*Journal of **AI** and **D**ata **M**ining*
Vol 6, No 1, 2018, 219-231

FDiBC: A Novel Fraud Detection Method in Bank Club based on Sliding Time and Scores Window

S. M.- H. Hasheminejad* and Z. Salimi

Department of Computer Engineering, Alzahra University, Tehran, Iran.

Received 17 November 2016; Revised 27 February 2017; Accepted 18 June 2017

**Corresponding author: SMH.Hasheminejad@Alzahra.ac.ir(S.M.-H. Hasheminejad).*

Abstract

One of the recent strategies for increasing the customer's loyalty in banking industry is the use of customers' club system. In this system, customers receive scores on the basis of financial and club activities they are performing, and due to the achieved points, they get credits from the bank. In addition, by the advent of new technologies, fraud is growing in banking domain as well. Therefore, given the importance of financial activities in the customers' club system, providing an efficient and applicable method for detecting fraud is highly important in these types of systems. In this paper, we propose a novel sliding time and scores window-based method, called *FDiBC* (Fraud Detection in Bank Club), to detect fraud in bank club. In *FDiBC*, firstly, based upon each score obtained by customer members of bank club, 14 features are derived, and then based on all the scores of each customer member, five sliding time and scores window-based feature vectors are proposed. For generating training and test dataset from the obtained scores of fraudster and common customers in the customers' club system of a bank, a positive and a negative label are used, respectively. After generating the training dataset, learning is performed through two approaches: 1) clustering and binary classification with the OCSVM method for positive data, i.e. fraudster customers, and 2) multi-class classification including SVM, C4.5, KNN, and Naïve Bayes methods. The results obtained reveal that *FDiBC* has the ability to detect fraud with 78% accuracy, and thus can be used in practice.

Keywords: *Financial Fraud Detection, Club System, Banking Industry and Sliding.*

1. Introduction

Fraud is an illegal action through which a person earns a property without the permission of its owner; electronic fraud is one of the prevalent crimes growing currently and associated officials have not been so far able to uproot it. In fact, financial fraud detection means separating financial data of fraudsters from financial data related to ordinary people. With the advent of modern technologies, the techniques of committing these crimes have become more varied, consequently, trapping the culprits and proving their crimes have become more difficult. In 2013, the report of 1.44 billion fraud in European banks, and 8% growth rate compared with 2012, clarifies expediting the e-fraud growth rate, especially in the banking industry [1].

When it comes to banking business, one of the strategies for increasing the customer loyalty in banking is applying the customers' club system. In this system, customers receive scores on the basis of financial and club activities they are performing, and due to the scores obtained, they get credits from the bank. However, by the advent of new technologies, fraud is growing in the banking domain as well. Data mining [2] has been used in different areas such as diagnosing heart diseases [3], text-mining [4], designing software architecture [5-7], selecting design pattern [8, 9], and so on. One application of data mining is to detect fraud. Fraud includes the crimes of credit card transactions, money-laundering, etc. [10]. In fact, using data mining helps abnormal scenarios identification. As a strategy, data mining can be learned as patterns using the past fraud data, and then by employing those patterns, future fraudsters can be predicted. The techniques of detecting financial fraud in banking can be divided into four categories: credit card, money-laundering, fake transactions, and false accounts. One of the primary studies in this domain can be referred to a study by [11], in which a method has been offered for detecting fraud in credit cards using the "fuzzy logic" technique. Most studies in financial domain have been related to credit card swindling that use the neural network approach [12]. Duman et al. [13] have employed a genetic algorithm (GA) for detecting credit card fraud. In [13], to each transaction a score is given, and based on this score, transactions are classified as legal or illegal. In [13], the aim is to reduce the number of transactions that are mis-classified. In another study [14], two algorithms of association rules and clustering have been used, and these two algorithms have been applied on a dataset from 114 firms. Additionally, in [15], Hidden Markov Model has been used for detecting fraud in fake transactions. West et al., in a review paper [10], have divided the fraud types into six categories including credit card fraud, securities fraud, financial statement fraud, insurance fraud, mortgage fraud, and money laundering but they did not consider the new fraud happening in bank club systems.

In this paper, a novel sliding time and scores window-based method called *FDiBC* (*Fraud Detection in Bank Club*) is proposed for detecting fraud in the customers' club system on the basis of data mining. In *FDiBC*, it is assumed that fraudsters attempt to carry out fraud using uncommon transactions and fake communications in the customers' club system. In *FDiBC*, on the basis of instance data of the scores of bank customers' club system, classifiers are learned, and then these classifiers are evaluated. In *FDiBC*, both the clustering and classifier approaches are employed and its novelties are as follow:

- Proposing a novel method based upon clustering and classification to detect fraud in the customers' club system (the proposed *FDiBC* method is described with more details in Section 3)
- Providing new sliding time and scores window-based feature vectors for summarizing the customer scores
- Identifying the best classifier in detecting fraud from the five classification methods OSCVM, SVM , C4.5, KNN, and Naïve Bayes [2]

Note that there are several techniques for classification but it cannot be said which one is better than the others. The reason is that there is no learning technique that achieves good results for all problems. In real life, some of them may achieve good results for some problems and bad results for the rest. Therefore, for a new problem (i.e. fraud detection) to find out the best learning technique, we evaluate several learning techniques so that the best learning technique is identified. Thus one of the contributions of this paper is to identify the best learning technique compatible with the fraud detection.

In the rest of this paper, at first, the bank club system will be introduced in Section 2. Then we proceed to explain the proposed *FDiBC* method in Section 3. In Section 4, the results of evaluating the proposed *FDiBC* method will be mentioned. Finally, in Section 5, conclusions would be done and further works would be stated.

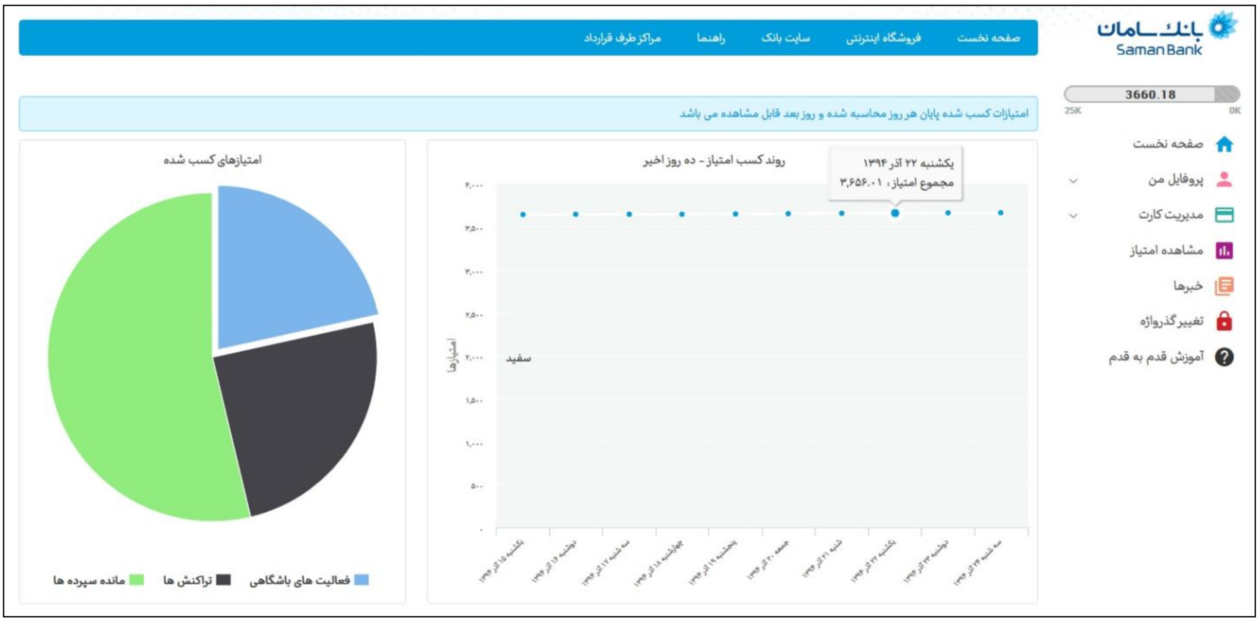
2. Bank club system

The loyalty of bank customers comes along with bank club systems. These systems, in an attempt to raise the level of customers' loyalty from its bottom and to establish effective and regular relationship with customers, provide new services appropriate with their needs to increase their satisfaction and faithfulness. Bank club systems have been usually designed for special and common customers. The customers obtain scores through activity in a bank using different ports such as Internet Bank, Mobile Bank, ATM, and POS. Then customers' loyalty cards are charged once their obtained scores reach a pre-defined limit. Furthermore, customers will be entitled to a discount for purchasing their given scores. Bank club systems usually have two panels for customers and management (both of which cover branches and administrators). Some of their most important features are mentioned below. The main features of management panel are as follow:

- Users management
- Types of deposits, their definition and management
- Types of deposit, transactional and club activity patterns, their definition and management
- Collective and individual registration of customers by bank as well as issuing and printing collective and individual password in the branches

The main features of customer panel are as follow:

- Completing the profile information
- Registration and card management
- Inviting friends
- News
- Observing the obtained scores (chart and graph)
- Sub-system of using the scores
- Issuing and managing loyalty card
- Management reports



امتیازات کسب شده پایان هر روز محاسبه شده و روز بعد قابل مشاهده می باشد

Figure 1. Customer panel of Saman Bank club system (In Persian).

In this paper, we use the archive customer scores of Saman Bank club as our case study. Saman Bank is one of the three famous private Iranian banks, and its club system has started since 2012. In figure 1, a customer panel of this club system is shown. Note that this system is displayed in Persian language. In figure 1, there is a pie chart revealing different kinds of obtained scores including deposit, transactional, and club activity.

3. Proposed *FDiBC* method

In this section, a proposed *FDiBC* method based on data mining is presented for detecting fraud, and its flowchart is illustrated in figure 2. The *FDiBC* method consists of the following eight steps: 1) Pre-processing, 2) Generating feature vectors based on sliding time and scores window, 3) Generating training and test datasets, 4) Separating the fraudster's data from common customer's data, 5) Clustering, 6) Learning several binary classifiers of OCSVM, 7) Learning multi-class classifiers for SVM, KNN, C4.5, and Naïve Bayes methods, and 8) Evaluating the learned classifiers. In the first step, some pre-processes including substitution of missed data, and noise and outliers deletion are applied on instances of scores obtained by customers from the bank club system. In the second step, according to the scores of each customer, feature vectors are built based on sliding time and score windows. In fact, the input of data mining in the proposed *FDiBC* method is a feature vector based on scores of each customer. In the third step, the entire data instances are divided into training and test sets. For this division, we used a ten-fold cross-validation procedure [20], in which 90% of all data instances are randomly regarded as training dataset and the rest are test dataset for 10 times. In the fourth step, data instances of fraudsters in training dataset are separated from the common customers in order to identify their hidden patterns. In the fifth step, feature vector of fraudster's data is clustered. The focal point in this step is that the number of clusters in the proposed *FDiBC* method is determined automatically, so there is no need to be specified by human experts. In the fifth step, the training dataset is re-labeled, and new classifiers are identified. In the sixth step, based on new classifiers of training dataset, binary classifiers are learned with the OCSVM method. In the seventh step, on the basis of two classes of fraudsters and common customers, some classifiers with SVM, KNN, C4.5, and Naïve Bayes methods are learned. In the eighth step, the learned classifiers are evaluated by test dataset. In the following sections, each step is explained in detail.

3.1. Pre-processing

In order to conduct a desired data mining, the lost values should be replaced in the first place, outliers are identified, and inconsistencies are modified. In this step of the proposed *FDiBC* method, two major activities are performed: 1) Replacing the lost data, and 2) Removing noise and outliers. The lost values are the data that is not available to the analyst for any reason at the time of analysis. Existence of such data makes their analysis difficult to deal with. In this case, there are lost values in the data; they should be estimated properly. In *FDiBC*, to replace the lost data for each score instance, linear regression of Weka is used [16]. Existences of noise, outliers, and unwanted data always cause dire effect on the results of clustering and classification; therefore, in using the data mining method, at first, it is attempted to eliminate these instances. In the proposed *FDiBC* method, for removing the noise instances and unwanted data, the method of "RemoveMisclassified" is employed and implemented in the Weka tool [17].

3.2. Generating window-based feature vectors

Feature vector is used to display a score instance. In this paper, at the outset, based on each score given to a customer in the customers' club system, fourteen features are derived. These features are illustrated in table 1. A pivotal point is that fraud cannot be detected from one score, and it is usually detectable from some sequential scores. Therefore, in this paper, some innovative features are proposed from sequential scores, which are calculable out of the features shown in table 1.

In this paper, to detect the scores of fraudsters, two sliding windows called *SSW* (*Sliding Scores Window*) and *STW* (*Sliding Time Window*) are proposed. A sliding window is, in fact, referred to as more general features calculated according to all the scores of a member.

SSW is a window of scores having size *N*. It must be noted that the size of each sliding window is referred to the number of scores that the features of the corresponding sliding window are computed based on them. In table 2, a list of 12 overall features computable from the sequence of customer scores is provided as window-based features. As it can be implied by table 2, features of a *SSW* can be calculated on the basis of the sequence of customer scores.

STW is a window of scores having the aggregated features of customer scores given to them during the time of window, e.g. a day, a week, and a month. It should be mentioned that in the *STW* feature vectors, the window size denotes the number of scores that are obtained by the customer during the time interval of *STW.*

Moreover, it can be seen in table 2 that some features are dependent on the size of the window. These features include 1) number of purchase transactions over the size of the window, 2) number of money transfer transactions over the size of the window, 3) number of purchase mobile phone charging transactions over the size of the window, 4) number of bill payment transactions over the size of the window, 5) number of card registrations and profile completion over the size of the window, and 6) number of friends introduced over the size of the window; in which all the activities performed are divided by the size of the window for normalization.

After deriving the features from table 2, for each feature vector, a label representing whether the sliding window-based feature vector is a fraud or a common one is provided. In figure 3, an example of a *SSW* feature vector with size 4 is given. As shown in this figure, label +1 denotes *SSW* of a fraudster customer and label -1 denotes *SSW* of a common customer.

Note that in this figure, at first, from score streams of each customer, for each score, 14 features are extracted in Step 1, and according to window size, each *SSW* feature vector is calculated in Step 2.

It should be noted that the minimum size of *SSW* equals one; however, for discovering the effective size of the *SSW* feature vectors, different values should be measured to identify the optimum size. In section 4 of this paper, evaluation of the effective *SSW* size is presented.

Due to the importance of transaction history, we proposed the *STW* feature vectors. In order to employ the historical data, we proposed four *STW* feature vectors including one-day *STW,* one-week *STW,* one-month *STW,* and hybrid *STW* with different time intervals. In figures 4 and 5, samples of these *STW* feature vectors are illustrated.

As shown in figures 4 (a-c), the 12 features of table 2 were calculated for all the scores obtained by a customer during a day, a week, and a month, respectively. Hybrid *STW* feature vector (shown in Figure 5) is generated by the three current day *STW,* week *STW,* month *STW* feature vectors. In fact, the goal of the hybrid *STW* feature vector is to consider the behaviors of a customer based on the current day and the last week and month. Note that in Section 4 of this paper, evaluation of the different *STW* feature vectors is presented.

It should be mentioned that the ABA feature shown in table 2 for the *STW* feature vectors is considered the average of balance of customer account during the *STW* time interval.

3.3. Generating training and test datasets

In the proposed *FDiBC* method, the training and test datasets are determined after conducting pre-processing and generating feature vectors. The purpose of generating the training dataset is to learn classifiers, and the aim of generating the test dataset is to evaluate the learned classifies.

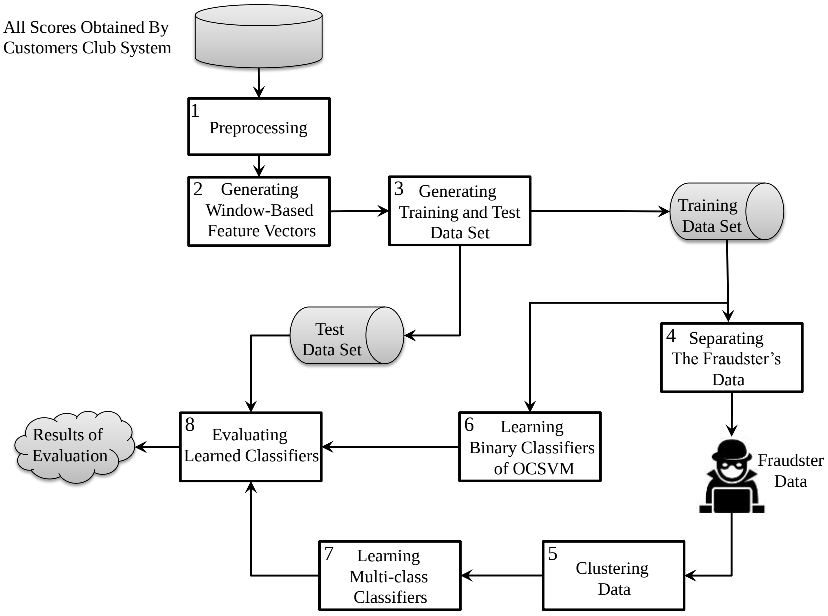

All Scores Obtained By Customers Club System
1 Preprocessing
2 Generating Window-Based Feature Vectors
3 Generating Training and Test Data Set
Training Data Set
Test Data Set
4 Separating The Fraudster's Data
Fraudster Data
5 Clustering Data
6 Learning Binary Classifiers of OCSVM
7 Learning Multi-class Classifiers
8 Evaluating Learned Classifiers
Results of Evaluation

Figure 2. Process of proposed *FDiBC* method.

Table 1. Extracted features for each score obtained by customers.

No.	Feature Name	Feature Type	ACRONYM
1	Score Type	Enumeration (including 1) Account-Based, 2) Card-Based, and 3) Club Activity-Based)	SCT
2	Amount	Numerical	AMN
3	Transaction Type	Enumeration (including 1) Purchase, 2) Money Transfer, 3) Pay Bill, and 4) Purchase Mobile Phone Charging)	TRT
4	Account Type	Enumeration (including 1) Saving Account, 2) Checking Account, 3) Money Market Account, and 4) Certificates of Deposit)	ACT
5	Account Balance	Numerical	ACB
6	Port Type	Enumeration (including 1) Net Bank, 2) Branch, 3) PinPad/PoS (Point of Sale), 4) ATM (Automated Teller Machine), 5) IVR (Interactive Voice Response), and 6) Mobile Bank)	POT
7	Customer Group Type	Enumeration (including 1) Regular, 2) Special, and 3) Honorary)	CGT
8	Acquire Account Number	Numerical	AAN
9	Date	Date	DAT
10	Time	Time	TIM
11	Introduce Friend	Binary (1 denotes a new friend introduction)	FRI
12	Register Card	Binary (1 denotes a new card registration)	RCA
13	Register	Binary (1 denotes customer registration)	REG
14	Customer Profile	Binary (1 denotes customer profile completion)	CUP

Table 2. Features proposed for a sliding window-based feature vectors.

No.	Proposed Feature Name	Feature Type	Acronym
1	Average Interval Time Between Scores	Time	ATT
2	Average Amount of Transactions	Numerical	AAT
3	Average Balance of Accounts for scores belonging to *SSW* and *STW*	Numerical	ABA
4	Difference Rate of Acquire Account Numbers	Numerical	DRA
5	Customer Group Type	Enumeration	CGT
6	Difference Rate of Ports	Numerical	DRP
7	Number of Purchase Transactions over Window Size	Numerical	NPT
8	Number Money Transfer Transactions over Window Size	Numerical	NTT
9	Number of Purchase Mobile Phone Charging Transactions over Window Size	Numerical	NCT
10	Number of Bill Payment Transactions over Window Size	Numerical	NBP
11	Number of Registration Card and Complete Profile over Window Size	Numerical	NRC
12	Number of Introduced Friends over Window Size	Numerical	NIF

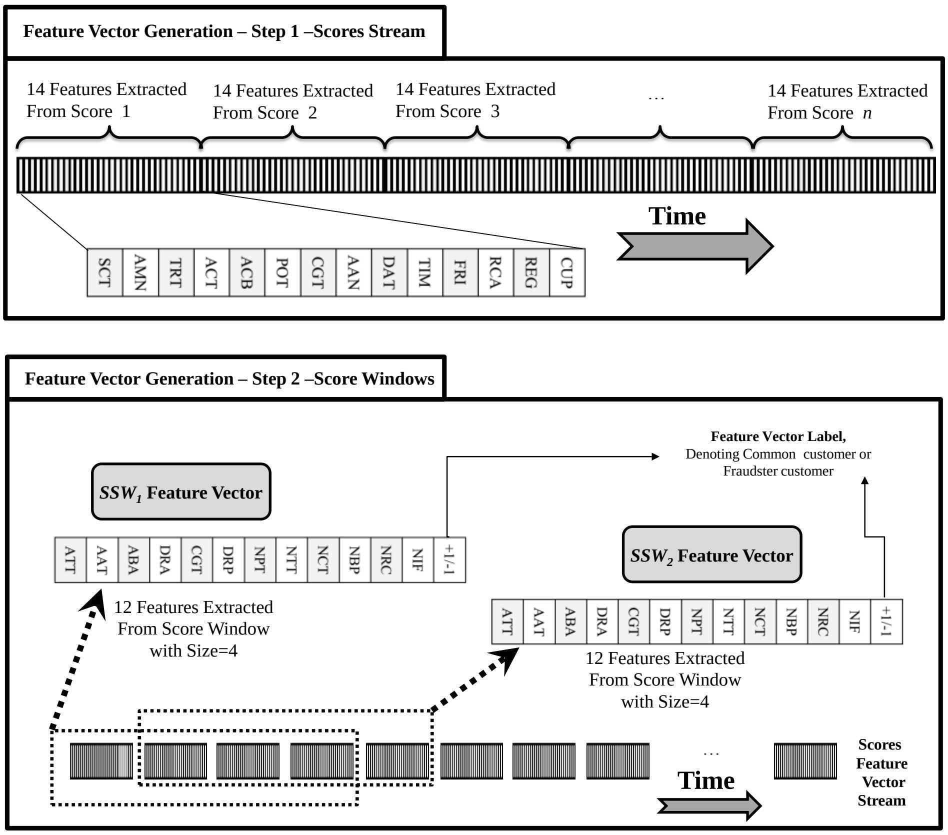

Feature Vector Generation – Step 1 –Scores Stream
14 Features Extracted From Score 1
14 Features Extracted From Score 2
14 Features Extracted From Score 3
…
14 Features Extracted From Score n
Time
SCT
AMN
TRT
ACT
ACB
POT
CGT
AAN
DAT
TIM
FRI
RCA
REG
CUP
Feature Vector Generation – Step 2 –Score Windows
Feature Vector Label, Denoting Common customer or Fraudster customer
SSW1 Feature Vector
ATT
AAT
ABA
DRA
CGT
DRP
NPT
NTT
NCT
NBP
NRC
NIF
+1/-1
12 Features Extracted From Score Window with Size=4
SSW2 Feature Vector
12 Features Extracted From Score Window with Size=4
…
Time
Scores Feature Vector Stream

Figure 3. A Sample of SSW feature vectors with size equal to 4.

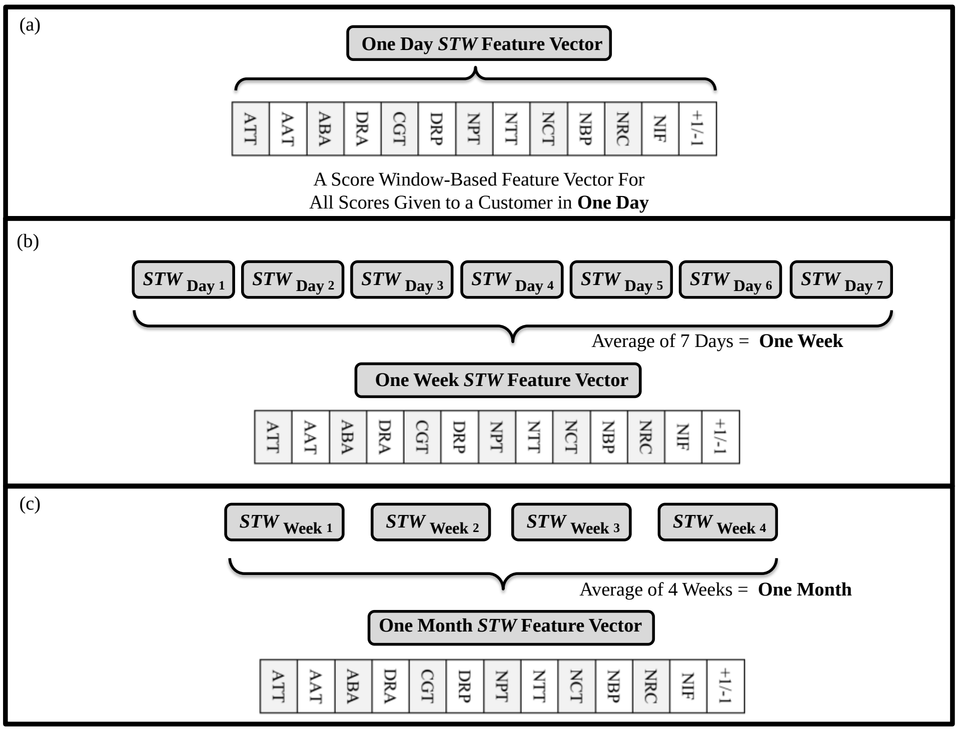

(a)
One Day STW Feature Vector
ATT
AAT
ABA
DRA
CGT
DRP
NPT
NTT
NCT
NBP
NRC
NIF
+1/-1
A Score Window-Based Feature Vector For All Scores Given to a Customer in One Day
(b)
STW Day 1
STW Day 2
STW Day 3
STW Day 4
STW Day 5
STW Day 6
STW Day 7
Average of 7 Days = One Week
One Week STW Feature Vector
(c)
STW Week 1
STW Week 2
STW Week 3
STW Week 4
Average of 4 Weeks = One Month
One Month STW Feature Vector

Figure 4. Samples of one-day STW (a), one-week STW (b), and one-month STW (c) feature vectors.

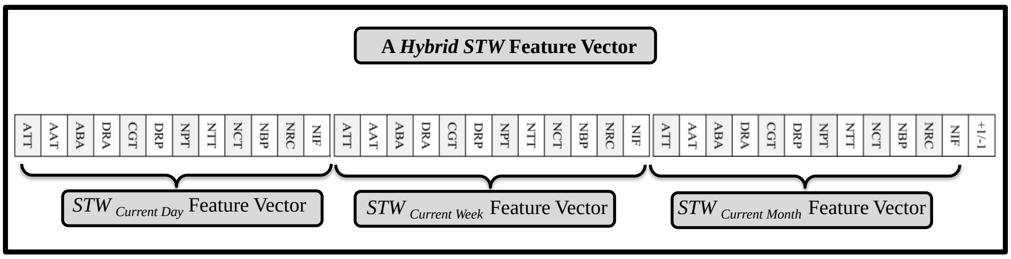

A Hybrid STW Feature Vector
ATT
AAT
ABA
DRA
CGT
DRP
NPT
NTT
NCT
NBP
NRC
NIF
ATT
AAT
ABA
DRA
CGT
DRP
NPT
NTT
NCT
NBP
NRC
NIF
ATT
AAT
ABA
DRA
CGT
DRP
NPT
NTT
NCT
NBP
NRC
NIF
+1/-1
STW Current Day Feature Vector
STW Current Week Feature Vector
STW Current Month Feature Vector

Figure 5. A Sample of a hybrid SSW feature vector.

3.4. Separating fraudster's data from common customer's data

As it can be seen in figure 2, after generating the training dataset, in order to detect the hidden patterns from fraudster's data, the data instance of fraudsters were separated from the common customer's data and sent to the clustering step (the fifth step of *FDiBC* shown in Figure 2).

3.5. Clustering

As it can be seen in figure 2, the fifth step of the proposed *FDiBC* method is to cluster the fraudster's data. The aim of this step is to re-label the labels of the fraudster's data. The reason is to find out the hidden patterns in the fraudster's data instances and to modify their labels. In order to calculate the similarity between a pair of feature vectors such as *SSW* or *STW*, we used the *Euclidean* similarity, as a popular similarity measure, which is defined in (1).

$$SIM_{Euclidean(i,j)} = \frac{1}{Dist_{Euclidean}(i,j)}$$
$$Dist_{Euclidean}(i,j) = \sqrt{\sum_{k=1}^{|FeatureVectorSize|} |x_k - y_k|^2} \qquad (1)$$

where, x_k and y_k represent the values of the i and j instances for the k^{th} feature, respectively. It is valuable to point out that to compute the similarity between a pair of feature vectors according to the *Euclidean* distance, after calculating the *Euclidean* distance between them, the inverse value of the *Euclidean* distance is considered as the similarity between them. The reason for this is that the *Euclidean* distance has an opposite relation to similarity.

In this step, the fraudster's instances are placed in several clusters, and for each fraud cluster, a separated pattern is provided, and then in the next step, for each pattern, one classifier is learned with the OCSVM classification method that is the indicator of that pattern in detecting fraud. In fact, through the clustering process, several patterns representing a group of instances can be identified, and then, in the classification step (the sixth step in the proposed *FDiBC* method shown in Figure 2), a classifier is learned for each cluster, and the fraud detection is performed through voting among these classifiers.

So far, several methods have been offered for data clustering such as *K-means, K-mediods*, which have been studied in [18]. Among the introduced clustering methods, the evolutionary clustering methods show higher precisions than the others. For this reason, these kinds of methods are used in the proposed *FDiBC* method. Meanwhile, in *FDiBC,* we require an algorithm for clustering, in which there is no need to determine *k* (number of clusters) before performing clustering. Therefore, the clustering methods should be used to determine the optimal value of *k* automatically. One of the new methods of clustering based on the PSO (*Particle Swarm Optimization*) algorithm is called the CPSOII algorithm [19]. The cause of using the CPSOII algorithm in the proposed *FDiBC* method for clustering fraudster's data is the high precision of this algorithm, and evaluation presented in [19] revealed that the CPSOII algorithm outperforms the classical clustering methods such as *K-means* and the evolutionary clustering methods such as GA. Another reason for choosing the CPSOII algorithm in *FDiBC* is that this algorithm is able to find the optimal number of clusters automatically. The important point to be noted in this section is that clustering is only applied to the fraudster's data belonging to the training dataset and the common customer's data is not involved in the process of clustering.

3.6. Classification

In order to detect fraud from the training dataset, some classifiers are learned. A classifier is a model through which the label of new data (test dataset) can be predicted. In the proposed *FDiBC* method, the two binary and multi-class classification methods are used. The output of the fraudster's data clustering was delivered to the binary OCSVM classification method (sixth step in the proposed *FDiBC* method, according to Figure 2). The OCSVM classification does not need any data with -1 label for learning a classifier but it can learn the classifier using data with +1 label. In the multi-class classification methods (seventh step in *FDiBC*, according to Figure 2), all instances of the training dataset are used; and the instances with +1 labels are used for fraudster customers, and -1 for common customers. After learning these classifiers, in the eighth step of the proposed *FDiBC* method, according to figure 2, each instance of the test dataset is given to all classifiers and their opinions are asked, and then it is compared with its real label, and finally, the performance of each classifier is evaluated. The point to be noted is that so far, several classification methods have been provided; however, it cannot be concluded that a particular method is the best for classification. In order to find out the best classification method in any area, different methods should be evaluated. Therefore, in the current study, the SVM, KNN, C4.5, Naïve Bayes, and OCSVM classification methods were employed, and one of the innovations of this paper is to identify the best classification method for the fraud detection.

4. FDiBC evaluation

In order to evaluate the proposed *FDiBC* method, 20388 instance scores during three years registered in the customers' club system of Saman Bank as a case study. These scores belonged to 2292 customers of the customer's club, 112 of which have committed fraud with 1933 score instances. Therefore, the scores of fraudster customers have been considered as +1 label, and -1 label have been used for common customers.

Figure 6 shows the score frequencies of the club system case study including the number of customers with *i* scores and the total number of instance scores. As shown in figure 6, the number of scores for each customer in this case study is in the range of [1, 31].

In the first step of the proposed *FDiBC* method, some filters introduced in the pre-processing section have been applied to the collected data of the case study, and the number of instances of this data have been declined from 20388 to 19823 scores. This process in the first step of *FDiBC* has considered 565 instances as noise or outlier instances, and removed.

After removing the noise and outlier instances, in the second step of the *FDiBC*, five proposed vectors including *SSW* with default window size equaled to 4, one-day *STW*, one-week *STW*, one-month *STW*, and hybrid *STW* are generated.

In the third step of the proposed *FDiBC* method, we used the ten-fold cross-validation procedure [20] to evaluate the performance of each classifier. In order to evaluate the clustering step of the proposed *FDiBC* method, the CPSOII algorithm with parameters of "number of particles" equaled 100, and the "maximum number of iterations" equaled 2500 have been applied on the fraudster instances belonging to the training dataset. Other parameters of the CPSOII algorithm have been supposed according to [19]. As for applying the CPSOII algorithm in the proposed *FDiBC* method, the important point is the function to be regarded from the provided functions of clustering fitness. So far, three fitness functions including 1) The sum of the squared errors (SSE), 2) Variance Rate Criterion (VRC), and 3) DBI Criterion have been used in the literature of clustering methods. Among these three methods, using the DBI criterion is suggested for automatically detecting the number of proper clusters. That is why in this paper, the CPSOII algorithm uses the DBI criterion as its fitness function. As mentioned earlier, five feature vectors including 1) *SSW* with the default window size equal to 4, 2) one-day *STW*, 3) one-week *STW*, 4) one-month *STW*, and 5) hybrid *STW* are used. Therefore, the application of the CPSOII algorithm over these five feature vectors leads to identify 5, 8, 7, 3, and 8 distinct clusters for *SSW*, one-day *STW*, one-week *STW*, one-month *STW*, and hybrid *STW*, respectively. It should be noted that the CPSOII algorithm has automatically achieved these results and converged into them; besides, the number of minimum clusters and the number of maximum clusters have been used, respectively, equal to 2 and the total number of instances. The results obtained by the CPSOII algorithm from the clustering of the fraudster instances equal to 0.58, 0.66, 0.47, 0.51, and 0.59 based on the DBI criterion for *SSW*, one-day *STW*, one-week *STW*, one-month *STW*, and one hybrid *STW*, respectively.

In the CPSOII algorithm, the *K-means* algorithm can be used in producing the primary population of particles as guided initialization. In figure 7, the convergence of the CPSOII algorithm for clustering of the fraudsters instances with a week *STW* feature vector have been illustrated for both modes including *guided & unguided* initialization (with *K-means* algorithm) and unguided initialization (without *K-means* algorithm).

In figure 7, the CPSOII algorithm has been implemented with *K-means* algorithm and without it, and the results, as it was expected, reveal that using *K-means* algorithm in the CPSOII algorithm leads to increase in the velocity of convergence; however, it has no impact on its precision.

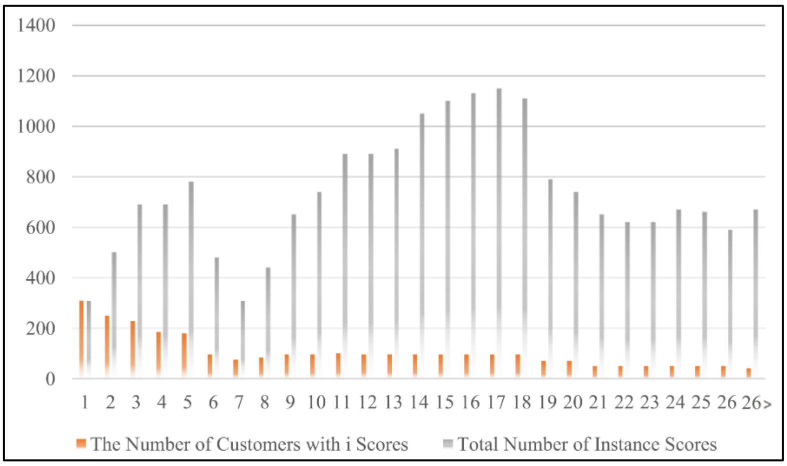


Figure 6. Score frequencies of Saman Bank club system.

Therefore, the CPSOII algorithm can identify the best or approximated to the best clusters without using the *K-means* algorithm. It should be noted that the average time of performing the CPSOII algorithm by a computer with Ci7 and main memory of 4 gigabyte is 6 minutes and 29 seconds, which can be regarded as a desirable time for pre-processing.

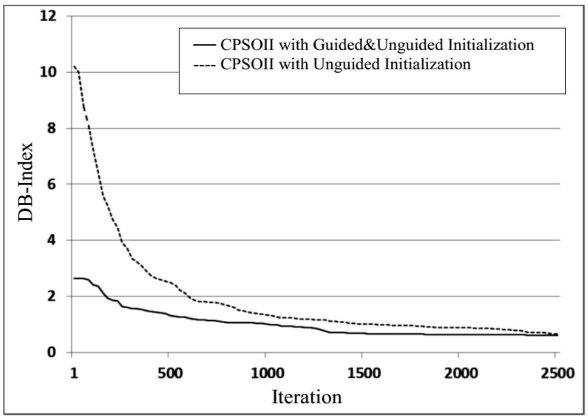


Figure 7. Comparing convergence of CPSOII with Guided & Unguided initialization toward qualified solutions (using DBI criterion) for a week STW feature vector.

In the sixth step of the proposed *FDiBC* method, a classifier is learned with the OCSVM method for each identified cluster in the fifth step. After learning, these classifiers (including 5, 8, 7, 3, and 8 distinct clusters for *SSW*, one-day *STW*, one-week *STW*, one-month *STW*, and hybrid *STW*, respectively) would be evaluated by the test dataset (including 5947 instances), which, in turn, were selected in the third step of *FDiBC* (according to Figure 2). Moreover, in the seventh step of *FDiBC*, for the training dataset, four classifiers are learned using four multi-class classifiers methods including the SVM, KNN, C4.5, and Naïve Bayes methods for five feature vectors proposed by *FDiBC* in Section 2. After learning these four classifiers for each feature vector, the test dataset (including 5947 instances), chosen in the third step of the proposed *FDiBC* method (according to Figure 2) are evaluated. The precision, recall, and accuracy criteria [20] have been employed to examine the performance of the proposed *FDiBC* method. These criteria are calculated using the results of classifiers on the test dataset. These three criteria have been defined in (2)-(4).

$$\textbf{Precision} = \frac{TP}{TP+FP} \tag{2}$$

$$\textbf{Recall} = \frac{TP}{TP+FN} \tag{3}$$

$$\textbf{Accuracy} = \frac{TP+TN}{TP+FP+FN+TN} \tag{4}$$

In these equations, *TP* (True Positive) means the number of fraudster instances, which is properly predicated as the fraudster label by classifiers. *FP* (False Positive) indicated the number of fraudster instances; however, they are predicated as the common customer label erroneously. *FN* (False Negative) indicated that the number of common customer instances is wrongly predicated as fraudsters. *TN* (True Negative) includes the number of common customer instances, which is predicated correctly as the common customer label. The results of evaluation of the learned classifiers are shown in figure 8 with five classification methods and five different feature vectors. As illustrated in this figure, among the five classification methods, the OCSVM classification method with the CPSOII clustering algorithm obtains the best results based on the precision, recall, and accuracy criteria, i.e. 0.79, 0.77, and 0.78, respectively. These results revealed that the hybrid of the clustering method with the classification method was able to learn classifiers efficiently according to the precision, recall, and accuracy criteria. From this evaluation, it can be concluded that using the clustering method for detecting the hidden patterns in the fraudster's data, the degree of classification method detection can be increased. Regarding the results achieved, shown in figure 8, the one-week *STW* feature vector outperforms the other four feature vectors. After all, the hybrid *STW* feature vector has better results in comparison with the other three feature vectors. These results reveal that considering the data of only current day, the results of fraud detection are not as good as the data of current week. In addition, the results of the aggregation of scores of one month are worse than the aggregation of scores of the current week. However, the results of hybrid *STW* feature vectors are better than one-day and one-month *STW* feature vectors. From this evaluation, it can be concluded that using the data of the current week of customer scores for detecting fraudster customer outperforms other four proposed feature vectors.

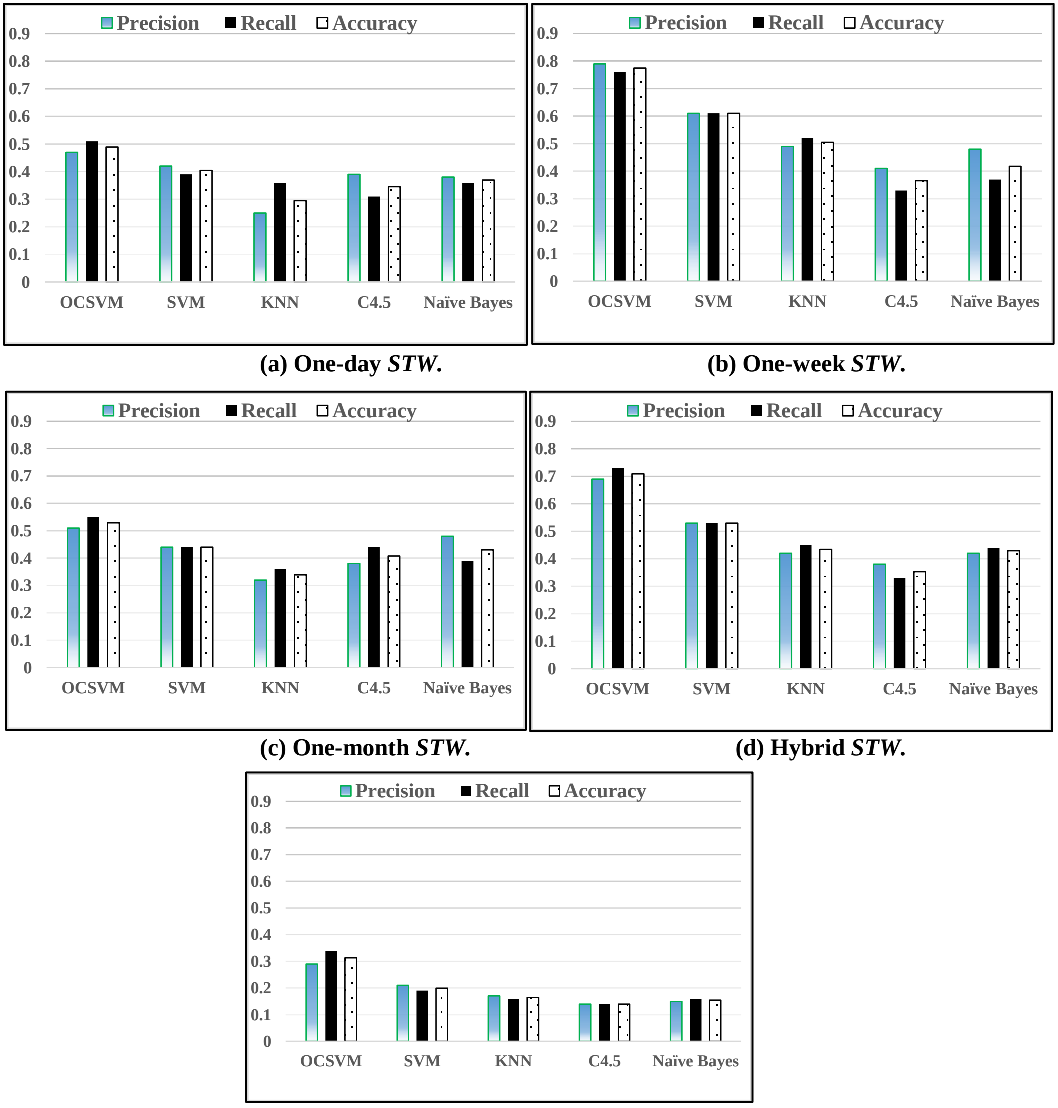
As it can be noted in Section 3.2, the size of sliding scores window (*SSW*) has an effective role on the efficiency of this feature vector; therefore, for the size of windows with 1 to 30 lengths, the values of accuracy criterion for the five classification methods are illustrated in figure 9. In this figure, only the accuracy criterion is taken into consideration, and as we can see, when the size of window is equal to 11, all the five classification methods provide the most efficiency and obtain the highest value of accuracy criterion.

Note that financial fraud in bank club system is an illegal action in which a customer achieves a score without deserving it. For instance, if we assumed that scores of remaining balance of customer account are calculated at 12:00 a.m, a fraudster can deposit 1,000,000 Rial to his account at 11:00 p.m., and after his scores are calculated, he withdraws all of the remaining balance of his account at 1:00 a.m. In fact, he can repeat these actions all midnights and achieve undeserved scores.

As mentioned earlier, using CPSOII, clustering algorithm in the proposed *FDiBC* method leads to increase in accuracy. Therefore, we investigate manually the clusters of fraudsters obtained by the CPSOII algorithm. Following on from what is said earlier, for six clusters detected by CPSOII in the fifth step of *FDiBC*, six patterns of fraud are mentioned:

1. *Obtaining scores through inviting friends (Cluster 1):*
   Regarding the fact that customers can get scores through inviting friends, the evaluation results in this section reveal that it is possible for customers of a bank club system to obtain scores via sending several invitation letters to fake e-mails. After detecting this fraud by the *FDiBC* method, in order to avoid this from happening, the inviter customer should obtain scores, whenever the invited member registers his/hers first bank card in the bank club system.
2. *Obtaining scores based on bank transactions (Cluster 2):*
   Customers can achieve scores on the basis of the amount of the transactions of the registered cards in the bank club system. According to the evaluation results in this section, some customers can obtain scores on the basis of different very low amount of transactions. After detecting this fraud by the *FDiBC* method, in order to avoid this from happening, giving scores on transaction basis should prevent by applying limitation on the amount of transaction at transaction registration time and at the time of registering the pattern of giving score on the transaction basis.
3. *Obtaining scores based on purchasing through POS (Cluster 3):*
   Based upon purchasing through POS, customers can obtain scores. According to investigations, customers having POS system can perform several transactions as purchase regulating their account on POS. These transactions are transferring money from his/her account to someone else's account. For this reason, after detecting this type of fraud by the *FDiBC* method, it is required to have a bank club alarm system, in which the information about how customers obtain score is examined, and if there is any fraud possibility, it will inform.
4. *Obtaining scores based on remaining balance of customer account (Cluster 4):*
   Customers can obtain scores on the basis of the remaining amount in their accounts. According to investigations, customers can obtain scores withdrawing from their accounts during day and completing their balance at the end of day (the system calculates customer scores during night and once a day). In order to prevent this fraud, after detecting this type of fraud via the *FDiBC* method, scores will be calculated on the basis of the minimum balance of every day or using data of more than one day to fraud detection like one-week *STW* feature vector.
5. *Obtaining scores based on registering bank cards (Cluster 5):*

Customers can obtain scores on the basis of registering their cards. Based on the investigations, if customers can register gift card, they can obtain scores without creating any added value to the bank.



(a) One-day *STW*.

(b) One-week *STW*.

(c) One-month *STW*.

(d) Hybrid *STW*.

(e)*SSW* with its window size equaled to 4

Figure 8. Results of evaluating different classification methods with different feature vectors.

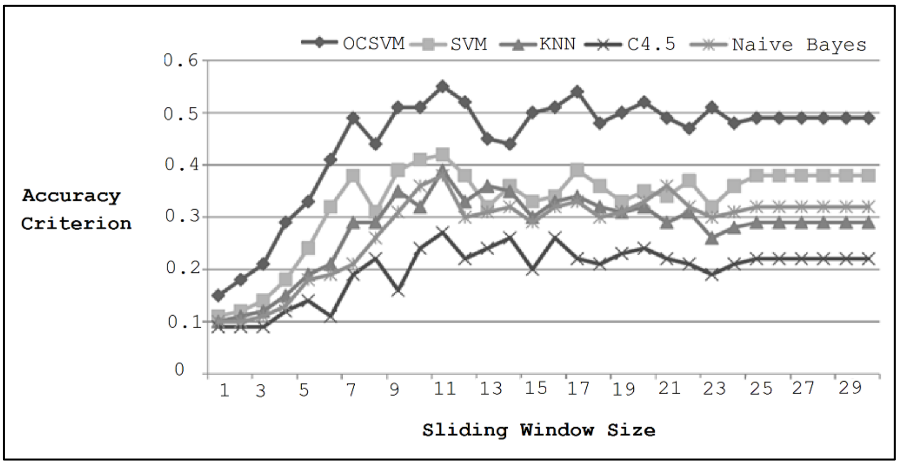


Figure 9. Impact of changing sliding window size on classifiers according to accuracy criterion.

For this purpose, after detecting this type of fraud through the *FDiBC* method, registration of this kind of cards have been prohibited in the customers' club system. In addition, each customer member can register at most 10 cards (according to the investigations) in his/her user account in order to prevent fraud based on the card registration except his/her own cards.

6. *Obtaining scores based on transactions of bill payment (Cluster 6):*
   Customers can obtain scores through bank POS by paying bills. According to the investigations on the processes of bill payment of some customers, the following frauds are detected:
   A- Creating different bills based on the formula of bill can identify with recognize from a major bill.
   B- Paying the bills of people who are not members of the bank club. Oftentimes, corporations offering bank services adopt this technique to commit fraud.

After detecting this kind of fraud through the proposed *FDiBC* method, a number of measures have been done to restrict every individual in paying bills.

As mentioned in the Introduction Section, there is no research work in which authors considered the fraud happened in bank club systems. However, the similar approach to *FDiBC* is Duman et al. [13] work. In [13], for each transaction, a score is given, and based upon this score, transactions are classified into legal or illegal. Indeed, in [13], each score has a label, and the classification process is applied on each score. Therefore, to compare *FDiBC* with the idea presented in [13], we assigned a label to 14 features obtained by each customer's score, and the results of five used classification methods are illustrated in figure 10.

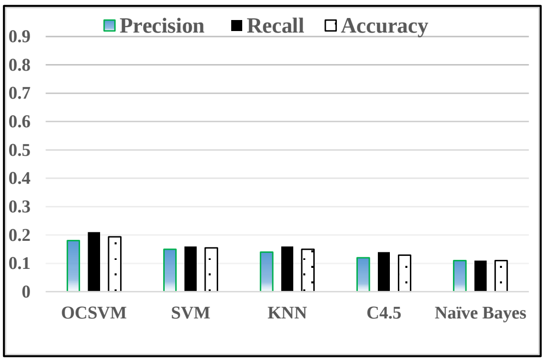


Figure 10. Results of different classification method evaluations with feature vector creating by 14 features mentioned in table 1.

As shown in this figure, as it is expected, the results of 14 features obtained by each customer's score is very poor. The reason for this case is that in bank club systems, a fraud usually happens by a sequence of scores, and one score alone cannot show a fraud.

5. Limitation

In the course of experiments during the evaluation, a number of limitations of FDiBC are apparent. First, employing an evolutionary search algorithm, i.e. the CPSOII algorithm for clustering, leads to an increase in complexity, particularly time complexity. In addition, when FDiBC calculates the STW and SWW feature vectors, the run-time of the FDiBC algorithm is greatly increased. Of course, we can assume that such systems could be optimized without concern for real-time performance because the systems could be run offline. Therefore, it seems likely that the run-time of the FDiBC algorithm is tolerable. Finally, as the clustering is a type of NP-complete problem [20], therefore, like the other existing methods, the FDiBC algorithm cannot guarantee to achieve an optimal solution. However, instead of the use of simple heuristics like K-means, FDiBC uses a powerful search-based algorithm, i.e. the CPSOII algorithm, as a crucial alternative to solve NP complete optimization problems [20]. As shown in figure 7, CPSOII outperforms the other heuristics methods like K-means according to DBI metric.

6. Conclusion

Fraud detection in the customer club system is one of the new challenges in the banking industry. In this paper, a novel sliding time and scores window-based method, called *FDiBC* (*Fraud Detection in Bank Club*), was proposed to detect fraud in bank club. In *FDiBC*, two models of feature vectors including time window-based, i.e. one-day *STW*, one-week *STW*, one-month *STW*, and hybrid *STW*, and scores window-based, i.e. *SSW*, were proposed for detecting fraud. Additionally, the dataset was divided into the training and test sets. The training dataset was learned through two approaches: 1) clustering with the CPSOII algorithm and classifying with the OCSVM binary classification method, and 2) classifying with multi-class SVM, C4.5, KNN, and Naïve Bayes classification methods. At the end, the learned classifiers were evaluated using the test dataset. The evaluation results presented in Section 4 of this paper revealed that out of the two approaches of the binary classification with clustering and multi-class classification, the binary classification with clustering provided more efficiency. Moreover, among the five proposed feature vectors, the evaluation results presented in Section 4 revealed that the one-week

STW feature vector outperformed the other four feature vectors. Meanwhile, the scores window size was more effective in the precision of the proposed *SSW* feature vector, and therefore, by changing this value and evaluating the results, the best value of the scores window size was 11.

Finally, evaluation of the proposed *FDiBC* method suggested that using the CPSOII clustering with the OCSVM classification method along with one-week *STW* feature vector detected financial fraud with an accuracy criterion equal to 0.78%. Applying *FDiBC* on the used dataset leads that six patterns of fraud are detected (see Section 4). Note that these patterns of fraud can be helpful for administration users of bank club systems to prepare solutions to deal with the problem.

For future works, we intend to use other classification methods to improve the performance of learning rate. In addition, we are going to propose a bank club alarm system like an IPS (*Intrusion Prevention System*) as a preventive measure against fraud and employ the proposed method in other banking areas [21].

7. References

[1] European Central Banking, (2014), Third Report on Card Fraud, https://www.ecb.europa.eu/pub/pdf/other/cardfraudreport201402en.pdf

[2] Han, J., Kamber, M., & Pei, J. (2011). Data mining: concepts and techniques: concepts and techniques. Elsevier.

[3] Homayounfar, E., Sepehri, M. M., Hasheminejad, S. M. H., & Ghobakhloo, M. (2014). Designing a chronological based framework for condition monitoring in heart disease patients-a data mining approach (DM-PTTD). Iranian Journal of Medical Informatics, vol. 3, no. 3, pp. 1-6.

[4] Sebastiani, F. (2002). Machine learning in automated text categorization. ACM computing surveys (CSUR), vol. 34, no. 1, pp. 1-47.

[5] Hasheminejad, S. M. H., & Jalili, S. (2013). SCI-GA: Software Component Identification using Genetic Algorithm. Journal of Object Technology, vol. 12, no. 2, pp. 1-35.

[6] Hasheminejad, S. M. H., & Jalili, S. (2014). An evolutionary approach to identify logical components. Journal of Systems and Software, vol. 96, pp. 24-50.

[7] Hasheminejad, S. M. H., & Jalili, S. (2015). CCIC: Clustering analysis classes to identify software components. Information and Software Technology, vol. 57, pp. 329-351.

[8] Hasheminejad, S. M. H., & Jalili, S. (2009). Selecting proper security patterns using text classification. In Computational Intelligence and Software Engineering, CiSE International Conference on IEEE, pp. 1-5.

[9] Hasheminejad, S. M. H., & Jalili, S. (2012). Design patterns selection: An automatic two-phase method. Journal of Systems and Software, vol. 85, no. 2, pp. 408-424.

[10] West, J., & Bhattacharya, M. (2016). Intelligent financial fraud detection: A comprehensive review. Computers & Security, vol. 57, pp. 47-66.

[11] Syeda, M., Zhang, Y. Q., & Pan, Y. (2002). Parallel granular neural networks for fast credit card fraud detection. In Fuzzy Systems, 2002. FUZZ-IEEE'02. Proceedings of the 2002 IEEE International Conference on IEEE (Vol. 1, pp. 572-577).

[12] Phua, C., Lee, V., Smith, K., & Gayler, R. (2010). A comprehensive survey of data mining-based fraud detection research. arXiv preprint arXiv:1009.6119.

[13] Duman, E., & Ozcelik, M. H. (2011). Detecting credit card fraud by genetic algorithm and scatter search. Expert Systems with Applications, vol. 38, no. 10, pp. 13057-13063.

[14] Gill, N. S., & Gupta, R. (2012). Analysis of Data Mining Techniques for Detection of Financial Statement Fraud. The IUP Journal of Systems Management, vol. 10, no. 1, pp. 7-15.

[15] Kumari, N., Kannan, S., & Muthukumaravel, A. (2014). Credit Card Fraud Detection Using Hidden Markov Model-A Survey. Middle-East Journal of Scientific Research, vol. 19, no. 6, pp. 821-825.

[16] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. ACM SIGKDD explorations newsletter, vol. 11, no. 1, pp. 10-18.

[17] WEKA, RemoveMisclassified, (2017), http://weka.sourceforge.net/doc.dev/weka/filters/unsupervised/instance/RemoveMisclassified.html.

[18] Hruschka, E. R., Campello, R. J., Freitas, A., & De Carvalho, A. C. (2009). A survey of evolutionary algorithms for clustering. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, vol. 39, no. 2, pp. 133-155.

[19] Masoud, H., Jalili, S., & Hasheminejad, S. M. H. (2013). Dynamic clustering using combinatorial particle swarm optimization. Applied intelligence, vol. 38, no. 3, pp. 289-314.

[20] Alpaydin, E. (2014). Introduction to machine learning. MIT press.

[21] Siami, M., & Hajimohammadi, Z. (2013). Credit scoring in banks and financial institutions via data mining techniques: A literature review. Journal of AI and Data Mining, vol. 1, no. 2, pp. 119-129.

نشریه هوش مصنوعی و داده کاوی

دانشگاه صنعتی شاهرود

روشی جدید برای تشخیص کلاهبرداری در باشگاه مشتریان بانک مبتنی بر پنجره‌های لغزان زمانی و امتیازی

سید محمد حسین هاشمی نژاد* و زینب سلیمی

گروه کامپیوتر دانشکده فنی و مهندسی، دانشگاه الزهرا(س)، تهران، ایران.

ارسال ۲۰۱۶/۱۱/۱۷؛ بازنگری ۲۰۱۷/۰۲/۲۷؛ پذیرش ۲۰۱۷/۰۶/۱۸

چکیده:

یکی از استراتژی‌های افزایش وفاداری مشتریان در سیستم‌های بانکی، استفاده از سامانه باشگاه مشتریان می‌باشد. در سامانه باشگاه مشتریان، هر مشتری بر اساس فعالیت‌های باشگاهی و تراکنش‌های مالی امتیازهایی کسب می‌کند که توسط آن امتیازها می‌تواند اعتباراتی از بانک بدست آورد. از طرفی دیگر با ظهور فناوری‌های جدید کلاهبرداری‌ها نیز در سامانه‌های بانکی رشد چشمگیری داشتند. بنابراین با توجه به اعتباراتی که فرد می‌تواند از باشگاه مشتریان یک بانک کسب کند، کشف کلاهبرداری در این سامانه‌ها نیز اهمیت زیادی دارد. در این مقاله روشی جدید مبتنی بر پنجره‌های لغزان زمان و امتیازی برای کشف کلاهبرداری در سامانه باشگاه مشتریان بانک ارائه شده است که در آن ابتدا از روی هر امتیاز مشتری ۱۴ ویژگی استخراج می‌گردد، سپس بر روی جریان امتیازهای مشتری، ۵ پنجره لغزان عبور داده شده تا بردارهای ویژگی از امتیازات مشتری بوجود آید. سپس دو مجموعه-داده آموزش و آزمون با برچسب مثبت برای نمونه امتیازات کلاهبرداری شده و برچسب منفی برای نمونه امتیازات عادی ایجاد می‌گردد. پس از ایجاد مجموعه آموزش در روش پیشنهادی، از دو رویکرد: ۱) خوشه‌بندی و دسته‌بندی تک کلاسه با روش OCSVM و ۲) دسته‌بندی چندکلاسه با روش‌های بیز ساده، SVM، KNN وC4.5 استفاده می‌شود. نتایج ارزیابی‌های روش پیشنهادی، نشان می‌دهد که این روش توانایی تشخیص کلاهبرداری با درستی ۷۸٪ را دارد و یک روش کاربردی محسوب می‌گردد.

کلمات کلیدی: تشخیص کلاهبرداری مالی، سامانه باشگاه مشتریان، صنعت بانکداری، پنجره لغزان.