An Approach to Detect Credit Card Frauds using Attribute Selection and Ensemble Techniques

Shivangi Sharma
Student
Baba Banda Singh Bahadur Engineering College
Fatehgarh Sahib

Puneet Mittal
Assistant Professor
Baba Banda Singh Bahadur Engineering college
Fatehgarh Sahib

Geetika
Assistant Professor
Guru Nanak Dev Engineering College, Ludhiana

ABSTRACT

Managing of an account part is an essential area in our present day era where practically every human needs to manage the bank either physically or on the web Credit-card fraud prompts billions of dollars in misfortunes for online shippers. With the advancement of machine learning calculations, analysts have been finding progressively complex ways to identify extortion, yet handy usage is infrequently detailed. In this paper we are working to identify the fraudulent accounts using classification algorithms and then to improve the accuracy of results using feature selection technique. Bee search and genetic algorithms has been used to select relevant features from large dataset. The reduced dataset has been studied for different aspects. The ensemble learning techniques are implemented to reduce the variance. The impact of bagging, stacking and voting present the optimal technique for fraud detection.

Keywords

Data mining attribute selection, classification, Ensemble techniques.

1. INTRODUCTION

Utilization of credit cards for online buys has significantly expanded and it caused a blast in the credit card extortion. Credit card extortion incorporates unlawful utilization of card or record data without the learning of the proprietor; subsequently it is a demonstration of criminal misleading. Many papers announced enormous measures of misfortunes in various nations. The term credit is used to describe the method of purchasing and vending supplies devoid of having money. Credit card is a small plastic card to present the credit service to client[1]. Credit card is very accepted and plays an imperative role in electronic exchange and online capital business area which is emerging every year. Credit card is considered as extremely decent focus of extortion as in brief term of time fraudsters can get part of cash without getting into much hazard[2]. To present credit card extortion, fraudsters become aware with data like Visa number and government supervised savings passwords. Credit card extortion is extraordinarily decisive issue as it holds parcel of cash measure of banks. Many supervised and unsupervised learning strategies have been employed to decide the forged exchanges and factual exchanges[3][4]. Out of these techniques, supervised learning computations give more accuracy[5]. Francisca Nonyelum studied data mining applications for credit card frauds. Neural networks were used to analyze the dataset. The technique used was unsupervised clustering. Four clustering techniques were used to detect the fraudulent and legitimate transactions form dataset [6] . Lin et al.[7]made a questionnaire to estimate fraud causing factors. The observations of this questionnaire showed that there are some factors that are said to be well suited to calculate frauds. Also they have studied different tools for studying different data mining techniques like logistic regression[8]. The observations of the researchers were as per according to the specialization of field. Many research papers describe various frauds and techniques for detecting these frauds[7]. Many algorithms like artificial neural networks, decision trees have been implemented to analyze the legitimate accounts.

Many studies have shown evolutionary algorithms [9][10] for fraud and spam mail detection. Many other nature inspired techniques were also studied in this context. The ant colony optimization algorithm was implemented to carry out analysis. They have used ANN for this study. These techniques were implemented due to versatility and ease of use.

2. LITERATURE SURVEY

Masoumeh Zarepoor et.al [11] suggested that Credit card fraud is increasing considerably with the development of modern technology and the global superhighways of communication. Credit card fraud costs clients and the economic corporations billions of dollars annually, and fraudsters endlessly try to find new rules and strategy to commit illegal actions. Thus, fraud detection systems have become necessary for banks and monetary foundation, to lessen their losses. The most commonly techniques used for credit card fraud detection are Naïve Bayes (NB), Support Vector Machines (SVM), K-Nearest Neighbors algorithms (KNN). The data set of USCD-FICO was analyzed. The performance of classifiers was examined and found that Ensemble classifier provide the most accurate results for fraud detection.

Nuno Carneiro et.al [12] described the development and deployment of a fraud detection system in a large e-tail merchant. They investigated the combination of manual and mechanical classification gives insights into the full upgrading process and evaluates different machine learning methods. They have inveterate that supervised learning methods are applicable to fraud detection in e-tail trader surroundings. The three machine learning algorithms tested (Logistic Regression, Support Vector Machines and Random Forests) provided fine results. Random Forests attained the maximum performance of the three.

According to Mohammad Sultan Mahmud et.al [13] over the past few years, credit card transactions have been experiencing considerably speedy expansion with the growth of e-commerce and shows marvelous promise of advancement in the future. Hence, due to explosion of credit card transaction, it is inevitable to secure transactions. The classification of an anomaly (bad transaction) as normal (good transaction) usually costs more than classification of a normal as anomaly. Data set of USCD-FICO data mining contest has been analyzed. They examined that Meta and tree classifiers give best results. The best results in terms of classification accuracy achieved by Bagging, RandomSubSpace, RotationForest, LMT, REPTree, and RandomCommittee.

Mohd Saberi Mohamad et. Al [14] suggested that the classification patterns require selection of subsets or some relevant feature patterns. The feature selection process is very significant which selects the edifying features for used classification process. This is due to the fact that performance of the classifier is at risk to the selection of the features used to build the good classifier from small or high facet data that are intrinsically noisy. They have worked on a capable feature selection method that finding and selecting informative features from small or high dimension data which enhance the classification accuracy. Genetic algorithm has been used to investigate and identify the impending revealing features permutations for classification. And hence the selected features have been used to justify the accuracy of SVM classifier.

3. CLASSIFICATION

Classification is the most frequently applied data mining procedure, which utilizes a set of pre-classified examples to enlarge a model that can classify the inhabitants of records at large. Fraud detection and credit risk applications are predominantly well apposite to this type of examination. The data classification process involves learning and classification. The techniques to be used for this purpose are Naïve Bayes, J48, random forest and k-nearest neighbor.

3.1 Naïve Bayes

Naïve Bayes classifier is an uncomplicated and prevailing algorithm for the classification task. Even if we are running on a data set with millions of accounts with some attributes, it is recommended to attempt Naïve Bayes approach. Naïve Bayes is supervised machine learning algorithm that uses training dataset with known target classes to forecast the class of prospect instances. In general words we can say that naïve Bayes technique presupposes the occurrence or lack of distinct attribute do not depend on the occurrence or lack of attributes in identical set. This technique is named as naive because it intelligently assumes the liberty of attributes specified the class. After that classification is done by using Bayes rule to check the probability of correct class. Naive Bayes is a type of classifier which uses the Bayes Theorem. It estimates membership probabilities for every class such as the probability that given record or data point belongs to a particular class. The class with the maximum probability is considered as the most liable class. This is also known as Maximum a Posteriori (MAP).

$$\boldsymbol{P(c|x) = \frac{P(x|c)P(c)}{P(x)}}$$

Here P(c|x) is posterior Probability of (target) class given attribute of class. P(c) is known as prior probability of class. P(x|c) is likelihood. P(x) is prior probability of predictor class.

3.2 KNN

The closest neighbor (NN) rule differentiates the classification of unidentified data point on the basis of its closest neighbor whose class is previously identified. K-nearest neighbor (KNN) algorithm is in which nearest neighbor is calculated on the basis of inference of k that designates the number of nearest neighbors to be measured to portray class of a sample data point. More than one closest neighbor is used to identify the class which is having given data points. The samples of data that are used should be present in memory at run time. These data samples are allocated with weights as per their distance from sample data. When given an unidentified sample, a k-nearest neighbor classifier investigates the pattern space for the k training samples that are closest to the unidentified sample. "Closeness" is defined in terms of Euclidean distance. The Euclidean distance between two points say,
X1 = (x11, x12, ……… ,x1n) and

X2 = (x21, x22,………,x2n), is

$$\mathrm{dist}(\mathrm{X1},\mathrm{X2}) = \sqrt{\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{x1i}-\mathrm{x2i})2}$$

3.3 Random Forest

Random Forest is collection of decision trees. Many decision trees are constructed and their results are compiled to get a final output. Whenever a new instance is classified it is put into as a tree in the forest. All the trees give their separate classification output for a class. The object having maximum output is selected. The classifier is having good computational speed and easy to use in case of large and unbalanced dataset having various attributes. The output obtained is accurate as it collect the differences in output from all the decision trees is forest and also includes more number of outputs to be included in final prediction.

3.4 Decision tree

Decision tree is an analytical model with a hierarchical or tree structure. It is used most in the area of classification and prediction methods. The assembly of decision tree based classifiers does not require any dominion knowledge or constraint setting, and therefore it is appropriate for exploratory knowledge discovery. The main advantages of Decision Trees are that this method provides a meaningful way of representing gained knowledge and hence makes it easy to extract IF–THEN classification rules. Decision trees are the most capable methodologies in learning and information mining. Let X is a tuple from any unknown class. Now X will be tested for decision tree. Now path will be made from root node to leaf that will be having calculated prediction for that tuple. The decision tree construction does not need to set any type of parameters so it is suitable to extract information from large amount of data. Some of the commonly known decision tree algorithms ID3, C4.5, and CART work on non backtracking approach by which trees are made in top down manner.

4. ATTRIBUTE SELECTION and ENSEMBLE TECHNIQUES

Attribute selection is the way toward finding the most applicable factors for a prescient model. These procedures can be utilized to distinguish and expel unneeded, superfluous and excess attributes that don't contribute or diminish the exactness of the prescient model. Generally, attributes are categorized as follows:

- Relevant

- Irrelevant
- Redundant

The techniques used for selecting relevant features in proposed work are genetic algorithm and Bee search.

4.1 Genetic Algorithm

Genetic Algorithm is a most important Heuristic Algorithm which imitates Darwin's theory of progression. The initial population needed at the time for start of algorithm is set of strings formed by generator. Every string describes solution for optimization problem.
Genetic Algorithm design is to include the following three important operators:

- Selection
- Crossover
- Mutation

The selection operator is usually designed to choose probabilistically good solutions (individuals with high Fitness Values) and remove other bad solutions. Here in this the individuals for next generation are selected.

Associated to each string fitness value is calculated. Fitness value describes the goodness for the results obtained. Genetic operator is assigned to transform set of strings to get high fitness value. Every individual string is copied from one to other as per the fitness value. In crossover there is gene combination of two parents to make new generation. The simple form of crossover is to cut the actual parent and combine it with randomly selected string. Mutation operation inhibits the process of combination of genes randomly selected for given chromosome.

4.2 Artificial Bee colony[15]

It is intelligence based algorithm and it mimics the behavior of honey bees. This model consists of three components: (i) food sources (ii) employed foragers (iii) unemployed foragers.

Food source describes the position of solution of the given problem. Food source can also be described as fitness of the solution.

Employee foragers: the employee bees are assigned a food source. A food source (or possible solution) is assigned to an employee bee. These employee bees are responsible for giving the information about food source to onlooker bees. After a food source has not enough resources anymore then the employee bee which assigned to that source become scout bees.

The unemployed foragers are of two types: onlooker bees and scout bees. Onlooker Bee: the onlooker bees are responsible for searching the better food sources around employee bees. Onlooker bees get information about food sources from employee bees and search for possible solution.

Scout bees: the food forces that are not searched by employee bees are reached by scout bees. They find the possible solution that are very far and become employee bees after they get a food source.

The Food source is chosen by artificial bee as per the probability values correlated to that food source $\boldsymbol{P_i}$ calculated by the following expression

$$\boldsymbol{Pi = \frac{fit_i}{\sum_{n=1}^{SN} fit_n}}$$

Also $\boldsymbol{fit_i}$ is the fitness value of the solution I comparative to the amount of the food source in the position i.
SN is actual number of food sources available which is similar to the number of employed bees or onlooker bees.

The ABC algorithm uses one more function to compute candidate food position from the old location.

$$\boldsymbol{V_{ij}=X_{ij}+\phi_{ij}(X_{ij}-X_{kj})}$$

Where k ∈(1,2,……SN) and j∈(1,2,3……..D) some random numbers chosen. Also $\boldsymbol{\phi_{ij}}$ is a random number that lies between [-1,1]. It is responsible for controlling the neighbor food sources. $\boldsymbol{X_{ij}}$ compares the different food positions described by a bee.

4.3 Ensemble techniques

Ensemble methods are the improved methods that combine multiple models to give more accurate results. Due to combination of multiple models that create multiple models these methods provide more accurate results[16]. The main motive of designing these methods is that here the model is ensemble as like a group of organizing team. The ensemble techniques used in this work are bagging, stacking and voting.

4.3.1 Bagging

It is a way to decrease the discrepancy of your calculation by generating supplementary data for training from your novel dataset. If we increase the size of our training set then it is not possible to increase the prediction of the model but it helps to reduce the variance. The ensemble technique is generally used with tree models but the analysis could be done with other models also. When used with tree models there are number of trees that are formed and the output is generated on the basis of majority vote of the trees[17]. Bagging classifiers are fast as they can simply handle unbalanced and noisy datasets.

4.3.2 Stacking

Stacking is also a method in which multiple models can be applied on same data at same time. Here no experimental formula is applied to data. The model generated is supplied to another Meta level and new approach uses it as input to other level.

4.3.3 Voting

Voting is used to generate majority votes from various models generated from the same data. We can apply multiple techniques on same data and separate models of that data are formed. Out of these models the best technique is selected on the basis of their majority, weight and average.

5. RESEARCH METHODOLOGY

Classification algorithms have been applied on dataset to detect financial frauds from dataset. The analysis has been done on dataset of European credit card holders. The dataset is highly imbalanced and having noise and redundant data. So it was necessary to remove the noise and to select the relevant attribute set that could give best results. Also it has been analyzed that in case of large datasets it takes more computational time and decrease efficiency of the classifier. So here two attribute selection techniques have been used to pick the most relevant attributes from data set. The genetic algorithm provides results on the basis of Darwin's theory. Also artificial bee colony algorithm is used. Both the algorithms have their own computational area on the basis of that analysis has been performed. The dataset is also analyzed using ensemble methods to provide better accuracy and efficiency.

6. Results and observations

The impact of various classification techniques and feature selection algorithms on given data has been analyzed. The impact of using bee search and genetic algorithms of feature selection has been analyzed. To analyze the effect of feature selection dataset of Credit Card Company has been taken. The dataset taken is very imbalanced. The datasets contains transactions made by credit cards in September 2013 by European cardholders. The impact of feature selection has been analyzed on different classification algorithms like Naive Bayes, IBK, J48 and Random Forest.

6.1 Impact on various classification algorithms

6.1.1 Impact on Naïve Bayes

Table 1 shows results of impact of attribute selection on Naive Bayes

classification Performance measures	classification	Using genetic algorithm	Using bee search
TP rate	0.952	0.96	0.997
FP rate	0.931	0.932	0.927
Precision	0.874	0.874	0.859
Recall	0.911	0.925	0.927
F measure	0.061	0.061	0.892
MCC	0.007	0.007	0.000
ROC	0.526	0.526	0.506
PRC	0.867	0.867	0.865

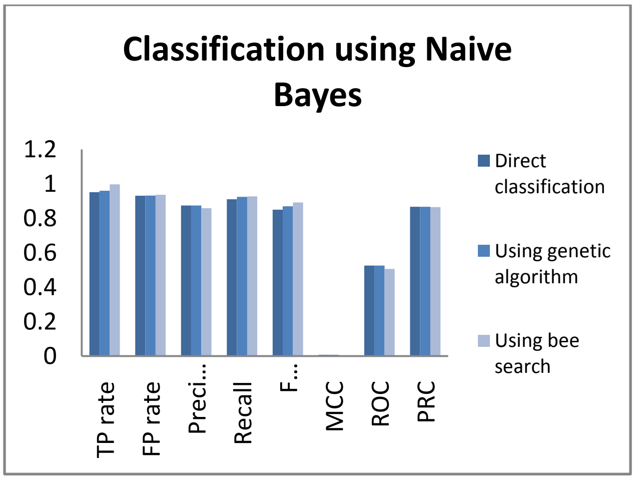

Fig.1 shows the impact of applying Naive Bayes before and after feature selection

6.1.2 Impact on IBK

Table 2 shows results of impact of attribute selection on IBK

classification Performance measures	classification	Using genetic algorithm	Using bee search
TP rate	0.911	0.927	0.927
FP rate	0.922	0.927	0.927
Precision	0.861	0.859	0.859
Recall	0.911	0.926	0.926
F measure	0.885	0.892	0.892
MCC	-0.022	-0.51	-0.51
ROC	0.52	0.509	0.509
PRC	0.865	0.866	0.866

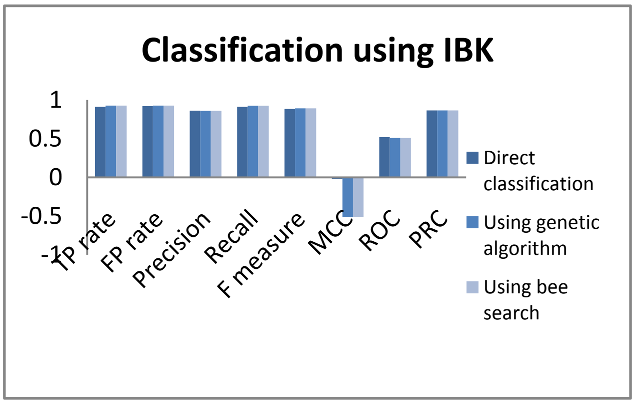

Fig.2 shows the impact of applying IBK classification before and after feature selection.

6.1.3 Impact on J48

Table 3 shows results of impact of attribute selection on J48

classification Performance measures	classification	Using genetic algorithm	Using bee search
TP rate	0.921	0.922	0.927
FP rate	0.911	0.927	0.927
Precision	0.842	0.858	0.859
Recall	0.911	0.927	0.927
F measure	0.870	0.888	0.892
MCC	0.000	0.000	0.000
ROC	0.496	0.496	0.496
PRC	0.861	0.863	0.867

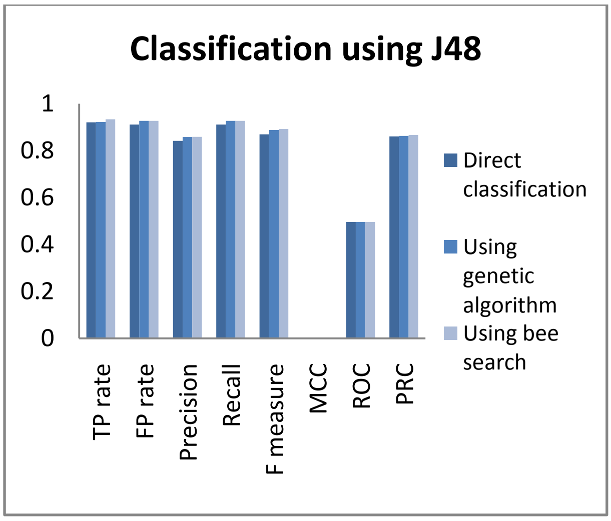

Fig.3 shows the impact of applying J48 classification before and after feature selection.

6.1.4 Impact on Random forest

Table 4 shows results of impact of attribute selection on Random forest

classification Performance measures	classification	Using genetic algorithm	Using bee search
TP rate	0.921	0.922	0.927
FP rate	0.911	0.927	0.927
Precision	0.842	0.858	0.859
Recall	0.911	0.927	0.927
F measure	0.87	0.888	0.892
MCC	0.000	0.000	0.000
ROC	0.496	0.496	0.496
PRC	0.861	0.863	0.867

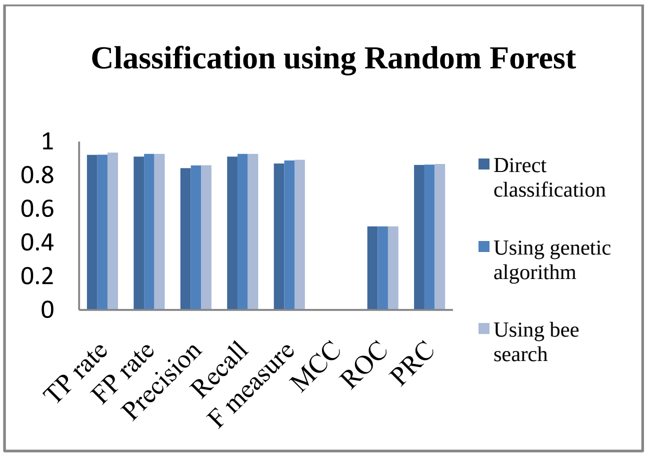

Fig.4 shows the impact of applying Random forest classification before and after feature selection.

The attribute selection methods used on the dataset selected the optimal features that have more impact on the results. These algorithms select the optimal features that help to detect the fraudulent transactions from dataset.

6.2 Impact of using Ensemble techniques

6.2.1 Bagging ensemble classifier

Table 5 shows results bagging classifier ensemble with other methods

Bagging ensemble with other algorithms				
Algorithms **Performance measures**	**Naïve bayes**	**IBK**	**J48**	**Random forest**
TP rate	0.738	0.991	0.927	0.927
FP rate	0.033	0.108	0.927	0.927
Precision	0.941	0.991	0.859	0.859
Recall	0.738	0.991	0.927	0.927
F measure	0.801	0.991	0.892	0.892
MCC	0.38	0.932	0.000	0.000
ROC	0.987	1.000	1.000	1.000
PRC	0.996	1.000	1.000	1.000

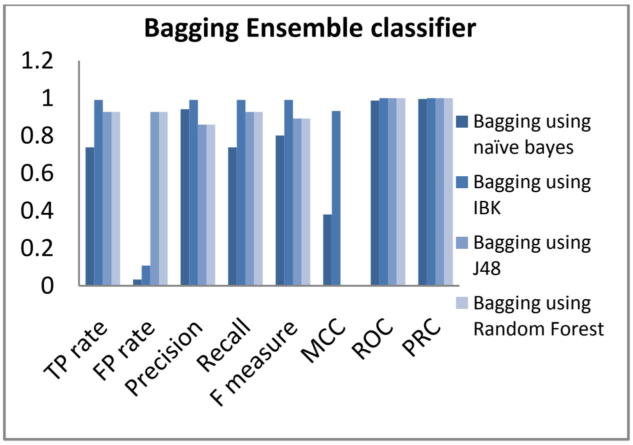

Fig.5 shows analysis of bagging classifier ensemble with other methods

From the above analysis it is clear that bagging classifier works best with tree models. The algorithm works well even if dataset is highly imbalances and reduce the variance and over fitting also.

6.2.2 Stacking and voting

Table 6 shows results of impact to ensemble various classification algorithms

Algorithms **Performance measures**	**Voting**	**Stacking**

TP rate	0.912	0.927
FP rate	0.922	0.927
Precision	0.861	0.859
Recall	0.912	0.927
F measure	0.885	0.892
MCC	-0.021	0.000
ROC	0.524	0.496
PRC	0.867	0.864

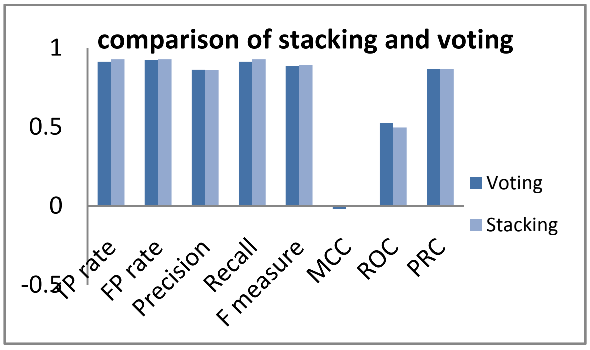

Fig.6 shows the comparison of results of stacking and voting

7. CONCLUSION AND FUTURE SCOPE

As with the advancement in technology there is increase in number of frauds in banking sector. There are number of reasons for this. So there is a need of system that could help to detect the illegal and legal monetary transactions. If we are able to detect such illegitimate transactions there are less chances of loss. In this paper the comparative analysis is done on various classification algorithms. As the dataset used was very imbalanced and noisy. Also there were large numbers of attributes that affect the computation. So it was in need to use techniques to reduce the dataset according to the relevance of attributes. The proposed methodology provided more accuracy in terms of precision and also shortened the time required for model building and computation. The ensemble models select majority voting and meta level learning techniques for the justification of results.

Further work can be done in this field to optimize the selection methods. Optimization for the selection function can be done. Also sampling approach can also be used to remove class imbalance.

8. REFERENCES

[1] Delamaire,L., Abdou,H. A. H. andPointon, J. Credit card fraud and detection techniques: a review 2009.

[2] Bhatla,T. P.,Prabhu,V., andDua, A. 2003Understanding Credit Card Frauds.

[3] Pippal,S.,Batra, L., Krishna,A., Gupta,H., and Arora,K. 2014Data mining in social networking sites: A social media mining approach to generate effective business strategies.

[4] Kaur,P., Singh,M., and Josan,G. S.,2015 Classification and Prediction Based Data Mining Algorithms to Predict Slow Learners in Education Sector.

[5] Abdallah,A.,Maarof, M. A., and Zainal,A. 2016Fraud detection system: A survey.

[6] OGWUELEKA,F. N.,2011Data mining application in credit-card Fraud detection system.

[7] Lin,C. C., Chiu,A. A., Huang,S. Y., and Yen,D. C.2015 Detecting the financial statement fraud: The analysis of the differences between data mining techniques and experts' judgments.

[8] Lee,K., Palsetia,D., Narayanan,R., Patwary, M. M. A., Agrawal, A. and Choudhary, A. 2011Twitter trending topic classification.

[9] M. Behdad, L. Barone, M. Bennamoun, and T. French, "Nature-inspired techniques in the context of fraud detection," *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.*, vol. 42, no. 6, pp. 1273–1290, 2012.

[10] Ravisankar,P., Ravi,V., Raghava Rao,G. andBose, I.2011 Detection of financial statement fraud and feature selection using data mining techniques.

[11] Zareapoor, M. and Shamsolmoali,P. 2015Application of credit card fraud detection: Based on bagging ensemble classifier.

[12] Carneiro, N., Figueira,G. and Costa,M 2017 A data mining based system for credit-card fraud detection in e-tail.

[13] Mahmud,M. S.,Meesad, P. and Sodsee, S.2016An evaluation of computational intelligence in credit card fraud detection.

[14] Mohamad, M. S.2004 Feature Selection Method Using Genetic Algorithm for the Classification of Small and High Dimension Data.

[15] Karaboga, D. and Akay, B.2009A comparative study of Artificial Bee Colony algorithm.

[16] Oza,N. C.2008 Ensemble Data Mining Methods.

[17] Bauer,E.2011An Empirical Comparison of Voting Classification Algorithms : Bagging , Boosting , and Variants.