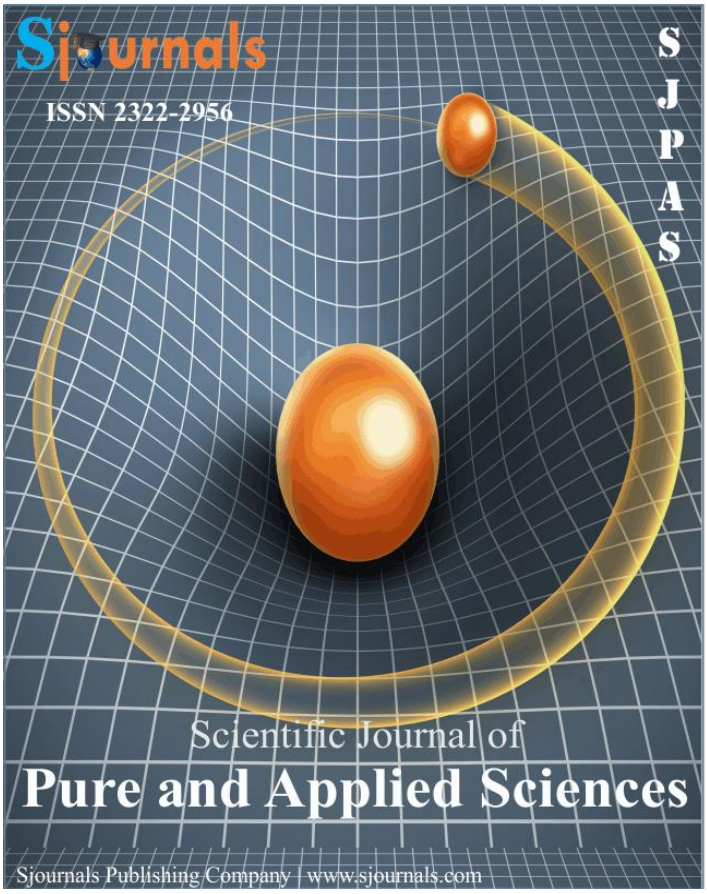
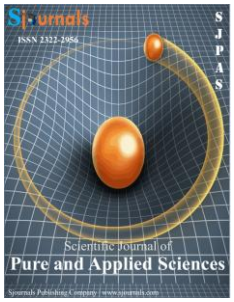


Scientific Journal of Pure and Applied Sciences (2017) 6(1) 544-555
ISSN 2322-2956
doi: 10.14196/sjpas.v6i1.2354



Contents lists available at Sjournals
Scientific Journal of **Pure and Appiled Sciences**
Journal homepage: www.Sjournals.com

Original article

A new method to detect deception in electronic banking using the algorithm bagging and behavior patterns abnormal users

Maryam Hassanpour[a,*], Ali Harounabadi[b], Mohammad Ali Naizari[a]
[a]*Faculty of Electrical and Computer, Institute Higher Eduction ACECR Khuzestan, Iran.*
[b]*Islamic Azad University Central Tehran Branch, Iran.*

*Corresponding author; ma.hassanpoor@yahoo.com

ARTICLE INFO

Article history,
Received 06 December 2016
Accepted 07 January 2017
Available online 11 January 2017
iThenticate screening 08 December 2016
English editing 05 January 2017
Quality control 09 January 2017

Keywords,
Electronic banking
Fraud detection
Clustering
Decision tree
Bagging algorrithm

ABSTRACT

Nowadays, large volumes of money transfers done in electronically channel and daily increasing grow in these services and transactions, on the one hand, and anonymity of offenders in the Internet on the other hand, encourage the fraudsters to enter to this field. One of the main obstacles in the use of internet banking is lack of security in transactions and some of abuses in the way of the financial exchanges. For this reason, prevent from unauthorized penetration and detection of crime is an important issue in financial institutions and banks. In the meantime, the necessity of applying fraud detection techniques in order to prevent from fraudulent activities in banking systems, especially electronic banking systems, is inevitable. In this paper, design and implementation system that recognizes suspicious and unusual behavior of bank users in the electronic banking systems. In this paper, we use data mining techniques to detect fraud in electronic banking. For this purpose, we use from a multi-stage hybrid method include: Clustering to separate customers and improve rankings and category for fraud detection. In the clustering method used from k center method and in the category method used from classification of C4.5 decision tree and also bagging's collective method of classification. Finally, the results indicate the high potential of the proposed method. The proposed method in compared with the previous method in the benchmark of accuracy

3.22 percent, in the benchmark of correctness 3.27 percent and in the benchmark of convocation 4.32 percent and in the benchmark of F1 3.81 been improved.

1. Introduction

Due to technological progress and technological development, possibility of fraud in various fields including banking, securities fraud and product fraud and other frauds is provide for profiteers. Fraud is one of important reason for failure in many organizations and also damaging to the capital markets; because investors and financial analysts in their decisions are rely on financial statements and trust them (Kashani, 2014). Fraud is targeted initiatives for gain financial illicit that contrary to the laws, regulations or usual policies (Bahador and Kazemi, 2010). In recent years financial fraud in banks and financial institutions has become a serious problem and attracted a lot of attention and concern to itself. Discovery of financial fraud in order to prevent the occurrence of devastating consequences arising from it are crucial. In other definitions, fraud considered as an abuse of profit of company or organization, regardless of its legal consequences. Fraud is also a process in which one or more persons for the sake of their personal interests deliberately excluded another from value of anything. Along with advances in information technology, Frauds and its variants are also expanding and it caused great losses for financial institutions and banks. Dishonest people due to weaknesses in the financial-bank electronic systems, enter to this systems and implementing their illegal objectives (Michalak et al., 2011).

Fraud is one of the factors that cause corruption in communities and keeps them from economic development. As in many countries has weak economic activity due to the use of illicit money, but what can shows ugly of fraud is criminal organs and their banking operations which sometimes as a flow outside the economic system, paralyzing the country's fiscal and monetary cycles. Since that bank is main core in protection network of financial system, thus the efficiency of an anti-fraud system depends on efforts of banks. Since that amount of create data in the banking industry with expansion of e-banking is increasing day by day, can with identify of data and analysis them, achieved to earlier detection of fraud. Due to the rapid growth of financial services and credit banks and various electronically financial institutions in the country, as well as increasing in use of electronic banking services by users, approaches of fraudsters towards e-banking is also on the rise (Kovach, 2011). Thus, lack of implement the mechanisms to detect and prevent fraud in electronic banking, we will see an increase in e-banking fraud.

Monetary and financial institutions are strictly tried to identifying activities of cheater. This is done because has direct impact on customer services for these institutions, reducing operating costs and stay as a trustworthy and reliable financial services provider. Therefore, using from techniques to detect fraud in order to prevent fraudulent practices in the banking system, especially e-banking systems, is inevitable. Intelligent criminal behavior, criminal's repetitive behaviors change is most important reasons that, inevitable the need a powerful and smart tools to identify and report the suspicious patterns among of customers mass behavioral information. The study is intended to use the unsupervised methods such as "Clustering" which can increase the accuracy of clustering methods in data collection. For this purpose first, data clustered and then classification methods apply on them. For this purpose, using from clustering algorithm of "k-medoids", C4.5 decision tree classification algorithm and also bagging cumulative techniques. Bagging's algorithm by different sampling with replacement, producing a random training series, that this type of sampling is known as the bootstrap sampling. With substituting different parts of data, different answers provide for education of categories. Using from many categories with alternative of data collection will bring a more accurately answer in this domain.

2. Theoretical consideration

E-banking is to provide opportunities for employees to increase speed and efficiency in providing banking services at the branch and interbank processes at all around the world and provide hardware and software facilities to customers that without having to be physically present in the bank at any hour of the day (24 hours), through secure communication channels, do their own banking operations (Majidi pour, 2011). In other words, e-

banking is use of advanced software and hardware technology based on networking and communications for exchange resource and financial information electronically and no need for physical presence of customer in the branch. Allow customers to conduct economic transactions on a secure website and small bank operations or virtual bank, credit and financial institutions or construction firms.

Nowadays with the development of modern technologies and global communications, Fraud is significantly on the rise and imposes large costs to business. Consequently, identifying fraud has become a very important issue. Financial systems based on information technology due to high potential in theft of money, often are easy goals for attackers. They use from multiple authentication flaws or weaknesses in security models that implemented in service and implement your goals. Weak authentication mechanisms that implement by signature, PIN, password and card security code cause be easier illegal financial transactions on behalf of attackers. As shown in Figure 1, generally fraud in the lifecycle of fraud can be used as a model, so that with analysis of that can will be given an appropriate response to fraud and again with the development of knowledge and proposing solutions and new protocols opened the way for fraudsters and new methods of cheating are formed and continues the life cycle fraud (Hatami Rad and Shahriari, 2010).

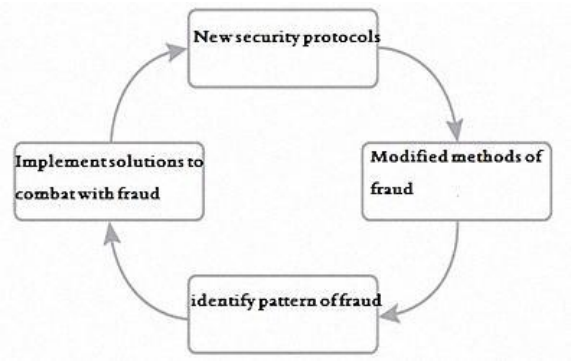


Fig. 1. The life cycle of financial fraud.

2.1. Decision tree C4.5

Trends of financial fraud generally are detected by analyzing and extracting information from the transactions database of financial institutions and this help to formulation and adoption of policies and new security protocols and authentication. C4.5 uses the benefit ratio of information to select the attribute of decision making. This algorithm provides values of attribute about predicted category.

- Make C4.5 decision tree
- Acquires information benefit for all adjectives
- Makes a decision node that is divided from best adjective
- Come back to list of divided adjectives and add this nude as child of divider adjective nude. This process recursively repeated on each subset of separation and when the separation is not more beneficial or can apply a category for all samples in achieved subset.
- Show C4.5 decision tree
- Each node inside the tree, test one adjective of sample
- Each branch out of internal node is appropriate for a possible value
- Each leaf node represents a category

2.2. Method of categories exit combination

There are different methods for combination of basic categories that most important includes: Majority vote, algebraic combinations, decision patterns, fuzzy integration, biz combination, space of knowledge and behavior. The most common methods are averages or use of majority vote. In general, it can be divided into two groups. First group are methods that work directly on the exodus of clauses so that they can obtain better results by combining the right clauses. Categories that participant to combine with this method usually have small numbers. The aim of second group is finding the best algorithm as well as best set of categories for compound. The number of categories used in these methods can be more. In this method often requires to a way to produce these items

with using a set of training data. Two popular techniques in this method to produce the categories is Bagging and Boosting (Polikar, 2006). In this article, we study the bagging method.

2.3. Bagging

Bagging method retrieved from Bootstrap rally is one of the simplest and at the same time the most successful method to improve the classification. This method is commonly used in the decision tree, but for other classification algorithms, such as Naive bayze, nearest neighbor and… can also be used. For example, a weak classifier, such as decision tree algorithm can be an unstable algorithm because a small change in the training data can be create a very different tree. The way to work with high dimensional volume data is very useful because in this case find a model due to the high complexity is not feasible. Bagging method for first time was introduced by Leo Berryman, in order to reduce the variance of a predictor (Wang, 2011). In this method, multiple copies with equal volume with volume of primary training collection data, randomly chosen and placement. Because the sampling of data collection is done by substituting, may some of the data several times appear in a sequence of training or did not appear in an education sequence. Each of these training sequences is used to train a weak classifier and constitute a model. The output of these models with using bagging techniques, are combined to obtain the final output (Syarif, 2012). This method is based on voting, and difference is that base learner is trained with different training data to have a slightly different. As a result, while this learner due to education will be identical with original set, due to random selection of sampling, training will also differ slightly. Bagging method for unstable learning algorithms namely Algorithms that with changes of data, change their results, Will have good performance (Neural network and decision tree are an example of these algorithms. The KNN is stable).

3. Related work

Banks with aware of their clients behaviors can prevent from theft, scams, fraud committed by clients. Banking fraud is a serious crime that necessitates the development of detection methods of transactions. Some research has been done, but the problem is not completely solved. Distinctive feature of conducted research in the field of prevent from fraud is different techniques and algorithms of learning as well as the application of these algorithms on a single, compound or group to classify of samples. In the following, we examined conducted research in the field of Fraud detection in banking.

Bahador et al. proposed a model with help of data mining with name of decision tree based on genetic algorithm that help to recognize of banks customer behavior from perspective of fraud and theft. Genetic algorithms can help by selecting the appropriate features and make a optimize decision tree to monitoring and CRISP validation of customers behavior. Classification model that proposed is based on clustering techniques, features choice, decision trees and genetic algorithm. This model focuses on selecting and combining the best decision tree based on optimality criteria and making the final decision tree to validate of customers (Bahador and Kazemi, 2010).

Vadoodparast et al. works based on detection of counterfeit electronic transactions by dynamic KDA model. Model KDA in this research; recognize 68/75% of online frauds and 82% of offline frauds. KDA clustering model is combines of three clustering algorithm include: K-MEANS, DBSCAN and AGGLOMERATIVE that are shown together as a dynamic solution. When be done a new transaction, the customer's behavior generated by this three algorithms, this means that every record used from three labels for anomaly detection. Each algorithm may use some or all of the parameters of the pre-processed. If the diagnosis were done by two or more algorithms show that the transaction is suspicious.

In above method, the final decision is considered based on comparing the output of all the algorithms together to reduce errors and increase accuracy. K-MEANS is fast and have great accuracy, but it has been fixed clustering. So the DBSCAN and AGGLOMERATIVE that have dynamic clustering is used. DBSACN is dynamic, but if fraud occurs outside the radius cannot be recognized, but K-MEANS and AGGLOMERATIVE can detect noise at all distances. AGGLOMERATIVE is dynamic but not fast enough and can put all objects in a cluster. When the number of parameters is large, K-MEANS and DBSACN have stop condition. So combination of all three algorithms has been used to better identify of fraud. When a new transaction was done, model of customer's behavior generated for these three algorithms and suspicious transactions placed between at least two algorithms. KDA model space common between these three algorithms and each algorithm tries to detect anomalies according to own way. The results of each algorithm separately written in database's tables and so you can easily compare abnormality

according to the results of the algorithms. Results suggest that detects KDA recognizing 96% of normal transactions (Vadoodparast et al., 2015).

Salehe Reza et al. provide a method that name is SARDBN that use from clustering and dynamic Bayesian network to Anomaly detection in suspicious transactions. SARDBN was used from dynamic Bayesian to get pattern from sequence of monthly transactions and calculates the anomaly index. Anomaly index is using of rank and entropy. Anomaly index calculates the degree of abnormality in a transaction and compared against a predefined threshold and specifies that the whether a transaction is ordinary or suspicious. The research on actual records has been tested and the results of that are shown. There are different approaches to fraud detection systems but mostly used from clustering to detect abnormalities and a few persons also used from neural networks, decision trees and support vector machine. In the above study used from a hybrid model based on clustering and dynamic bayesian networks for anomaly detection on transactions. Any deviation from this pattern considered as a deemed suspicious behavior. Transactions identify as a suspicious transaction that anomaly index exceeds from certain threshold (Reza et al., 2011).

4. Proposed methods

As we told according to importance of fraud in banking, in this article, we will look for a suitable method to identify abnormal banking users in order to prevent of fraud in electronic banking by clustering algorithms, ranking and bagging cumulative classifier. In the proposed method after clustering data collection used from decision tree classifier with different sampling with replacement from the set of training data. Then majority of voting technique was used to combine the results of classifier. The purpose of elementary clustering of data is find similar records that reduced values of the variables of each record as much as possible. In this article, was used hierarchical clustering algorithm of k-Medoids for clustering and decision tree algorithm C4.5 and C4.5v2 and as well as bagging cumulative algorithm. C4.5 decision tree with using from concept of entropy separates the data in a manner that selected best feature with minimize irregularities and selected threshold for the split data and decision-making about them. C4.5v2 decision tree can be used for testing. Basically can use from bagging techniques to assess the accuracy of estimates that applied in data mining methods through sampling with replacement from the training data. In this technique, it is assumed that the training data set are representative of population under study and variety of materialized scenarios of society can be simulated from this datasets. So, with using of twice sampling by employing multiple data sets, diversity is happening. When a new sample entered to each of the clauses, majority agreement applied to identifying the classes.

4.1. Data set

The statistical society in this study includes characteristics of electronic banking users. In this data set for each user is considered a feature vector based on banking record. Several features are announced for risk of customers and as well as being abnormal of their behavior by banks such as:

- Number of user errors when entering the system
- Number of internet remittances that was done by user
- Amount of internet remittances that was done by user
- Different IP number that is recorded in system during log in of user
- Hours of the day that system used by user
- Time of user familiarity with the system
- Type of user's browser in terms of the conventional
- Output: behavior that allocated to a user

In this article, used from eight reduced features so that last feature introduced as a target feature. Target feature have five values include: normal, a little suspicious, suspicious, very suspicious and dangerous. Due to terse of these features, not intend any reduction in features. By tracking the user features and emphasis on them can be identified user's behaviors. Number of input data in this research is 4000 and to evaluate the proposed method used from evaluation of 10-fold cross-validation method. Figure 2 shows the architecture of the proposed method.

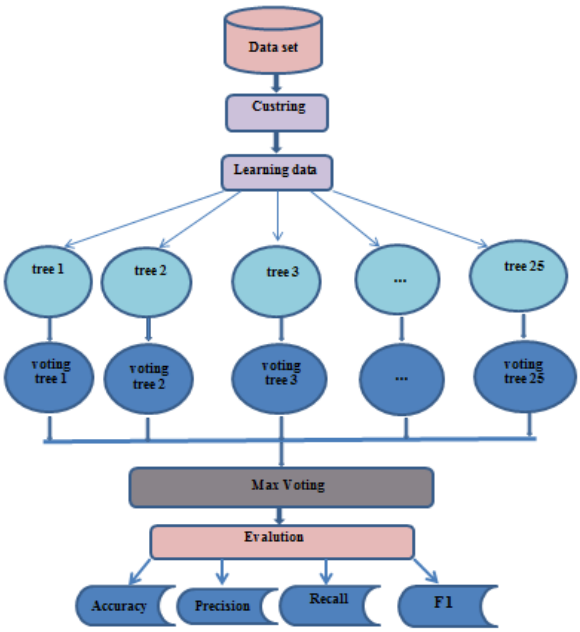


Fig. 2. The architecture of the proposed method.

4.2. K-Mediod's clustering algorithm for clustering data set in the proposed methodology

In this study, we want to use the methods of surveillance, such as clustering, to increase classification accuracy in data collections. For this purpose, first data are clustered, and then in a hybrid model, classification methods are applied on them. In this study, the aim of using data mining models such as clustering and classification to predict non-fraudulent and fraudulent behavior. After selecting algorithms, we should change and set their parameters, in order to increase their accuracy. For example, in the study of clustering method, K-Medio's method is used for clustering records. In this clustering method, parameter of number of clusters has to be specified the this can be done by setting different values for the parameters of the number of clusters. Then in each step that the value of this parameter was optimized, the number of relevant clusters is optimized. Silhouette index is used for validation of the clustering. The mean value of Silhouette index is used for evaluation of clustering validity and also for choosing number of optimized class, which this value calculated by observations and clusters distance**.** The value of S (i) is calculated using the following formula (Ghiyasi et al., 2015).

$$S(i)= \frac{((b_i)-(a_i))}{\max((b_i),(a_i))} \quad (1)$$

a (i) the average distance between observation i with other observations in a similar cluster and b (i) is the average observation distance i with all other observations in the cluster.

Based on the above formula, the value of S (i) am between -1 to +1. If S (i) is closer to +1, meaning that clustering of example is good and the proposed cluster is suitable for the sample, but if S (i) is closer to -1 means clustering of example is not well suited to data. By comparing the silhouette index, the value of parameter K is optimized with 4 clusters. Table 1 shows the optimal value of the parameter k.

Table 1
Optimal value for the parameter k.

Parameter k	Silhouette
2	0.8
3	0.3
4	0.9
5	0.7
6	0.5

4.3. Proposed method steps

After the optimal parameters for clustering and classification have been identified, three-stage operation takes place to determine user behavior on the data as follows:

First step: Clustering data based on the available features of the users, doing so have to be in a way that acceptable laws and results can be derived from that. Then, based on specified characteristics, determine existing clusters among customers. The number of clusters that has been set is 4 per label. Our data label is from ordinary to dangerous and the total number of clusters will be 20. New values of labels equivalent to the value of the new label (between 1 and 4) plus previous labels that multiplied by 10**.** So data labels with previous label 1 in the cluster 1 will be 11 and data label with previous label 2 in the cluster 1 will be 21 and so on. So they don't mix up with each other**.**

Second step: Training phase. Each sample from the data collection categorized by starting with root node of the tree and attribute test that specified by the node and moving to corresponding branch with specified attribute value and repeat this until reaching to leaf node. Also, when using decision tree models, required number of trees must be determined. In this research, we used different trees that if n = 25, they show high accuracy and if n is higher, considering evaluations there's no significant difference. The number of classifiers (trees) considered 25 and the practice is intended to be 10 fold. Bagging or group classification is based on high classifiers, which considered between 10 to 100 and there's 25 trees in each fold. The trees learn data with new labels; in fact we only used labels for easier learning of the tree. Each tree learns a series of samples. This feature ensures that each tree to become dominant on one data series and thus provide resulting diversity at the time of voting and increase confidence**.**

The third step: Review, determine and sorting behavior of users. Now we categorize all customers based on the cluster that have been set and first and second stage features. When a test sample enters a tree it simultaneously enters all of the trees and each of them comment the sample separately. The results of voting are stored. The reason we divide prediction labels to 10 is that in cluster section, the tens suggest actual label and the ones suggest cluster label. For example, tag 23 means the data belongs to little suspicious group (class 2) and cluster 3. In fact, the true label is 2 and clustering is used solely to improve the classification and at the end, Average results 10 fold for maximum voting that has been calculated.

Table 2
Number of appropriate classifiers of decision tree for proposed method.

The number of classifiers	Accuracy
10	80/65
15	83/21
20	89/78
25	96/14
35	96/08
40	94/54
50	95/83
80	96/34
100	95/74

4.4. Bagging technique to combine classifiers in the proposed method

In the proposed method, bagging technique is used for combining decision tree output classifiers, at every time, random sampling is done by replacing volume equal to educational series to create a new training complex. Each of the new training set, train existing classifiers that given to them. Classifiers output are combined with each other to reach an agreement with a majority of voting method. For each sample, each classifier export a vote, the highest vote represents its sample class of that record. For classification of sample x in the K class, each base classifier Ct produces one vote to class (Ct (x) = K) and the class that have highest vote will be selected as sample class x.

4.5. 10-fold cross-validation method in order to evaluate the proposed methodology

In anticipation of fraud in electronic banking, cross-validation method is used that in some sources also called rotary encoder, in order to estimate the accuracy of the proposed method**.** In 10 fold cross-validation method, the data set is divided into 10 equal parts. 9 parts used for training data set, and on the basis of those classifiers are made and with remaining part, the testing operation gets done. In order to assess, we must compare the output that classifier assign to training samples with the category that samples belongs to. The aforementioned process repeat 10 times, in such a way that each of the 10 parts, once used for the evaluation. In each iteration, the assessment criteria that have been made for model are calculated. In this way of evaluation the final accuracy of the proposed method is equal to the average of 10 calculated in each iteration. The most common value in that repeat this method in scientific texts is considered equal to 10. Obviously if this value is greater, then the amount calculated for classifier is more reliable and resulting knowledge will be more comprehensive, but, in this case increasing of classifier's evaluation time is the most important problem of it.

5. Experimental consideration

As mentioned in the previous section, in the proposed method, using random sampling with training data set replacement, several educational series are created. In this paper, after studying clustering has been used to improve classification and the proposed method**.** Implementation of the method takes place in three phases, in the first step clustering is made in order to improve the quality of the categories, and in the second step, classification algorithm is used and finally the method of majority voting.

5.1. Simulation experiment

In this research, MATLAB version R2014a is used for implementation. MATLAB is a high-level language and an attractive environment. This software is made by Math works company.

5.2. Criteria evaluation of the proposed method

Among the criteria used in assessing a classifiers we can mention precision, accuracy, classification error, calling and F1 (Hossin and Sulaiman, 2015). Continued on in this section we present how to calculate these criteria. The most important criterion for determining the efficiency of an clustering algorithm is the Accuracy. This criterion presents the total accuracy of classifier. This criterion reflects this issue, which what percent of the total data has been classified correctly, equation (2) shows how to calculate the accuracy.

$$\text{Accuracy}= \frac{\text{Total number of correctly diagnosed cases}}{\text{Total number of cases}} \quad (2)$$

The criterion of categories can be determined from equation (3). This relationship is opposite of the Accuracy. Most low value equal to zero is equivalent to best performance, and maximum value equal to one is equivalent to lowest performance.

$$\text{Error}= \frac{\text{Total number of incorrectly diagnosed cases}}{\text{Total number of cases}} \quad (3)$$

Precision criterion show percent of samples that from all samples that assigned to that by classifier are correctly classified. How to calculate this criterion is shown in equation (4).

$$\text{Precision}= \frac{\text{Number of correctly diagnosed for classi}}{\text{Number of diagnosed cases for classi}} \quad (4)$$

Recall criterion show percent of samples that from all samples that assigned to that by classifier, are correctly classified. How to calculate this criterion is shown in equation (5).

$$\text{Recall}= \frac{\text{Number of correctly diagnosed cases for classi}}{\text{Number of cases for classi}} \quad (5)$$

F1 or F-measure criteria is obtained from combination of Precision and Recall criteria and used in cases that we can't have particular importance for each of the two criteria for Recall and Precision. Equation 6 shows how to calculate this criteria.

$$F1= \frac{2*Precisioni*Recalli}{Precisioni+Recalli} \quad (6)$$

5.3. Evaluation of proposed method with K-Mediod's clustering

For evaluating the efficiency of the proposed method, the 10-part mutual evaluation procedure is used. In a 10-part method, in each section 90% of the data considered as educational data and 10% of the data as test data. performance of classifier at every step is reviewed by accuracy, authenticity, calling and F1 criteria. As well as average criteria for the evaluation of overall is also calculated**.** Table 3 is evaluation results of proposed method using the K-Mediod's clustering that represents high-performance of proposed method with average accuracy equal to 15/92%, average call 59/88%, average F1 is equal to 34.90% and authenticity criteria equals to 14.96%. Using cross-validation of 10-part evaluation method, the percentage of criteria for each section is calculated. Bar chart 1 presents different evaluating criteria on the proposed method.

Table 3
Evaluation results of proposed method using the K-Mediod's clustering.

Fold	Precision	Recall	F-measure	Accuracy
1	90/88	88/60	92/70	89/72
2	94/81	90/78	92/75	92/75
3	87/45	83/12	88/27	85/23
4	91/48	85/58	90/00	88/43
5	95/61	90/32	92/88	92/88
6	94/94	89/47	87/10	92/12
7	85/34	82/92	87/71	84/11
8	96/71	92/52	89/63	94/56
9	94/83	93/12	90/81	93/96
10	89/51	89/55	89/87	89/53

Average-precision of max voting: 92/15
Average-recall of max voting: 88/59
Average-F-Measure of max voting: 90/34
Average-accuracy of max voting: 96/14
Classification_error: 03/86

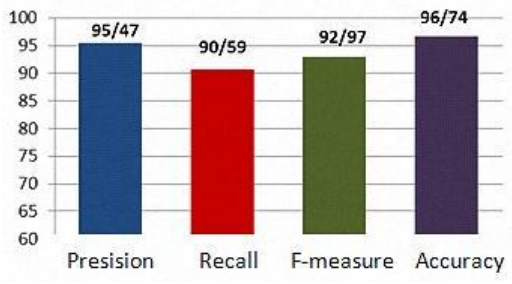


Chart 1. The results of the evaluation of different criteria by proposed method.

5.4. Comparison of proposed method with previous work

In this section, the results obtained from the recommended method for a comparison with some of the previous works. In the table 4 some of the previous tasks carried out for comparison with the proposed method are presented. To compare proposed method with previous works, we also used the same data set that is in the

proposed working method. 10-fold cross-validation of evaluation method of criteria has been examined by accuracy, Precision, Recall and F1 in accordance with the proposed methodology compared with previous work. As we can see the proposed method has higher efficiency, compared to the previous works about the accuracy and precision, Recall and F1. Chart 2 to compare the rate of accuracy, chart 3 to compare the level of precision of the criteria, chart 4 to compare the amount of Recall and chart 5 to compare the level of precision of the criteria in the proposed method and the previous comparison tasks in ten repetitions. The proposed method is compared to reference (Adeyiga et al., 2012), improved in precision criterion is 13.8% and the accuracy criterion is 65.9% and in Recall is 75.7% and in F1 is 93.7. Also, in comparison with the reference (Vadoodparast et al., 2015), is improved 22.3% in precision criterion and 27.3% in accuracy and 32.4% in Recall and 81.3% in the F1.

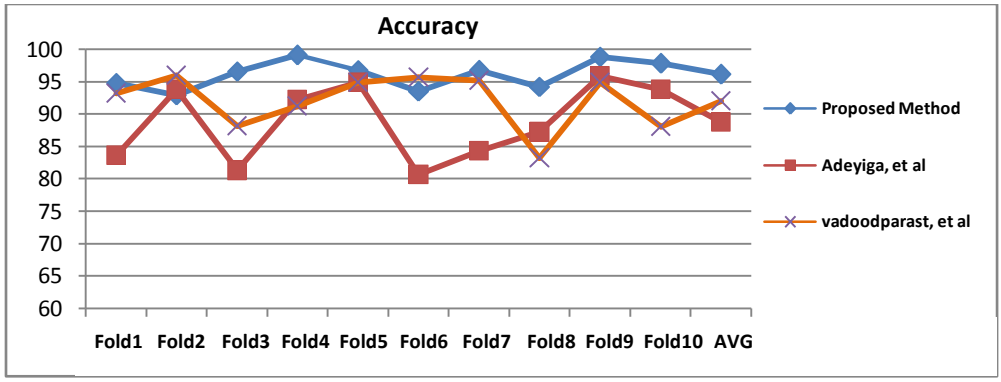


Chart 2. Compare the level of accuracy of the criteria in the proposed method and the previous comparison tasks in ten repetitions.

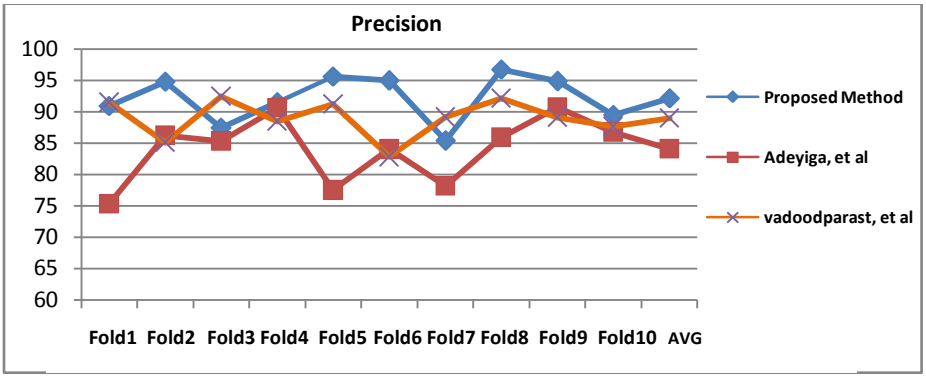


Chart 3. Compare the level of precision of the criteria in the proposed method and the previous comparison tasks in ten repetitions.

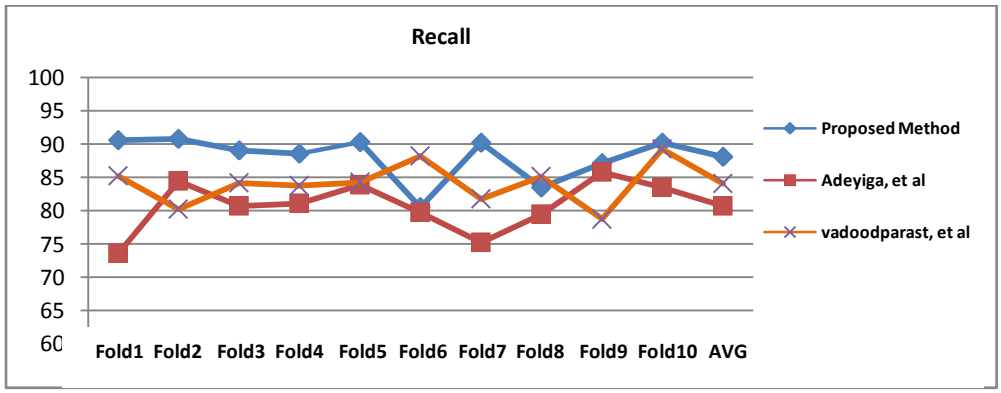


Chart 4. Compare the level of recall of the criteria in the proposed method and the previous comparison tasks in ten repetitions.

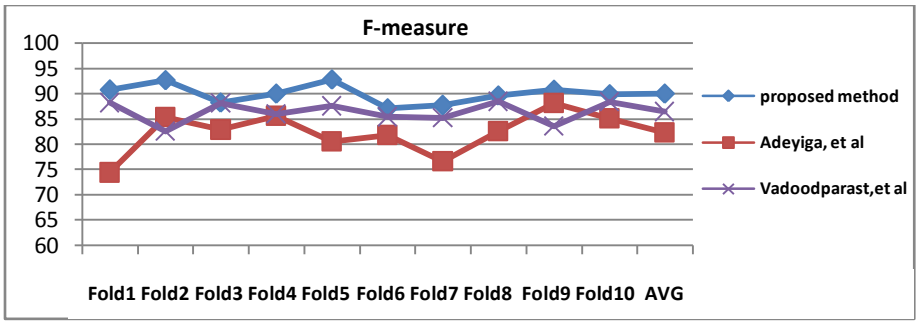


Chart 5. Compare the level of F1 of the criteria in the proposed method and the previous comparison tasks in ten repetitions.

Table 4
Comparison of proposed method with previous work.

Reference (Author / Year)	Approachs	Accuracy	Precision	Recall	F1
Porposed method	k-medodids,C4.5, bagging algorithm	96/14	92/15	88/49	90/28
Adeyiga et al. (2012)	Neural Networks	86/49	84/02	80/74	82/35
Vadoodparast et al. (2015)	K means,DBSCAN, Alggoromative	92/87	90/82	86/97	88/86

6. Conclusion and future work

The purpose of this article is to provide a way to prevent fraud in electronic banking so in this way we can recognize suspicious behaviors with high accuracy. Several algorithms are available for clustering and classification, which in this article, the K-Mediod's clustering algorithm and C 4.5 decision tree has been used. Method of clustering has been used to group the data set. The purpose of initial clustering of data is to find same records to decrease distinction between the values of variables of each record and thus allow classification methods categorize data easier based on this distinction. In the proposed method after clustering data collection that has been used, we used decision tree classifier with different sampling with replacement of the collection of educational data. Then we use majority of voting technique to handle composition of the classifiers, the largest number of votes for each class specifies the sample class. The proposed method with use of data collection to prevent fraud in electronic banking and using a 10-part cross-validation of the evaluation method and different criteria for accuracy, authenticity, calling and F1 is implemented and evaluated.

In this paper, the proposed method implemented and evaluated by the previous work in terms of performance. The basic model used on the data contained in each cluster, which was shown that clustering model with tree C 4.5 and bagging technique in predicting suspect behavior in data collection was more accurate from previous methods of the comparison. Results shows high efficiency of the proposed method with a mean precision equal to 15/92%, average call 59/88%, F1 34/90% and accuracy equal to 14.96%. The proposed method compared with reference 16, has improvement of 13.8% precision criterion and accuracy 65.9% and calling 75.7% and F1 93.7%. Also, in comparison with the reference 13, is improved to the extent of 22.3% precision criterion and 27.3% accuracy and 32.4% calling and 81.3% F1. The results of this research can be used to get more accuracy in the detection of suspicious behavior. The method that examined in this paper was a group method, which has good flexibility in the field of selection of algorithms so that it can be distinguished from other categories for other algorithms, especially the ones that have the instability of property. What of this research can be used as proposal for future research and works?

- Research on other features related to suspicious transaction

- The articles can use any other classification algorithms, especially algorithms that have the instability of property, such as decision trees, or other types of neural networks.
- Other combinations of techniques, including genetic algorithm, other types of decision trees and so on to obtain the rules can be used. As well as the composition of heterogeneous classifiers.

At the stage of combining classification algorithms we can also use Boosting technique which has not been raised in this article. Also, various other methods can be used in order to select the best characteristics or instead use the sampling method.

References

Adeyiga, J.A., Ezike, J.O., Omotosho, A., Amakulor, W., 2012. A Neural network based model for detecting irregularities in e-banking transactions. Afr. J. Comput. ICT Ref., 4(3,2), 7-12.

Ameri, F., Walden couples, M.J., 2007. The various techniques unsupervised clustering method. Geomatics 86, Tehran, national mapping agency, 1-10.

Bahador, H., Kazemi, A., 2010. A model for the identification of bank customers in terms of bank robbery with the use of data mining fuzzy.The fourth Iran Data Mining Conference, Tehran, Sharif University of Technology, 1-12.

Ghiyasi, F., Nezafati, N., Shokohyar, S., 2015. Clustering users of marine data by using data mining techniques. Letters processing and management, 30(4), 1037-1039.

Hatami Rad, A., Shahriari, H.R., 2010. Methods and strategies to detect fraud in electronic banking. Economic News, 134, 219-225.

Hossin, M., Sulaiman, M.N., 2015. A review on evaluation metrics for data classification evaluations. Int. J. Data. Min. Knowl. Manag. Process (IJDKP), 5(2), 01-11.

Kashani, S., 2014. Detect fraud in electronic banking using data mining. National Conference on Computer Engineering and Information Technology Management, University of Sistan and Baluchestan, 1-12.

Kovach, S., 2011. Online banking fraud detection based on local and global behavior. ICDS: The Fifth International Conference on Digital Society, 166-171.

Majidi pour, M., 2011. Evolution and common methods of electronic banking. Management magazine, 30, 37-41.

Michalak, K., Korczak, J., 2011. Graph mining approach to suspicious transaction detection. Proceedings of the Twenty-Ninth Federated Conference on Computer Science and Information Systems, IEEE, 69-75.

Polikar, R., 2006. Ensemble based systems in decision making. Circ. Syst. Mag., 6(3), 21-45.

Reza, S., Haider, S., 2011. Suspicious activity reporting using dynamic bayesian networks. Procedia. Comput. Sci., (3), 987-991.

Syarif, I., Zaluska, E., Prugel-Bennett, A., Wills, G., 2012. Application of bagging, boosting and stacking to intrusion detection. 8th International Conference on Machine Learning and Data Mining in Pattern Recognition, Springer, Berlin Heidelberg, 7376, 593-602.

Vadoodparast, M., Hamdan, A.R., Sarim, H.M., 2015. Ftaudulent electronic transaction detection using dynamic KDA model. (IJCSIS) Int. J. Comput. Sci. Inform. Secur., 13(2), 1-8.

Velmurugan, T., Santhanam, T., 2010. Computational complexity between K- Means and K-Medo ids clustering algorithms for normal and uniform distributions of data points. J. Comput. Sci., 6(3), 363-368.

Wang, G., 2011. A comparative assessment of ensemble learning for credit scoring. Exp. Syst. Appl., 38(1), 223-230.

How to cite this article: Hassanpour, M., Harounabadi, A., Naizari, M.A., 2017. A new method to detect deception in electronic banking using the algorithm bagging and behavior patterns abnormal users. Scientific Journal of Pure and Applied Sciences, 6(1), 544-555.