amount, balance amount, transaction time, and all other customer profile details using this all values ECSVM techniques preprocessing the dataset and calculate the overall accuracy for the transactions and generate the class type for card user.

Keywords: AdaBoost, classification, machine learning, hybrid method, credit card, fraud detection, predictive modeling, voting.

Date of Submission: 18-11-2018 Date of acceptance: 04-12-2018

I. INTRODUCTION

The credit card fraud detection is an expensive problem for all financial institutions. In lot of domains efficient classified techniques to be built a fight fraud and other activities. The term fraud detection is defines as the abuse of a profit organization's system without inevitably leading to direct legal effects. In a competitive industry, credit card fraud can become a business biggest problem if it is very large and if the action of stopping procedures is not fail-safe. The financial transactions made in e-commerce field have very easy but the same time it is the risk related to fraudulent activities. It is a risk for transfer directly with the use of credit cards; taking into consideration that almost all the users will use the e-commerce that offer goods or services network space allows users to making transactions. In this proposed system consider a many number of techniques to manage this risk, although their effectiveness is failure for most common risks. The general principle used in almost all the modern approaches of fraud detection is significantly extent based on the comparison between the set of foregoing authorized transactions of a user and the new transactions under validation. In order to overcome this problem, a fraud detection approach should be able to use as much as possible information about the transactions during the validation process, but this is not always to be done due to the state of being unable of some approaches to manage some data. In this credit card fraud detection system using AdaBoost and majority voting's method. AdaBoost is to construct a strong model for frequently combining multiple weak models. In this AdaBoost and majority voting is used to apply the hybrid models. The main terms of this proposed system is to contribute a variety of machine learning techniques which predict the real world fraud detection in datasets. The fraudulent used a publicly available datasets. The data set is used to extract from the real credit card transactions for the last three months. This system machine learning technique is used for financial applications and detects fraudulent activity.

II. PROBLEM DEFINITION

In the current year more number of people can be using the ***credit card or debit card***. Because the card transaction is very fast and easy transaction and it give more advantages. But the credit card fraud is rising up so many different types of credit card scams increased. The credit card scammers are retrieve smarter so can have more tricks such as phone calls, emails to credit card skimmers and the Wi-Fi hotspots is to obtained the personal informations. If any fraudulent transaction on the credit card first thing should immediately contact the credit card company. The main elements of the credit card fraud is such as ***(a) Credit card theft, (b) Credit card forgery, (c) Credit card fraud***. In that credit card fraud is attending in four ways are, ***(i) Bankruptcy fraud, (ii) Theft fraud, (iii) Application fraud, (iv) Behavioral fraud.***

In this paper (Joseph King-Fung Pun, 2011) has presented the paper and improves the detection manner. In this paper introduce the concept of "Meta-Learning Strategy", which is used for improving credit card fraud detection. In this system is to reduce the big issues of the credit card frauds such as, (i) the transactions labeled or pointed out the fraudulent transaction it is fact manner, (ii) To predict the false alarms, (iii) The credit card transaction data from daily transaction and to determine the savings improvements, this concept based identifying the fraudulent transactions. The meta-classifier model is includes the three base classifiers are (i) k-nearest neighbor, (ii) decision tree, and (iii) naïve Bayesian algorithms. It follows the two main processes are (a) modeling techniques (Neural Network-NN), (b) updated data set.

In this paper (Maira Anis, Mohsin Ali, 2015) have presented the decision tree algorithms for class imbalanced learning in credit card fraud detection, which is reduces the financial problems. This concept is to apply comparative method in decision tree techniques. The paper concept introduce the term, "Resampling", it is related to the imbalanced data. In the paper concept, aim is to find out the best classifier based on distribution. In this concept is applied in to the RUS and feature selection for the family of classifier that is "Decision tree classifier". In this concept finally, provides the result is denotes the improved performance for the decision tree classifiers are already known, so for this system is very efficient to detect the fraud.

In this paper (Tamanna Chouhan, Ravi Kant Sahu, 2018) has presented data mining concept, "classification technique", which is used for credit card fraud detection. The mining is a term, to predict some that means, but the data mining is a technique, is to find out the hidden predictive information. The data mining introduce and describe the technique is called "Prediction analysis". The SVM is a classifier and it is proposed into the detecting the credit card fraud. In the concept is to take the input data and it is divided into test and training sets, and it is predicted the precision and recall.

In this paper (V.Dheepa, R.Dhanapal, 2012) has presented the SVM for credit card fraud detection. The concept is to detect the credit card fraud that means behavior based analyze the fraud using the ***support vector machine (SVM)***. In the developing world every day use the credit card so it is an unavoidable one, but in this time increase the frauds are already known. In this approach is adopted for efficient ***feature extraction method***. It analyze and predict the ***behavior transaction pattern***, if suppose this pattern differs from other to find out the fraud, that is doubted pattern is occur in behavior pattern it is predicted.

In this paper (Vijayalakshmi Mahanra Rao, 2013) have presented the decision tree induction for financial fraud detection by using the ensemble learning techniques. The credit card fraud problem is mostly affected the banking industries. The rise of web services give the advantages (banking) at the same time raises the banking frauds. The banking systems every second have robust, safe and secured one. It is order to detects and prevents the fraudulent activities of physical and virtual (any kind of) transactions. But the machine learning technique is to minimize the frauds. In this paper aim is to reduce the banking frauds. This system is used to (i) ensemble tree learning techniques, and (ii) genetic algorithm. These two techniques are to indicate the ensemble of decision trees, so the bank transaction datasets are identified and to prevent the bank fraud.

The (Wee-Yong Lim, Amit Sachan, 2014) kindly presented the concept [6] namely, "***conditional weighted transaction aggregation for credit card fraud detection***", which reduce the problem of substantial losses for credit card companies and consumers. The conditional weighted transaction aggregation technique describes to identify this issue used the supervised machine learning techniques, so it is to identify the fraudulent transactions. This technique is effective and better than existing system.

This concept (Wei Fan, Salvatore, 2012) has presented the main concept are "***AdaCost: Misclassification cost-sensitive boosting***", which is used to the cost of misclassifications updated by training distribution based on successive boosting rounds. This AdaCost is differs from AdaBoost. This concept main goal is to reduce the cumulative misclassification cost, it is more than AdaBoost. It moreover significantly reduce the total (cumulative) misclassification cost is over the method of AdaBoost (without consuming the additional computing power).

III. PROPOSED SYSTEM

Analyzing the credit card risk and finding correct solution is the major aim of this proposed system. To perform the above, eMtended classification support vector machine algorithm has been proposed. ECSVM

involves incrementally building an ensemble by training each new model instance to emphasize the training instances that previous models wrongly classified and misplaced. The proposed ECSVM often leads to a powerful improvement in predictive accuracy by providing an effective feature selection, when a data is classified and decision selected. The all applicable rules are found and help to calculate the priority; based on the priority the selection has been made. This finally progress in the two type of rule sets and their predictive accuracy

3.1 Advantages of the proposed System

- ECSVM has been developed to get better accuracy than eMisting combined classification svm and all other techniques in the eMisting system.
- It is an alternative decision support decision and accuracy detection tool, which helps to the financial institute and banking domain for accurate decision making.
- The proposed system overcomes the problem of over-fit the training data. And reduces training overhead.
- The proposed ECSVM algorithm has the ability to manage the credit card transaction class values.
- Performs card details like card number, eMpiry date, card verification number based feature selection for reducing the classification results.

3.2 Contributions

The system implements a new decision tree based classification algorithm with the use of effective enhanced clustering support vector machine to find the maMimum accuracy. The system introduces a new Bank transaction data Classification algorithm with enhanced ECSVM Advanced boosting and majority voting algorithm.

- This also creates a advanced boosting classification decision tree structure and find out the maMimum accuracy of the majority voting. The system developed with the intension of high accuracy and less training overhead.
- Analyzes the customer profiles
- Predicts the credit card transaction from customers profile
- Finds the customer types by their transaction. The customer's types are categorized into three types, general, valuable and risky.
- Identifies the highest transaction made by the customer per day or month and year, analyze their transaction details and find out the transaction is secured or fraud transactions.
- Brings the appropriate decisions to the credit card transactions based on the above features.

Data Preprocessing: Data Preprocessing involves removing the fake and manage the transaction values. In this preprocessing work the credit card details are verified with the set of rules in the eMisting samples.

Attribute Analysis: Database may also have the unsuitable transaction attributes. Association and correlation analysis is used to find out whether any two given attributes are related like as customer transaction limit for the card number and transaction amount.

Feature reduction: With the help of enhanced classified support vector machine clustering this can reduce the features for classification. So initially this need to be creates the dataset into initial analysis phase.

Classification and ECSVM: The data can be transformed to the classification process by using the following steps. The proposed system utilizes enhanced decision tree algorithm for decision making process.

IV. ECSVM Approach

The study is limited to the credit card transaction limit given by banking sector. Initially the system collects numerous customer records from UCI repository. The data set includes several fields related to the particular district. The details are enclosed in the previous chapter. And it also contains some card values of the transaction details. So initially the system performs data preprocessing. Proposed system generate the output of the pre-processed data were converted into a proper format to apply data mining concepts. The following fig 3.1shows the overall process involved in the proposed system. The first stage in the proposed work is the process initialization, where the data collection, selection and transformation process are done.

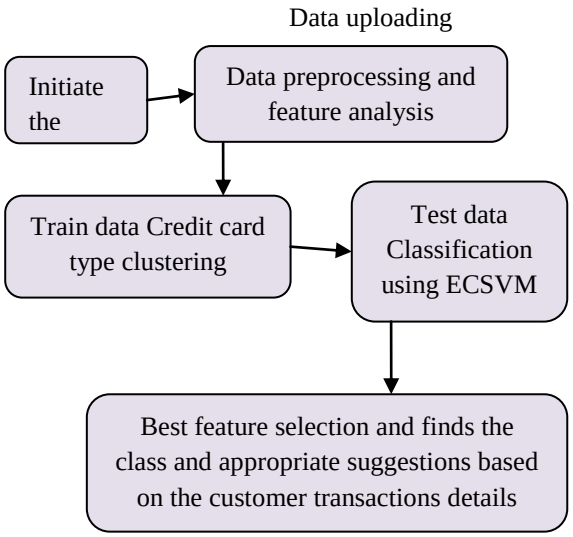

Fig 1.0 Overall processes of ECSVM

Data collection and uploading

The data used in this study was prepared from the financial department. The dataset included credit card details and customer profile details, location attributes (latitude, longitude), etc. Data was collected for particular district. The data format is presented in Table 3.1.

Attribute Information:

Table 1.0 Bank Customer Credit Card dataset

ID	Pan no	Amount limit	billing	Card no	Balance amount
1	AAAPL1234C	100000	35000	14785236985231	75000
2	ANSBH2356K	200000	30000	1856974123065	170000
3	BHIHS25689I	150000	20000	157894512352	130000

After analyzing the customer profiles and predicts of credit card limit for transactions, a method to predict the credit limit and also to estimate the transaction type by the customer type.The bank dataset of customer details which are required for data mining are collected and got familiarized with. Various attributes needed are also studied.

Data Preprocessing

Data preprocessing steps are applied on the new set of customer credit card transaction data and they are converted to categorical values by applying filters using unsupervised clustering algorithm named an enhanced classification support vector machine. After the operations are carried out, a total of input instances of individual locations are presented for analysis.

The attributes in the bank data set are filtered and the related attributes needed for prediction are selected. After that the complete transaction records in the dataset are analized and prepared for mining.

Data Selection

At this position, data applicable to the analysis was decided on and retrieved from the eMtracted customer data set. The eMtracted customer data set had several attributes; their type and description are presented in Table [3.5].

Input variables:

bank credit card data:

1 – Pan Number (numeric)

2-Transaction Limit : in this transaction limit intimates the maMimum purchase amount limit provided to the customer from the banking sector.

3 Billing amount: in this billing amount maintained the purchase amount by the customer while purchasing and amount entered by the billing section.

4- Card number: in this card number given to the customer by the banking side, it contains 16 or 18 digit numerical number in the credit card.

5 – Balance amount: in this field contains the balance amount available from the credit card after successful transactions.

6 – Transaction type: this transaction value contains the type of the transaction like as shop or ecommerce, shopping.

7 – Purchase type: purchase type values in the customer purchase type

8 - Contact: contact communication type (categorical: 'cellular','telephone')

9 – Customer type: credit card customer type (categorical: 'general', 'valuable', and 'risky')

10-class: this class field contains the numerical values 0 and 1, if the transaction class value contains 0 is secure transaction, or else contains 1 class value it's an fraud transaction.

Customer data descriptions

In this stage the system is developed in an efficient and user-friendly manner so that even those users with minimum technical knowledge can also use it generally. The system provides the most related attributes that help in formative whether to secure or fraud transaction in the credit card transaction. This aids in predicting the security of future customers.

The data set used in this study is split into classifying and testing data sets. All training cases are set by default taking into account the banks' guidelines for personal credit card details in the banks. Data used is of 1000 customer's data. The data required for the current study was collected from UCI repository. It contains of the different private variables and one reliant variable. Variables are the conditions or characteristics that he researcher controls or observes. It is needed to optimize variables by using ECSVM. Variables are classified as dependent and independent variables. An independent variable is the condition or characteristic that affects one or more dependent variables: its size, number, length or whatever eMists independently and is not affected by the other variable. A reliant variable modifies as a result of changes to the private variable. Private Variables: Using this data set, a model is built, which consists of a decision tree model ECSVM to predict whether a future applicant's a credit card is secured or fraud activity attempted. This chapter can use the decision tree node to classify observations by segmenting the data created according to a series of simple rules. This can use the entropy gain reduction method to build the tree. The return node fitted the logistic return model to the data. The ECSVM successfully implemented to the banking domain.

Proposed algorithm: ECSVM

After the segmentation, the system performs the rule set definition by implementing the mean, median and variance, the correlation is calculated using Pearson distribution.

Algorithm ECSVM **(labeled eMample S, set of variables M)**

Input: A set S of labeled eMamples, a set M of variables.

Output: Feature set

1. Let B = empty.
2. Get the training data D into C subsets Dc by the class value c or customer profile attribute.
3. for each training data set Dc

- Compute the Mean M (M_i;M_j) and the Mode φ(M_i;M_j) between each starting to end tuple of variables M_i and M_j.
- Compute W (M_i) for each variable M_i.

4. Return B.

The Bank decision tree algorithm made a number of changes to improve C4.5 and C5 algorithms some of these are:

• The proposed system handles training data with transaction values of attributes. Finally we get the calculation will be more accuracy and valueable.

• Handling attributes with separate and continuous values Let the training data be a set R= R1, R2 ... of already classified samples. Each sample Ki = kl , K2... is a vector where Kl , K2 ... represent attributes or features of the sample.

The training data is a training vector TV= TV1, TV2..., where TV1, TV2... represent the class to which each sample behinds also. At each node of the tree, D4.5 chooses one attribute of the data that most successfully splits data set of samples T into subsets that can be one class or the other. It is the standardize information secure that results from choosing an attribute for splitting the data. The attribute factor with the highest regulated information values is considered to make the right decision for the selected test samples.

V. RESULT AND ANALYSIS

The experiments are basically designed so that the different parts of the work could be evaluated easily and effectively. To this aim, first the fraud detection important terms are discussed. Second, all the dataset is processed and apply the different approaches are theoretically analyzed over the classification model entire dataset completely. Finally, algorithms are implemented this proposed work was implemented using the C#.net programming language. The performance of this proposed work scheme was compared with the existing algorithms based on the following parameters

5.1 Data Sets

The experiment uses the real time data and as well as synthetic data sets for experiments result. In particular most of the scenario, initially the data set has been collected from different web sources. The credit card customer dataset are collected from the UCI web repository etc., the dataset includes several attributes' such as location, customer details and card information and appropriate decision are collected from the literature. The system can have n number of tuples for experiments.

Dataset 1: (a) Customer Feature dataset with several properties for three types of customers named as Normal and fraud.

Dataset Description:

Dataset Name: credit card

URL: http:\\www.UCIrepostory.org\ credit card

Table 2.0 Dataset Description

Metrics	**Dataset**	**Existing AdaBoost**	**Proposed ECSVM**
	DS1(100)	95	99
Detection Accuracy (%)	**DS2(150)**	93	98.8
	DS3(200)	93	98.5
	DS4(250)	90	98

Table 2.0 Dataset Sample

Code	Description
DE002	Primary Account Number(PAN)
DE004	Amount, transaction
DE006	Amount, cardholder billing
DE011	System trace audit number
DE012	Time, local transaction
DE013	Date, local transaction
DE018	Merchant Type
DE022	Point of service entry mode
DE038	Authorization identification response
DE049	Currency code, transaction(ISO 4217)
DE051	Currency code, cardholder billing(ISO 4217)

The data set has 11 attributes. The experiment takes credit card fraud detection dataset from UCI repository. The dataset contains 11 attributes considered are: PAN, amount, Transaction, cardholder billing, Time, local transaction, Date, local transaction, Point of service entry mode, etc. There are a total of 5000 transactions are recorded in the database.

Experimental Results

This section describes the implementation process. Implementation is the realization of an application, or execution of plan, idea, model, design of a research. This section explains the software, datasets and modules which are used to develop the research.

Then the experimental term is performed on an AMD Sempron (tm) 140 Processor 3.70 GHz with RAM capacity 4 GB. The algorithms are implemented in Dot net along with sqlserver2008.Finally proposed algorithms are implemented implemented using C#.net.

.**A. prediction accuracy -** Determines the correctness of fraud detection

B. Time taken- Determines the processing time involved.
C. Precision- Repeated process same result.

5.2 Performance Evaluation
This experiment has been done through the credit card Dataset. The dataset is processed by ECSVM. The proposed system detects objects and its classes with significant improvement in terms of high classification accuracy this can be analyzed with different set of data's and results are shown in Figures 2.0.
Performance comparison of proposed ECSVM with existing approaches based On Credit card Detection Result accuracy

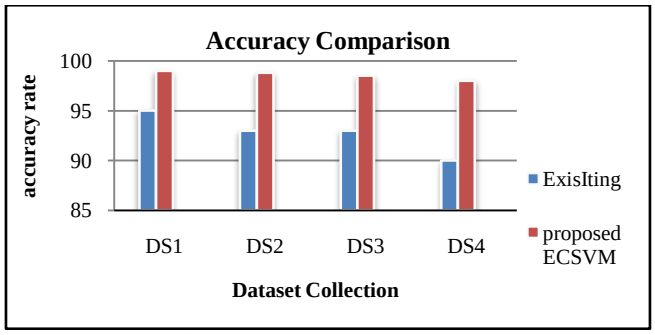

Fig 2.0 Accuracy Comparison

From the results shown in the graphs, it can be observed that the proposed ECSVM based approaches provides better accuracy and increased true positive rate when it is analyzed with different number of datasets. The system finally performs the analysis to show the accuracy of the proposed system.

Processing Delay
In this section describe about the comparisons of existing method1 AdaBoost, and the proposed method ECSVM. Based on the three metrics such as processing delay, iterations and result accuracy the following comparisons are made. Performance comparison of proposed ECSVM with existing approaches based on processing Delay

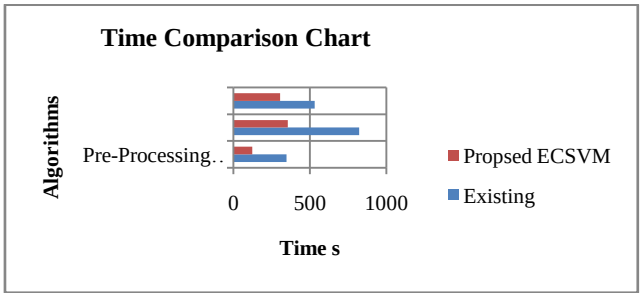

Fig: 3.0 Time comparison between

Existing and proposed
This time graph observed that the performance is very promising compared to the existing methods that have been explored in the previous chapter. The next chapter deals with the presentation of the conclusion and enhancements.

Result Iteration and accuracy:
Performance comparison of proposed ECSVM using stacking with existing approaches based on Result accuracy. In this study for conducting experiment the experimental phase used two different data sets, they are dataset 1, dataset 2files, the dataset1 is collected from the proposed website and the second dataset2 has been used with synthetic dataset. The performance study of the proposed method is compared with two different existing AdaBoost. The metrics used for comparison, so it comparing to the accuracy, precision, recall, f-measure and time taken.

The accuracy is calculated using the equation.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

TP =Number of true positives, TN =Number of true negatives, FP =False positives, FN=False negatives.

Precision or Confidence as it is called in the Data Mining component so that is denotes the proportion of Predicted Positive cases that are correctly Real Positives. Nonetheless, analogously called True Positive Accuracy (TPA), it measures the accuracy of Predicted Positives.

In difference with the rates of discovery of Real Positives (taper):

$$precision = \frac{No.ofcorrect\ risky\ customers\ Predicted}{Totalno.ofrisky\ custoemrspredicted}$$

Recall or the Sensitivity as it is called in Psychology; it is the proportion of Real Positive cases that are correctly Predicted Positive. To measures the Coverage of the Real Positive cases by the +P (Predicted Positive) rule. In this desirable feature is that it reflects how many of the relevant cases the +P rule picks up:

$$recall = \frac{No.ofaccurate\ risk\ values\ Predicted}{Totalno.ofrosky\ customerspredicted}$$

F-measure is a measure of a test's accuracy. It considers both the precision *p* and the recall *r* on the test to compute the score: *p* is the number of correct results divided by the number of total returned results and *r* is the number of correct results divided by the number of results that should have been returned. The F_1 score can be interpreted as a weighted average of the precision and recall, where an F_1 score reaches its best value at 1 and worst score at 0:

$$F - measure = 2\frac{precision * recall}{Precision + recall}$$

Table 3.0 Iteration comparison table

Metrics	**Existing AdaBoost**	**Proposed** ECSVM
iterations	5	3

Iterations Comparison Chart:

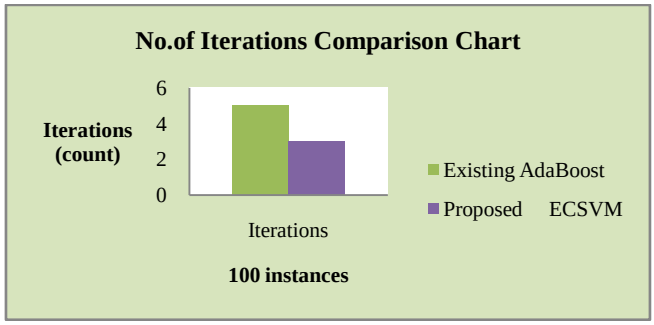

Fig 4.0 Iteration comparison chart

From the chart it shows the performance measure based on the iterations taken for the accurate decision selection in the existing and proposed approach. ECSVM took less iteration while comparing the other methods.

VI. CONCLUSION

Banking credit card fraud risk analysis is become difficult to learn, when the user doesn't know anything about the customer prior information. In this robustness is by reason of the high dimensional and dynamic data. This thesis presented an overview of the banking customer credit card transaction data required for detect credit card fraud and provides an interactive GUI based tool to analyze the credit card fraud detection using machine learning algorithms. This includes the enhanced ECSVM algorithm for fast detection process. This finds the appropriate solutions and find fraud detection based on the given attributes and values. Conducting the experimental research with various conditions and using the factors then to evaluate the output of the proposed system. The study proposed a new classification and prediction scheme for bank credit card data. The system studied the main two problems in the literature, which are prediction accuracy and classification delay. The study overcomes the above two problem by applying the effective enhanced ECSVM. The experimental results are evaluated using the C#.net.The experimental result shows better quality assessment compared to traditional techniques. From the experimental results, the execution time calculated for classification object is almost reduced than the existing system.

REFERENCES

[1]. Joseph King-Fung Pun. "Civilizing Credit Card Fraud Detection Using A Meta-Learning Strategy", Diss. 2011.
[2]. Maira Anis, Mohsin Ali, Amit Yadav. "A Proportional Study Of Decision Tree Algorithms For Class Imbalanced Learning In Credit Card Fraud Detection". International Journal Of Economics, Commerce And Management. Vol. III, Issue 12, December 2015.
[3]. Tamanna Chouhan, Ravi Kant Sahu. "Classification Technique For The Credit Card Fraud Detection". International Journal Of Latest Trends In Engineering And Technology Vol.(10)Issue(2), Pp.283-286. April 2018.
[4]. V.Dheepa, R.Dhanapal. "Performance Based Credit Card Fraud Detection Using Support Vector Machines". Ictact Journal On Soft Computing, Volume: 02, Issue: 04, July 2012.
[5]. Vijayalakshmi Mahanra Rao, Yashwant Prasad Singh. "Decision Tree Induction For Financial Fraud Detection Using En Masse Learning Techniques". Proceeding Of The International Conference On Artificial Intelligence In Computer Science And ICT (AICS 2013). 2013.
[6]. Wee-Yong Lim, Amit Sachan, Vrizlynn Thing. "Conditional Weighted Transaction Aggregation For The Credit Card Fraud Detection". IFIP International Conference On Digital Forensics. Springer, Berlin, Heidelberg, 2014.
[7]. Wei Fan, Salvatore J.Stolfo, Junxin Zhang, Philip K.Chan. "Adacost: Misclassification Cost-Sensitive Boosting". Icml. 1999.

Anusree.B. " A Novel Machine Learning Approach to Detect Credit Card Fraud Using ECSVM." IOSR Journal of Engineering (IOSRJEN), vol. 08, no. 11, 2018, pp. 54-62.