Journal of Computer and Knowledge Engineering, Vol. 2, No.2, 2019.
DOI: 10.22067/cke.v2i2.82152

SSLBM: A New Fraud Detection Method Based on Semi- Supervised Learning

Research Article

Zahra Karimi Zandian[1] Mohammad Reza Keyvanpour[2*]

Abstract. The increment of computer technology usage and rapid development of the Internet and electronic business lead to an increase in financial transactions. With the increase of these banking activities, fraudsters also use different methods to boost their fraudulent activities. One of the ways to cope their damages is fraud detection. Although, in this field, some methods have been proposed, there are essential challenges on the way. For example, it is necessary to propose methods that detect fraud accurately and fast, simultaneously. Lack of non-fraud labeled data and little fraud labeled data for learning is another challenge in this field particularly in banking. Therefore, we propose a new fraud detection method for bank accounts called SSLBM. In this method, after preprocessing phase, a helpful learning method called SSEV is used that is based on semi-supervised learning and evolutionary algorithm. The results imply improvement of detection by using SSLBM with 68% accuracy and acceptable speed.

Keywords. Fraud, Fraud detection, Semi-supervised learning, Evolutionary algorithm, Feature extraction.

I. Introduction

Wang et al. [1] introduced fraud as a meaningful activity to obtain unauthorized financial benefits and contrasts with law, rule, or policy. In more general terms, fraud is a deliberately deceptive and misleading activity that is different from definitions of normal behavior. Fraud is said to be an abnormal behavior that perpetrators attempt to portray as normal [2].

Using the Internet for different purposes has become one of the daily activities of almost all people around the world who research, shop, use applications, and do many other things online [3].

In general, areas in which fraudsters engage in fraudulent activities include credit cards, online auctions, insurance and telecommunications, and e-business. Fraud detection aims to deal with damages of fraudulent activities. It is part of the overall control fraud, coming into play once prevention has failed to aim at stopping the abuse in progress as quickly as possible after its first occurrence [4].

An investigation of the methods proposed in this area, clarifies that the main challenge in fraud detection is to employ a method to detect frauds quickly and accurately [5], [6], inasmuch as what is important in the fraud detection area is to identify frauds and non-frauds correctly. The process of identification has to be done quickly as well to prevent next fraudulent activities.

Another challenge in fraud detection especially in banking is lack of data with non-fraud label and shortage of data with fraud label to learn [7].

Therefore, using supervised methods seems to be inappropriate. Furthermore, labeled data in the learning process eventuates accurate detection. Consequently, learning to use some unlabeled data and some labeled data is efficient. This is the main basis of semi-supervised learning [8].

In this study, a new fraud detection method called SSLBM, which helps to detect banking fraud quickly, accurately, academically, and practically, is proposed to encounter these challenges due to the inability of certain to examine these detection problems at the same time, whether theoretically or practically. According to this method, in the first phase useful features are extracted by the extraction method. In the second phase, a novel learning method called, SSEV is proposed which is based on semi-supervised clustering and evolutionary algorithm. The purpose of SSEV method is to combine these two methods in order to increase the efficiency of each of them for fraud detection accurately and quickly and using just little fraud labeled data. In this step, learning occurs based on the features obtained from the previous step and SSEV as the learning method. According to experimental results, using SSLBM creates a trade-off between the accuracy and velocity and improves them.

This article proceeds as follows: In section two, related works are discussed. Section three introduces SSLBM. Section four presents the experiments. Finally, section five proposes the concluding remarks.

II. Related Work

In this section, we provided an overview of the related work in fraud detection. Reviewing the proposed methods and their classifications in [2] fraud detection are divided into four categories based on the strategy ahead: data mining-based methods, social network analysis-based methods, formal methods, and statistics-based methods (Fig. 1). In recent years, many efforts have been initiated in the area of fraud detection, which are often based on either data mining or social network analysis or a combination of them.

Manuscript received July, 23, 2019; accepted March, 3, 2020.

[1] Z. Karimi Zandian, MSc, Data Mining Lab, Department of Computer Engineering, Faculty of Engineering, Alzahra University, Tehran, Iran. E-mail: z.karimizandian@yahoo.com.

[2*] M. R. Keyvanpour, Associate Prof., Department of Computer Engineering, Faculty of Engineering, Alzahra University, Tehran, Iran. E-mail: keyvanpour@alzahra.ac.ir

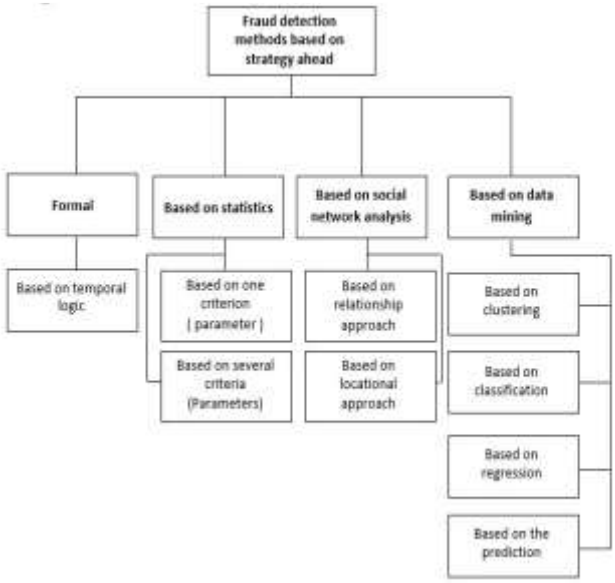

Fig. 1.Classification of fraud detection methods based on the strategy ahead [2]

A. Data mining-based methods

Panigrahi et al. [9] have developed a hybrid approach to fraud detection in credit cards. This method has combined evidence from current as well as past behaviors. The proposed fraud detection system comprises a rule-based filter, a Dempster-Shafer adder, a transaction history database and a Bayesian learner. Chang et al. [10] have suggested an early fraud detection mechanism in which an innovative two-phased modeling framework has been developed. It integrates hybrid-phased models with a successive filtering procedure to identify latent fraudsters by examining the phased features of potential fraudsters' lifecycles in online auctions.

Awad [11] has proposed a novel framework based on behavioral biometric factors. The proposed model works on both security and performance factors. This characteristic leads to increased security of the fraud detection system against hacker attacks. Jain et al. [12] have suggested a method to extract a new set of features based on analyzing the periodic behavior of the time of a transaction using the von Mises distribution for credit card fraud detection. In a study by Ram and Gray [13], the use of a variant of density estimation trees is recommended for fraud detection using distributional properties of the data, both categorical and numerical features.

Sarno et al. [14] used a method combining process mining and association rule learning for fraud detection. Baader et al. [15] proposed a new hybrid method to reduce false positive results. In this method, the red flag approach is combined with process mining. Kunda et al. [16] introduced a method obtained from the combination of BLAST and SSAHA algorithms.

FU et al. [17] presented a technique based on CNN for fraud detection in credit cards to obtain exclusive patterns of labeled data. After extracting the patterns, hidden patterns of each sample were identified by neural network technique. In a study conducted by Behera and Panigrahi [18], a two-stage neuro-fuzzy expert system was proposed for credit card fraud detection. In the first step, using a pattern-matching system transactions were processed and a score was assigned to each transaction. A fuzzy inference system computes a suspicious score by combining the values and accordingly classifies the transaction as genuine, suspicious, or fraudulent. If a transaction is labeled as suspicious, in the second step a previously trained neural system is used to investigate whether it was an actual fraudulent action or an occasional deviation by the legitimate user.

Khodabakhshi and Fartash [19] have proposed a method combining KNN and association rule mining algorithms to detect frauds of banking transactions.

Chen et al. [20] have integrated natural language processing, queen genetic algorithm and SVM to develop a novel fraud detection method to increase the accuracy of fraud detection in the narratives of annual reports. Carneiro et al. [21] have proposed a combination of manual and automatic classifications to detect frauds in credit cards area. Zoldi et al. [22] presented a three step-method for fraud detection to obtain frequency information for at least one account, converting frequency information to a frequency variable and predicting whether an activity is fraudulent in response to the frequency variable.

The researchers in [23] proposed a technique to detect fraud in large scale online auction networks as an incremental semi-supervised anomaly detection. They tried to solve low detection performance or slow convergence of fraud detection in online auction. This method combines semi-supervised anomaly detection with belief propagation to detect collusive frauds.

Taha et al. [24] used an optimized light gradient boosting machine to detect fraud in credit card. In this algorithm, a Bayesian-based hyper parameter optimization algorithm is intelligently integrated to tune the parameters of the light gradient boosting machine algorithm.

Beigi et al. [25] proposed a new fraud detection method in credit card that combine datamining and statistical methods. In this method, after identifying useful features, the resampling strategy is determined based on the design of experiments and response surface methodologies. In this paper the cost-sensitive C4.5 algorithm is used as the base learner in the Adaboost algorithm.

B. Social network analysis-based methods

Subelj et al. [26] have developed an expert system to detect and investigate groups of collaborating automobile insurance fraudsters. This system has focused on detection of groups of collaborating fraudsters and the relations between them. As a result, the networks have been used to show these relations and calculate a score as the suspicious score for each entity. Lin et al. [27] used a ranking method to evaluate how dangerous the fraudsters were. Thereby, a process based on social network analysis can provide a method to detect collusive fraud groups in online auctions.

Sylla et al. [28] have focused on the creation of new coding models based on the extensions of SQL and MapReduce and using path concept in graphs on a large scale.

Jamshidi et al. [29] have developed a data enrichment scheme that focuses on social network analysis in order to help the detection systems by providing information on hidden relations between entities.

Vlasselaer et al. [30] have extracted network level features using social network analysis and proposed a new propagation algorithm in order to measure impressibility of nodes from fraudster nodes.

Jiang et al. [31] designed a network called GCN to detect anomalous behaviors of users and malicious threat groups. In this network, the relationship between entities and features of each of them are specified. The researchers studied and analyzed this network to detect both anomalous behaviors of individuals and associated anomalous groups. This network is applicable for fraud detection.

C. Data mining and social network analysis- based methods

To detect fraudsters, Lin et al. [32] suggested an approach based on neighbor diversity. In this method, some classification techniques like J45, decision tree, neural network and SVM and social network analysis were used in order to calculate neighbor diversity.

Yu et al. [33] have proposed a hybrid method to detect fraud in online auctions. In this paper, they used social network analysis to obtain behavioral features. These features convert to fuzzy rules that can show fraud detection rules. Then, they optimized the fuzzy rules using genetic algorithm.

Van Vlasselaer et al. [34] presented a method has used from both exclusive features of transactions and their network based features for fraud detection in credit cards. This method was supervised in real time.

Lebichot et al. [35] have developed a novel fraud detection system based on network and semi-supervised learning to examine the effect of fraudulent nodes on other nodes. Chiu et al. [36] have proposed a hybrid method consisting of network criteria and classification techniques in order to detect fraud in online auctions.

Lin et al. [37] proposed a model called COSIN for fraud

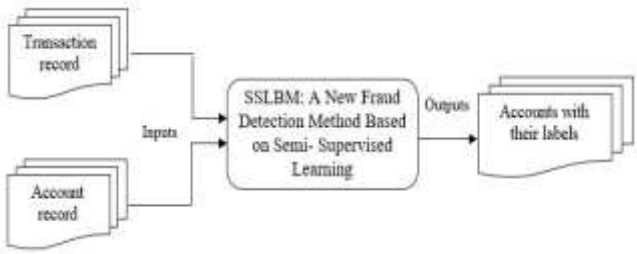

Fig. 2. An overview of the proposed method

detection that uses both sequential and interaction behaviors of users. They used a dynamic interaction network to model and study these behaviors and introduced a hierarchical Hidden Markov Model (HMM) to capture the sequential schema-dependent transitional patterns.

III. SSLBM: A New Fraud Detection Method Based on Semi-Supervised Learning

As shown in Fig. 2, bank transactions and accounts records are inputs of SSLBM method separately. Given that one of the challenges in banking fraud detection is lack of data with non-fraud label and that little data exist with fraud label to learn [7], a few accounts have fraud label (called positive label). Finally, outputs of SSLBM method are accounts with their labels – fraud label or non-fraud label. The general structure of the proposed fraud detection method is shown in Fig. 3. SSLBM method comprises two phases: pre-processing and learning.

Generally, in the first phase, pre-processing operation is done by receiving bank transactions and accounts records and finally the outputs obtained from this phase in the format of features vectors (FVs) as inputs are sent to the second phase. In the learning part, fraud accounts are distinguished from non-fraud accounts.

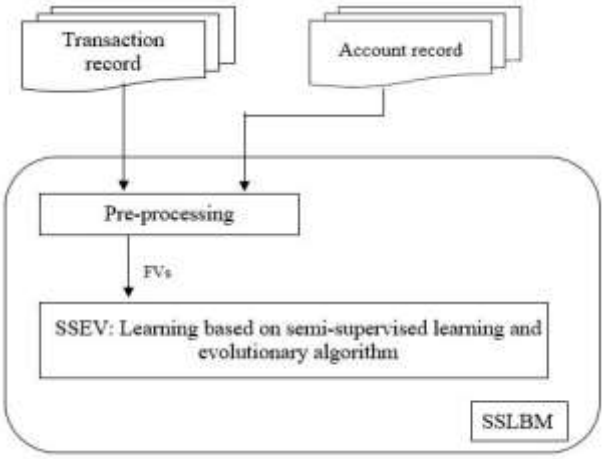

Fig. 3. The general structure of SSLBM method

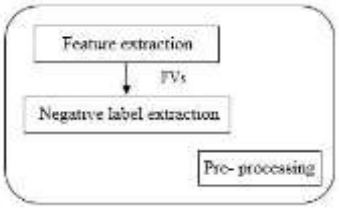

Fig. 4. Pre-processing steps

A. Pre-processing

To increase accuracy of fraud detection, it is needed to use all useful features of accounts. On the other hand, regarding existing unbalanced classes of accounts based on their labels, it is necessary to have pre-processing step before learning phase. As shown in Fig. 4, in pre-processing phase, in first step, the useful features of the accounts are extracted and then some unlabeled accounts are labeled as negative accounts (non-fraud accounts) in the second step using feature vectors (FVs) extracted from the first step.

1) Feature extraction: As mentioned in [38], features of the components can be network-based or user-based. Network-based features of each component are achieved by taking other components and relationships between them into account, whereas user-based features are obtained based on the characteristics of the component. Combination of algorithms, each of which has focused on various aspects of information hidden in the data, can help detect fraudulent accounts more accurately [38]. Therefore, we use the FEMBSNA method [39] to extract features in the first phase of the pre-processing step, which employs both feature types to detect fraudulent accounts.
2) Negative label extraction: Fraud detection is a binary

problem in which there are two sets of examples. One includes positive examples and the other includes negative examples. As mentioned before, there are very few fraudulent (positive) instances available since labeling non-fraud (negative) instances is costly and there may not be any negative instances [7], [40]. Therefore, semi-supervised learning could be very helpful. Nevertheless, lack of negative data is a challenge for traditional semi-supervised learning. There are many different methods to extract negative instances which can be divided into two approaches: the methods using only positive samples for learning and the methods using both positive and unlabeled data. The advantages of the second type of these methods include higher accuracy, better detection rate, using the reduced unlabeled data for final learning, and higher performance [7], [40]. Therefore, in this paper, we use the method developed in [7], based on KNN method, to extract negative labeled data.

According to this method, the distances of unlabeled samples from k-nearest neighbors of each positive sample are computed. The resulting values can be used to sort the classified examples, where closer unlabeled instances take positions ahead of the ones that are further away [41]. Consequently, the extracted samples include ones with more similarity to positive samples appear at the end of the list and N-components that are less similar to positive samples ahead of the list. The selected samples are used in the learning phase as negative samples.

B. SSEV: Learning based on semi-supervised learning and evolutionary algorithm

In this paper, in the learning phase, a new method called SSEV is proposed. This method that is based on semi-supervised learning and evolutionary algorithm, receives features vectors obtained from the first phase as its inputs and finally the method detects fraud and non-fraud accounts. Semi-supervised learning uses many unlabeled data and a few labeled data to learn [42] [43]. Generally, many methods based on super-vised and unsupervised learning are employed to detect fraud. However, as mentioned before, one of the basic challenges in fraud detection is the lack of sufficient labeled data in the real world due to the time-consuming, costly and difficult process of labeling training data [44], [45]. Furthermore, unsupervised learning usually suffers from high false alarm rate and low detection rate without labelled information [46]. In this paper, we use semi-supervised clustering.

CHARACTERISTICS OF THE DATASET USED

Characteristic	Quantity
Number of accounts	387
Number of transactions	2070
Number of features of transactions	5
Number of features of accounts	3

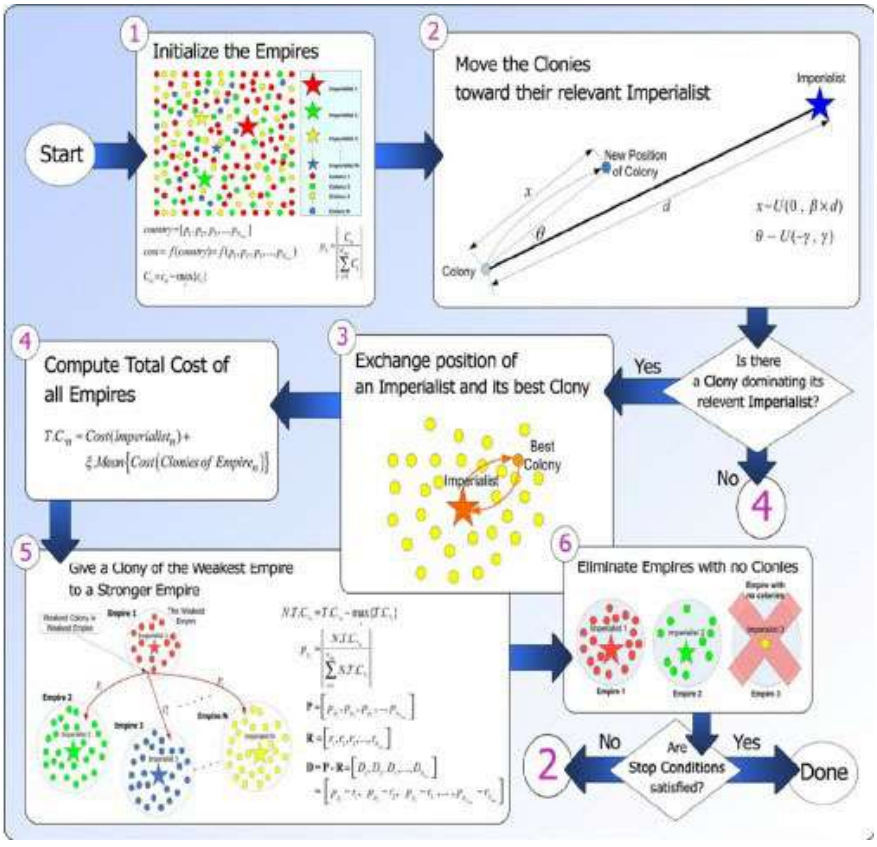

Fig. 5. The operation of the Imperialist Competitive Algorithm [49]

```
Algorithm: SSEV
Input: D, negativeData, positiveData, unlabeledData, numCountries,   numImperialist,
numDecades
Output: bestSolution
For i=1: size of D
    Select two data from D randomly and if these data are in the same class labeled as Must-link,
    otherwise  Cannot-link
End
For j=1: numCountries  //Generate initial population
    C(j)=Select first feature from positiveData and unlabeledData randomly and Select second
    feature from
    negativeData and unlabeledData randomly
    Cost(j)=Calculate cost(j) by objective function
End
Empire=Create Initial Empires
For t=1: numDecades
    Empires= Calculate PCK-means for all of counties in all of empires
   //attachment of countries toward an Imperialist with best position for them
    Empires=Assimilate Countries in each empire  //Movement of Colonies Toward Imperialists
    (Assimilation Policy) in different directions
    Empires= Revolve Countries in each empire //A Sudden Change in the Socio-Political
    Characteristics
    Empires= Recalculate Costs of countries in each empire //New Cost Evaluation
    Empires= Select best countries with lowest cost in each empire as emperor
                //Power Possession, Empire Possession
    Empires=calculate Total Cost for Empires
    Empires=Unite Similar Empires or weakest Empire with strongest
    Empires=all imperialists compete to take possession of colonies of each other// Imperialistic
    Competition
End
For t=1: number of unlabeledData
    [unlabeledData(t),label]=train and calculate labels of unlabeledData as bestSolution
End
```

Fig. 6. Pseudo code of SSEV algorithm

One of the most well-known and simple unsupervised algorithms is K-means and one of the semi-supervised algorithms that improves it, is PCK-means. This method merely ensures the achievement of local responses [47]. To solve this problem, we propose a combination of PCK-means and imperialist competitive algorithm as one of the evolutionary algorithms. The imperialist competitive algorithm has higher search and detection power, obtains better solutions and its speed of convergence to optimal solution is more than that of other evolutionary optimization algorithms [48].

Fig. 5 shows the routine operation of the imperialist competitive algorithm. As mentioned before, in our proposed learning algorithm called SSEV, a combination of PCK-means and imperialist competitive algorithms is used. The presentation of this new algorithm is another innovation of this paper to achieve more accurate and speed method.

The imperialist competitive algorithm first runs on random initial population. In the SSEV algorithm, in each decade before the movement of colonies toward imperialists, the PCK-means algorithm is applied on all colonies in order to ameliorate the performance of the ICA algorithm. This change causes emperors and colonies to be placed in better and more suitable positions, thereby increasing the speed of achieving an optimal solution and accuracy. Finally, running ICA algorithm proceeds until an optimal solution is found. In the ICA algorithm, a mutation step is added. The pseudo code of SSEV algorithm is presented in Fig. 6.

IV. Experiments

A. Dataset

In the absence of public data sources in the financial domain, especially transactional datasets with information about social relations, we used the financial data of PKDD'99 [50]. This dataset has been used to evaluate many methods in different fields [51-54]. Due to the availability of financial transaction data, demographic information, and validity of this dataset, the dataset has been used here to test our proposed method. We used transactions table to form a social network and accounts table to extract simple data. We have also made some changes to transactions table like eliminating transactions that were not transactions for transferring money. As shown in Table 1, our dataset consists of about 387 accounts selected from the accounts table and 2070 transactions from the transactions table. Each transaction has five features: trans-id, source account-id, destination account-id, amount and date. Each account also has three features: account-id, district-id, and date.

B. Evaluation criteria

To evaluate the performance of the proposed fraud detection method, popular criteria are used: True Negative (TN) rate, False Positive (FP) rate, False Negative (FN) rate, precision, recall (also called True Positive (TP) rate), F1score and accuracy.

- TNrate: as Eq. (1) shows, it is the proportion of non-frauds (negatives) that are correctly identified as such.

$$TNrate = \frac{TN}{TN+FP} \quad (1)$$

- FPrate: as stated in Eq. (2), it is the proportion of non-frauds (negatives) that are wrongly identified as frauds (positives).

$$FPrate = \frac{FP}{FP+TN} \quad (2)$$

- FNrate: the proportion of frauds (positives) that are wrongly identified as non-frauds (negatives) (Eq. (3)).

$$FNrate = \frac{FN}{FN+FP} \quad (3)$$

- Precision: as shown in Eq. (4), it is the number of accounts correctly labeled as belonging to the fraud (positive) class (TP) divided by the total number of accounts labeled as belonging to the fraud (positive) class (i.e. the sum of true positives and false positives, which are items incorrectly labeled as belonging to the class).

$$precision = \frac{TP}{TP+FP} \quad (4)$$

- Recall: the number of true positives divided by the total number of accounts that actually belong to the fraud (positive) class (i.e. the sum of true positives and false negatives, which are items and were not labeled as belonging to the positive class but should have been) (Eq. 5).

$$recall = TPrate = \frac{TP}{TP+FN} \quad (5)$$

- F_1score: as stated in Eq. (6), it is the harmonic mean of precision and recall.

$$F_1score = \frac{2.precision.recall}{precision+recall} = \frac{2.TP}{2TP+FP+FN} \quad (6)$$

- Accuracy: the proportion of frauds (positives) and non-frauds (negatives) that are correctly identified as such (Eq. (7)).

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (7)$$

C. Experimental results

The purpose of proposing the method in this paper is to improve accuracy and speed of fraud detection. Therefore, we evaluate SSLBM method and report its results based on two tests: evaluation of different fraud detection methods based on the evaluation criteria expressed in previous section and evaluation of different fraud detection methods based on runtime. Furthermore, we study impact of different parts of the proposed method on the performance of fraud detection. Therefore, in each test we compare SSLBM method with four other methods: our method with the feature extraction method proposed in [55], which has been concisely explained in [29], our method without using ICA algorithm, our method without using PCK-means method and our method without the feature extraction phase.

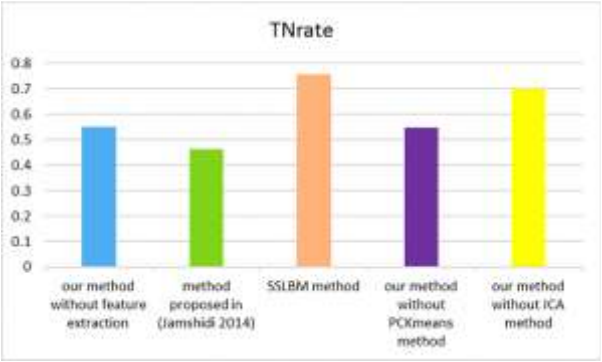

Fig. 7. Comparison between SSLBM method and other methods based on TNrate

Our method is compared to what is proposed in the study by Jamshidi [55], because this is the only method that is similar to our method in banking. This method employs features obtained from a social network created between accounts for fraud detection. Furthermore, in this paper, we detect fraud accounts, while in the other papers in banking field, fraud transactions are detected.

1) The first test is the evaluation of different fraud detection methods based on the evaluation criteria expressed in previous section. As shown in Figs 7. and 8., the SSLBM method has significantly improved TNrate and decreased FPrate. Using the FEMBSNA method [39] as a feature extraction method in the pre-processing phase, as well as PCK-means in SSEV have remarkably increased non-fraud detection rate. Using the SSLBM, our proposed method without the ICA shows a better TNrate. This shows that the existing PCK-means in SSEV contributes to the performance of our method based on TNrate. Using the method proposed in the study by Jamshidi [55] as a feature extraction in the pre-processing phase reduces the TNrate of our method more considerably compared to the absence of feature extraction in our method. This means that using FEMBSNA and PCK-means distinguishes non-frauds from frauds and correctly distinguished as non-fraud. As mentioned before FPrate is the complement of TNrate, and thus the FPrate of SSLBM is lower than that of others. Taking advantage of the ICA algorithm and combining it with PCK-means named SSEV has reduced FNrate and decreased recall more significantly than not using ICA algorithm at all (Figs 9 and 10). Existing restrictions on obtaining new features in FEMBSNA like the paths length that can cause loss of useful information for distinguishing fraud accounts have reduced the amount of recall in our method compared

to the method proposed by Jamshidi [55] as feature extraction and absence of feature extraction.

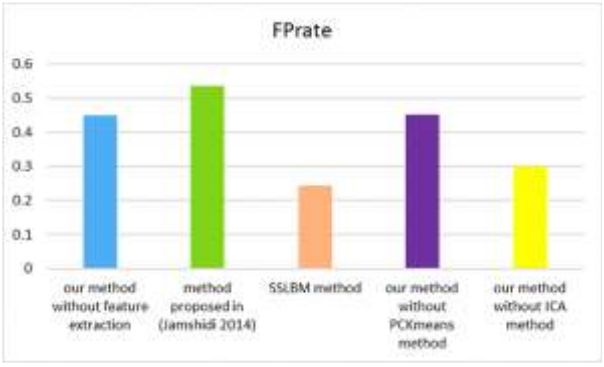

Fig. 8. Comparison between SSLBM method and other methods based on FPrate

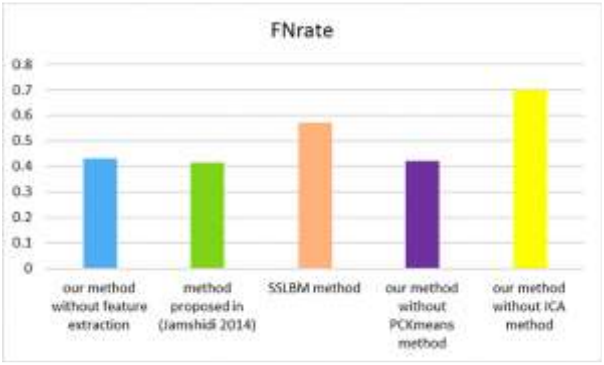

Fig. 9. Comparison between SSLBM method and other methods based on FNrate

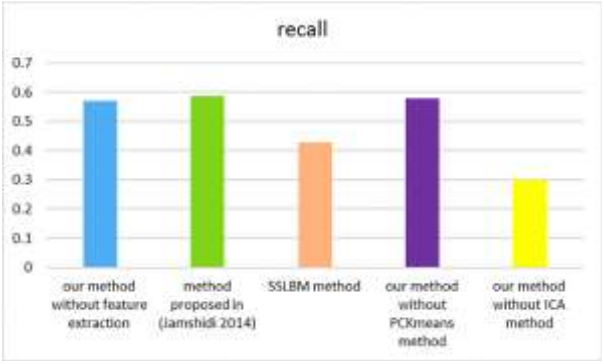

Fig. 10. Comparison between SSLBM method and other methods based on recall

The results also show that the proposed method by Jamshidi [55] used as a feature extraction method in the pre-processing phase of our method is precise in fraud detection while some non-frauds are also incorrectly detected as frauds. Thus, the FNrate of this method is low and recall is high. In this paper, an effort has been made to pay attention to correct detection of both fraud and non-fraud. It seems that lack of feature extraction phase in the detection process produced results similar to the results obtained using the method proposed by Jamshidi [55]. Given that precision is influenced by both TP and FP, SSLBM has improved the precision of detection more than while using the method proposed by Jamshidi [55]as a feature extraction method and not using any feature extraction. Using ICA algorithm and FEMBSNA has been effective in achieving this result (Fig. 11).

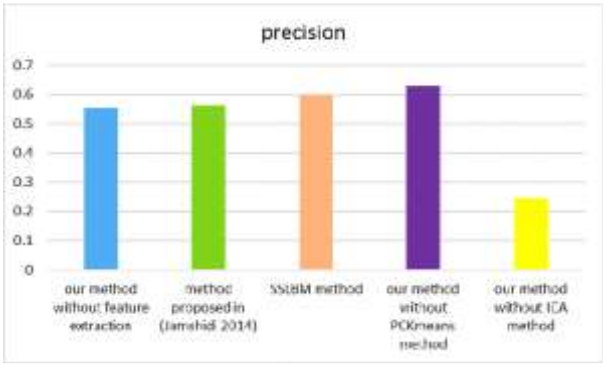

Fig. 11. Comparison between SSLBM method and other methods based on precision

Generally, to create a trade-off between precision and recall, F1score is used. It is the harmonic mean of precision and recall. Thus, F1score of all methods is deducible from their precision and recall. As shown in Fig. 12, our proposed method produces better results based on this criterion than others concerning its acceptable levels of recall and TNrate.

As shown in Fig. 13, SSLBM method is more accurate. The proposed method aimed to detect fraud and non-fraud correctly and simultaneously. To reach this purpose, both features based on network level and user level are used in the pre-processing phase using FEMBSNA method. Furthermore, we propose a new learning algorithm called SSEV, which helps to achieve this goal. However, the TNrate of our method using the method proposed by Jamshidi [55] as a feature extraction method is much lower than that what obtained in the SSLBM method and in our method without feature extraction. So its accuracy is the lowest.

2) Test2 is the evaluation of different fraud detection methods based on runtime: The main challenge in this area is to suggest a method to detect frauds accurately and quickly at the same time [5], [6]. As shown in Fig. 14, given that the implementation of ICA algorithm is time-consuming and SSEV is a combination of ICA and PCK-means, the implementation speed of the proposed method (SSLBM) is acceptable as the speed of achieving optimized solutions using features obtaining FEMBSNA method is increased by SSEV. Nevertheless, combining the method proposed by Jamshidi [55] and SSEV increases its runtime to reach optimal results. It is clear that the elimination of each part of the method reduces its runtime. Specially, eliminating ICA algorithm from the learning phase minimizes the runtime of the method. As mentioned before, the implementation of ICA algorithm is time-consuming. Finally, because of running PCK-means repeatedly, omitting it reduces the runtime of the method compared to the runtime of the method without any feature extraction phase once.

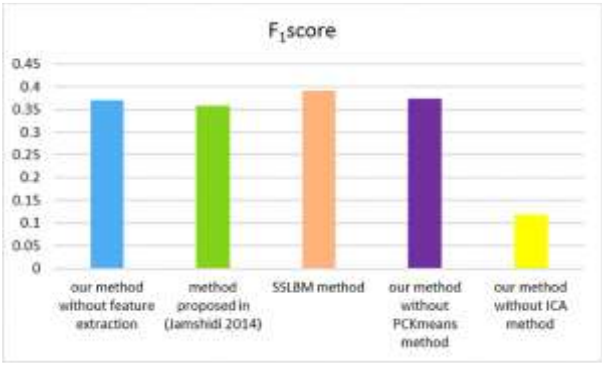

Fig. 12. Comparison between SSLBM method and other methods based on F_1score

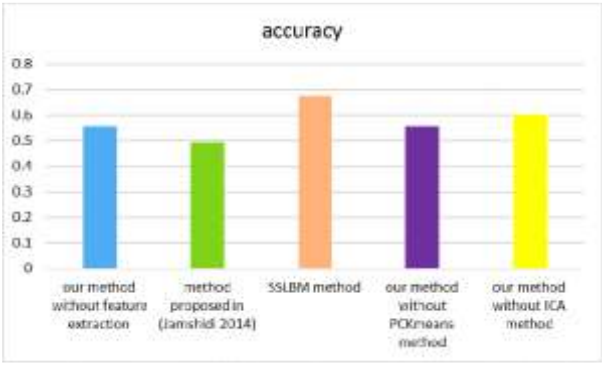

Fig. 13. Comparison between SSLBM method and other methods based on accuracy

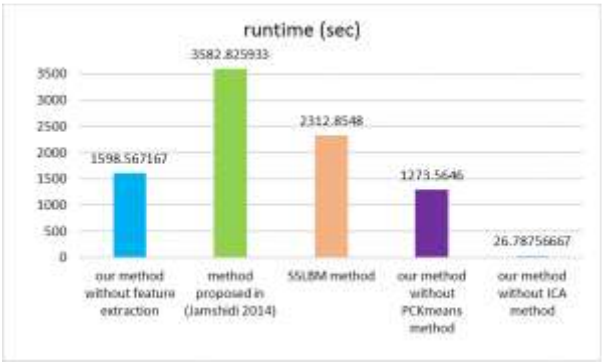

Fig. 14. Comparison between SSLBM method and other methods based on runtime

V. CONCLUSION

Fraud detection copes with the damages of fraudulent activities that have become common due to the rapid development of the Internet technology and electronic business. Lack of fast and accurate methods [5], [6], along with shortage of non-fraud labeled data are main challenges in fraud detection field particularly in banking. To address the shortcomings theoretically or practically, the SSLBM method was proposed. According to this method, in the first phase by using a feature extraction method called FEMBSNA [39], new features based on user level and network level are extracted. In the second phase, a novel learning method called SSEV is proposed, which is based on semi-supervised clustering and evolutionary algorithm. In this step, learning occurs based on the features obtained from previous step and using SSEV as the learning method. The results indicate improvement in accuracy and acceptable speed of fraud detection using our proposed method. Therefore, using this method can significantly help detect frauds especially in banking accurately and quickly. It uses merely a little fraud labeled data.

References

[1] J. H. Wang, Y. L. Liao, T. M. Tsai, and G. Hung, "Technology-based financial frauds in taiwan: issues and approaches," *in International Conference on Systems, Man and Cybernetics*, IEEE, Vol. 2, pp. 1120–1124 (2006).

[2] Z. Karimi Zandian and M. Keyvanpour, "Systematic identification and analysis of different fraud detection approaches based on the strategy ahead," International *Journal of Knowledge-based and Intelligent Engineering Systems*, Vol. 21, No. 2, pp. 123–134 (2017).

[3] M. Moradi and M. Keyvanpour, "Captcha and its alternatives: A review," *Security and Communication Networks*, Vol. 8, No. 12, pp. 2135–2156 (2015).

[4] M. Krivko, "A hybrid model for plastic card fraud detection systems," *Expert Systems with Applications*, Vol. 37, No. 8, pp. 6070–6076 (2010).

[5] S. B. E. Raj and A. A. Portia, "Analysis on credit card fraud detection methods," *in International Conference on Computer, Communication and Electrical Technology* (*ICCCET*). *IEEE*, pp. 152–156 (2011).

[6] K. Seeja and M. Zareapoor, "Fraudminer: A novel credit card fraud detection model based on frequent itemset mining," *The Scientific World Journal*, Vol. 20, (2014).

[7] A. Daneshpazhouh and A. Sami, "Semi-supervised outlier detection with only positive and unlabeled data based on fuzzy clustering," *International Journal on Artificial Intelligence Tools*, Vol. 24, No. 3 (2015).

[8] L. Xie and R. Yan, "Extracting semantics from multimedia content: challenges and solutions," *Multimedia Content Analysis*. *Springer*, pp. 1–31 (2008).

[9] S. Panigrahi, A. Kundu, S. Sural, and A. K. Majumdar, "Credit card fraud detection: A fusion approach using dempster–shafer theory and bayesian learning," *Information Fusion*, Vol. 10, No. 4, pp. 354–363 (2009).

[10] W. H. Chang and J. S. Chang, "A novel two-stage phased modeling framework for early fraud detection in online auctions," *Expert Systems with Applications*, Vol. 38, No. 9, pp. 11244–11260 (2011).

[11] A. Awad, "Collective framework for fraud detection using behavioral biometrics," in Information Security Practices. Springer, pp. 29–37 (2017).

[12] N. Jain and V. Khan, "Credit card fraud detection using recurrent attributes," People, Vol. 5, No. 2 (2018).

[13] P. Ram and A. G. Gray, "Fraud detection with density estimation trees," in KDD 2017 Workshop on Anomaly Detection in Finance, pp. 85–94 (2018).

[14] R. Sarno, R. D. Dewandono, T. Ahmad, M. F. Naufal, and F. Sinaga, "Hybrid association rule learning and process mining for fraud detection." *IAENG*

International Journal of Computer Science, Vol. 42, No. 2 (2015).

[15] G. Baader and H. Krcmar, "Reducing false positives in fraud detection: Combining the red flag approach with process mining," *International Journal of Accounting Information Systems*, Vol. 31, pp. 1–16 (2018).

[16] A. Kundu, S. Panigrahi, S. Sural, and A. K. Majumdar, "Blast-ssaha hybridization for credit card fraud detection," *IEEE transactions on dependable and Secure Computing*, Vol. 6, No. 4, pp. 309–315 (2009).

[17] K. Fu, D. Cheng, Y. Tu, and L. Zhang, "Credit card fraud detection using convolutional neural networks," *in International Conference on Neural Information Processing*. *Springer*, pp. 483–490 (2016).

[18] T. K. Behera and S. Panigrahi, "Credit card fraud detection using a neuro-fuzzy expert system," in Computational Intelligence in Data Mining. Springer, pp. 835–843 (2017).

[19] M. Khodabakhshi and M. Fartash, "Fraud detection in banking using knn (k-nearest neighbor) algorithm," in International Conf. on Research in Science and Technology (2016).

[20] Y.-J. Chen, C.-H. Wu, Y.-M. Chen, H.-Y. Li, and H.-K. Chen, "Enhancement of fraud detection for narratives in annual reports," *International Journal of Accounting Information Systems*, Vol. 26, pp. 32–45 (2017).

[21] N. Carneiro, G. Figueira, and M. Costa, "A data mining based system for credit-card fraud detection in e-tail," *Decision Support Systems*, Vol. 95, pp. 91–101 (2017).

[22] S. M. Zoldi, H. Li, and X. Xue, "Fraud detection based on efficient frequent-behavior sorted lists," Google Patents (2012).

[23] M. Dadfarnia, F. Adibnia, M. Abadi, and A. Dorri, "Incremental collusive fraud detection in large-scale online auction networks," *The Journal of Supercomputing*, pp. 1–22 (2020).

[24] A. A. Taha and S. J. Malebary, "An intelligent approach to credit card fraud detection using an optimized light gradient boosting machine," *IEEE Access*, Vol. 8, pp. 25579–25587 (2020).

[25] S. Beigi and M. Aminnaseri, "Credit card fraud detection using data mining and statistical methods," *Journal of AI and Data Mining*, Vol. 8, No. 2, pp. 149–160 (2019).

[26] L. Subelj, S. Furlan, and M. Bajec, "An expert system for detecting automobile insurance fraud using social network analysis," Expert Systems with *Applications*, Vol. 38, No. 1, pp. 1039–1052 (2011).

[27] S.-J. Lin, Y.-Y. Jheng, and C.-H. Yu, "Combining ranking concept and social network analysis to detect collusive groups in online auctions," *Expert Systems with Applications*, Vol. 39, No. 10, pp. 9079–9086 (2012).

[28] Y. Sylla, P. Morizet-Mahoudeaux, and S. Brobst, "Fraud detection on large scale social networks," in *2nd International Congress on Big Data*, pp. 413–414 (2013).

[29] S. Jamshidi and M. R. Hashemi, "An efficient data enrichment scheme for fraud detection using social network analysis," *in Sixth International Symposium on Telecommunications* (*IST*). *IEEE*, pp. 1082–1087 (2012).

[30] V. Van Vlasselaer, T. Eliassi-Rad, L. Akoglu, M. Snoeck, and B. Baesens, "Gotcha! network-based fraud detection for social security fraud," *Management Science*, Vol. 63, No. 9, pp. 3090–3110 (2016).

[31] J. Jiang, J. Chen, W. Huang, and P. Mohapatra, "Anomaly detection with graph convolutional networks for insider threat and fraud detection." *IEEE Military Communications Conference (MILCOM)*, pp. 109–114 (2019).

[32] J.-L. Lin and L. Khomnotai, "Using neighbor diversity to detect fraudsters in online auctions," *Entropy*, Vol. 16, No. 5, pp. 2629–2641 (2014).

[33] C.-H. Yu and S.-J. Lin, "Fuzzy rule optimization for online auction frauds detection based on genetic algorithm," *Electronic Commerce Research*, Vol. 13, No. 2, pp. 169–182 (2013).

[34] V. Van Vlasselaer, C. Bravo, O. Caelen, T. Eliassi-Rad, L. Akoglu, M. Snoeck, and B. Baesens, "Apate: A novel approach for automated credit card transaction fraud detection using network-based extensions," *Decision Support Systems*, Vol. 75, pp. 38–48 (2015).

[35] B. Lebichot, F. Braun, O. Caelen, and M. Saerens, "A graph-based, semi-supervised, credit card fraud detection system," in International Workshop on Complex Networks and their Applications. Springer, pp. 721–733 (2016).

[36] C. Chiu, Y. Ku, T. Lie, and Y. Chen, "Internet auction fraud detection using social network analysis and classification tree approaches," *Inter-national Journal of Electronic Commerce*, Vol. 15, No. 3, pp. 123–147 (2011).

[37] H. Lin, G. Liu, J. Wu, Y. Zuo, X. Wan, and H. Li, "Fraud detection in dynamic interaction network," IEEE *Transactions on Knowledge and Data Engineering* (2019).

[38] Z. Karimi Zandian and M. Keyvanpour, "Helpful and Efficient Framework for Classification and Analysis of various Fraud Detection Approaches from the perspective of Time and Features," *in 4th International Conference on Applied Research in Computer Engineering and Signal Processing* (2016).

[39] Z. Karimi Zandian and M. R. Keyvanpour, "Feature extraction method based on social network analysis," *Applied Artificial Intelligence*, Vol. 33, No. 8, pp. 1–20 (2019).

[40] A. Daneshpazhouh and A. Sami, "Entropy-based outlier detection using semi-supervised approach with few positive examples," *Pattern Recognition Letters*, Vol. 49, pp. 77–84 (2014).

[41] J. Hroza, J. Zizka, B. Pouliquen, C. Ignat, and R. Steinberger, "Mining relevant text documents using

ranking-based k-nn algorithms trained by only positive examples," *in Proceedings of the Fourth Czech-Slovak Conference Knowledge*, pp. 29–40 (2005).

[42] O. Chapelle, B. Scholkopf, and A. Zien, "Semi-supervised learning," IEEE *Transactions on Neural Networks*, Vol. 20, No. 3, pp. 542–542 (2009).

[43] H. Hassanzadeh and M. Keyvanpour, "A variance based active learning approach for named entity recognition," in Intelligent computing and information science. Springer, pp. 347–352 (2011).

[44] H. Hassanzadeh and M. Keyvanpour, "A two-phase hybrid of semi-supervised and active learning approach for sequence labeling," *Intelligent Data Analysis*, Vol. 17, No. 2, pp. 251–270 (2013).

[45] M. R. Keyvanpour and M. B. Imani, "Semi-supervised text categorization: Exploiting unlabeled data using ensemble learning algorithms," *Intelligent Data Analysis*, Vol. 17, No. 3, pp. 367–385 (2013).

[46] B. Scholkopf, R. C. Williamson, A. J. Smola, J. Shawe-Taylor, and J. C. Platt, "Support vector method for novelty detection," in Advances in neural information processing systems, pp. 582–588 (2000).

[47] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM computing surveys (CSUR)*, Vol. 31, No. 3, pp. 264–323 (1999).

[48] M. Koohzadi, "Event mining in video data with semi-supervised learning," Ph.D. dissertation, Alzahra University, Tehran (2012).

[49] E. Atashpaz-Gargari and C. Lucas, "Imperialist competitive algorithm: an algorithm for optimization inspired by imperialistic competition," *in Congress on Evolutionary computation, CEC 2007*. IEEE, pp. 4661–4667 (2007).

[50] P. Berka, "Pkdd'99 discovery challenge guide to the financial data set," https://sorry.vse.cz/ berka/challenge/pkdd1999/berka.htm9 (1999).

[51] T. S. Buda, T. Cerqueus, C. Grava, and J. Murphy, "Rex: Representative extrapolating relational databases," *Information Systems*, Vol. 67, pp. 83– 99 (2017).

[52] R. Frank, F. Moser, and M. Ester, "A method for multi-relational classification using single and multi-feature aggregation functions," in European Conference on Principles of Data Mining and Knowledge Discovery. Springer, pp. 430–437 (2007).

[53] R. Zall, "A semi-supervised learning based method for classification of multi-relational data," Ph.D. dissertation, Alzahra University, Tehran (2015).

[54] J. Zhang and Y. Tay, "Dscaler: Synthetically scaling a given relational database," *Proceedings of the VLDB Endowment*, Vol. 9, No. 14, pp. 1671–1682 (2016).

[55] S. Jamshidi, "Developing a dynamic multi-level model for creating a behavioral profile to detect fraud in electronic payments," Ph.D. dissertation, Tehran University, Tehran (2014).