*Journal of **AI** and **D**ata **M**ining*
Vol 7, No 2, 2019, 213-224

DOI: 10.22044/JADM.2018.6311.1746

MEFUASN: A Helpful Method to Extract Features using Analyzing Social Network for Fraud Detection

Z. Karimi Zandian[1] and M.- R. Keyvanpour[2*]

1. Data Mining Lab, Department of Computer Engineering, Alzahra University, Vanak, Tehran, Iran.
2. Department of Computer Engineering, Alzahra University, Vanak, Tehran, Iran.

Received 10 November 2017; Revised 09 April 2018; Accepted 16 May 2018

**Corresponding author: keyvanpour@alzahra.ac.ir (M. R. Keyvanpour).*

Abstract

Fraud detection is one of the ways to cope with damages associated with fraudulent activities that have become common due to the rapid development of the Internet and electronic business. There is a need to propose methods to detect fraud accurately fast. To achieve accuracy, fraud detection methods are required to consider both kinds of features, features based on the user level and features based on the network level. Therefore, in this paper, a method called MEFUASN is proposed to extract features based on social network analysis. After extracting these features, both the obtained features and the features based on user level are combined together to detect fraud using semi-supervised learning. Evaluation results show that using the proposed feature extraction as a pre-processing step in fraud detection improves the accuracy of detection remarkably, while it controls runtime in comparison with other methods.

Keywords: *Feature Extraction, Fraud Detection, Social Network Analysis, Semi-supervised Learning, Network Level Features, User Level Features.*

1. Introduction

Fraud detection describes the methodologies deployed in order to investigate allegations of fraud. It is more reactive than proactive [1]. In other words, when fraud occurs, it can be detected by different methods, as in the event of unauthorized use of another person's personal information. Therefore, fraud detection involves a review of historical transactions to identify indicators of a non-conforming transaction [2].

On one hand, in a fraud detection area, research works show that one of the challenges of many existing methods is not to consider features based on the user level and network level simultaneously to learn, while investigating these two kinds of features can help to increase the accuracy of fraud detection methods.

On the other hand, the computer industry has seen a large growth in technology, particularly in access, storage, and processing. This, combined with the fact that there are huge amounts of data to be processed has paved the way for data analysis and mining to derive potentially useful information. Various demands range from commercial to military needs to analyze data in an efficient and fast manner [3]. Data mining is a process that uses data analysis tools to uncover and find patterns and relationships among data that may lead to extraction of new information from a large database [4, 5]. One of the issues related to data is to convert raw data into a set of useful features, and another one is to identify the best and most useful features to analyze and extract [6]. Therefore, before applying learning algorithms to datasets, it is usually necessary to preprocess the data properly. Data preprocessing is a crucial, still neglected step, in data mining [7]. Feature extraction can be the pre-processing step of data mining. Feature extraction is to extract patterns and derive knowledge from large collections of data with identification and extraction of unique features for a particular domain. Though there are various features available, the aim is to identify the best features, and thereby, extract relevant information from the data [3]. Today, feature extraction is used in many fields such as image processing, text mining, signal processing, and pattern recognition.

In this paper, in order to cope with this challenge, we propose a novel and efficient feature extraction method based on the social network analysis for fraud detection in banking accounts. In the proposed feature extraction method, features based on both network and user level are extracted, and then with these features, learning starts. One of the basic methods based on semi-supervised learning is the PCKmeans method, which will be used in this paper to evaluate our feature extraction method.

The rest of the paper is organized as what follows. In Section 2, the related works are discussed. In Section 3, the proposed feature extraction method is introduced. Evaluation results are presented in Section 4, followed by the concluding remarks in Section 5.

2. Related works

Jamshidi et al. [8] have proposed a new feature extraction method based on social network analysis called bad-score to improve fraud detection. The proposed method is created from 4 phases: building social network, analyzing patterns, storing patterns, and updating. In this work, various features of transactions are used to detect fraud. Carneiro et al. [9] have developed a method to detect fraud in credit cards that combine manual and automated classification. In this work, the features and properties of credit cards are used. Save et al. [10] have devised a novel system for credit card fraud detection based on decision tree with combination of Luhn's algorithm and Hunt's algorithm using features of credit cards. Behera et al. [11] have proposed a fraud detection method based on fuzzy clustering and neural network using features of credit cards. Botelho et al. [12] have developed a feature that is obtained from social network called badRank to help improve the fraud detection using semi-supervised learning. Chiu et al. [13] have proposed features extracted from social network as the input of fraud detection classifiers. In [14], by analyzing social network, the patterns that are common to fraudulent entities are identified, and each entity is described by its original features plus another one for each pattern. Finally, these features are used in classification methods. Subelj et al. [15] have used some features extracted from social network to detect fraud. In [16], the use of features obtained from transaction history databases and the current and past behaviors of credit cards to detect fraud is proposed. Sadaoui et al. [17] have proposed a real-time framework that observes the progressing auctions to be able to take actions on time and set a fraud score for each user. This fraud score represents the user's behavior in past auctions. In [18], a fraud detection method based on neural network is proposed. Self-organizing map algorithm is used to extract cardholders' behavior and to learn and classify this behavior. Krivko [19] has used features based on the user to propose a model to detect fraud. The proposed data-customized approach combines elements of supervised and unsupervised methodologies aiming to compensate for the individual deficiencies of the methods. Chang et al. [20] have used changes of behavior in each user to detect fraud. In this work, clustering techniques are used to distinguish changes of behavior. Reviewing the proposed methods in the fraud detection area and classification, we proposed in [21] fraud detection methods which, based on the features of entities, can be divided into two categories: fraud detection methods based on the user level features and fraud detection methods based on the network level features. In fraud detection methods, based on the user level features, it is sufficient to investigate inherent and exclusive features derived from a specific component [21].

According to the classification presented in [21], in methods based on the network level features, features of each component are obtained considering a component position along with other components. The features then participate in fraud detection. These methods usually use the connections between components to obtain new features. In order to achieve this goal, social networks comprising these components are paid attention to, and useful information is obtained from them [21]. According to [21], the speed of detection in methods based on the user level is higher than that in methods based on the network level. In contrast, the complexity of methods based on the network level is more than that of the methods based on the user level [22]. Also the accuracy of these two kinds of methods is not high [8].

As a result, the feature extraction method proposed in this paper uses a combination of these two kinds of features to increase the speed and accuracy of fraud detection.

3. MEFUASN: a helpful method to extract new features of banking accounts using analyzing social network

As mentioned in [21], features of the components under consideration can be divided into two categories: network-based and user-based.

Features based on the network level are features that consider components obtained in the presence of other components, and include a set of components that are related to each other according to their relationships with others [15, 23, 24], while features based on the user level are features that belong to a certain component with no regard to the relationship between that component and others [16, 20, 25, 26]. Combining algorithms, each of which has focused on various aspects of information hidden in the data, can help detect fraudulent accounts more accurately [21] because in the first type of algorithms, existence or non-existence of relationships between the data is ignored, and in the second type of algorithms, individual frauds are not considered to be important. Therefore, in the proposed feature extraction method, we have used both feature types to detect fraudulent accounts. The challenge of many existing methods in this area is not to consider these two feature types simultaneously.

The block diagram for the proposed system of feature extraction (MEFUASN) is shown in figure 1.

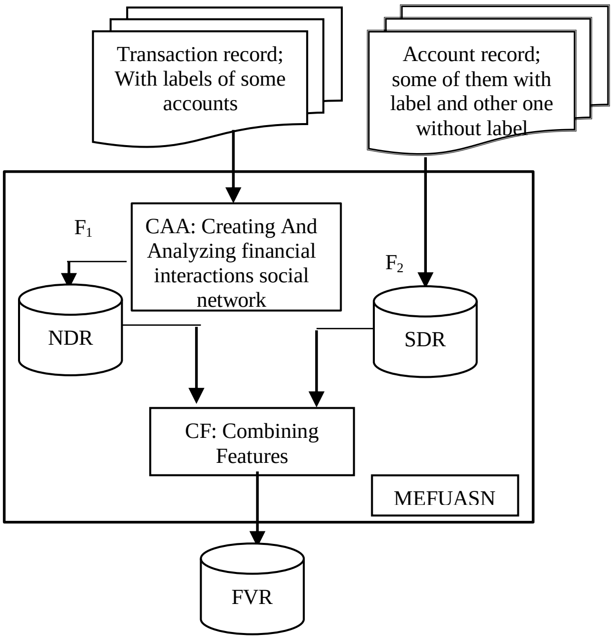

Transaction record; With labels of some accounts
Account record; some of them with label and other one without label
F$_1$
CAA: Creating And Analyzing financial interactions social network
F$_2$
NDR
SDR
CF: Combining Features
MEFUASN
FVR

Figure 1. Block diagram of MEFUASN.

Accordingly, as shown in figure 1, MEFUASN involves two steps:

- In step 1, in order to provide and obtain new features, a social network of financial interactions is created and analyzed. The method proposed in this paper uses transaction record as input for creating and analyzing financial interactions social network (CAA) phase.
- In step 2, features obtained from CAA phase, namely features 1 (F_1) in figure 1, and saved in network data repository (NDR) and account record namely features 2 (F_2) in figure 1, stored in simple data repository (SDR), are used as inputs for the combining features (CF) phase. In the CF phase, features and network data belonging to accounts obtained from Phase 1 are combined with features based on the user level of the existing accounts record called simple. These features are then shown as accounts features vectors and saved in feature vectors repository (FVR).

Using the criteria that demonstrate possible scenarios of fraudulence and the factor increasing the risk of fraud accounts can have an essential role in increasing fraud detection accuracy [27]. In order to achieve this goal, CAA is proposed and used in this paper. More specifically, the feature that is extracted from this network shows the score of fraudulence of each account because of the relationships with other fraudulent accounts.

CAA phase is based on creating and analyzing financial interaction social network. In the first step, financial interaction social network is created by receiving transaction record, and network data is achieved and saved in NDR. This repository is used to save network data in the next step.

3.1. Creating financial interactions social network

In this work, an implicit social network called financial interaction social network is used. What is important in detecting fraud in financial interactions is financial transactions between accounts [8]. Thus by considering accounts as the nodes [28] of financial interaction social network and financial transactions as edges [28], the hidden features in this network can be extracted. According to the method proposed in this paper, accounts for which at least a single transaction exists have a relationship [29]. Another remarkable property in the proposed network is that if the account of the receiver of the transaction and its sender account are different, the relationship between accounts is directed [30]. In figure 2, an example of this kind of network is shown.

Another important feature of the proposed network, in addition to being directed, is having weight. The relationship between any two nodes with each other is not equal, and as a result, to obtain the nodes' fake scores, they will not have the same influences. These weights [30] depend on various factors but what is considered in this paper is the number and total value of transactions between any two accounts affecting the weight of the edge between them. It seems likely that an account that is related to with a fraudulent account for many times with a high total value has a higher possibility of being fraudulent than an account that is related to a fraudulent account less frequently with a lower total value. As a result, edges between nodes are also weighted.

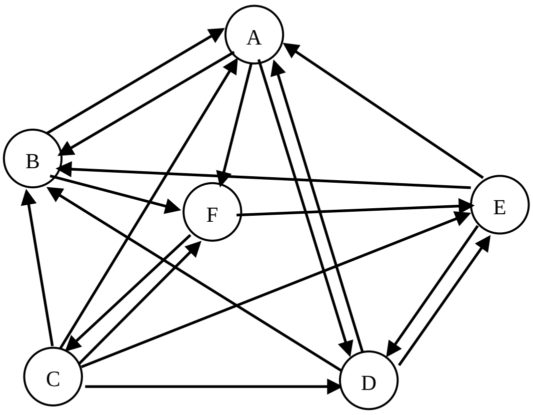


Figure 2. An example of directed social network.

As mentioned earlier, the proposed network is weighted and the weights of the edges affect the number and total value of transactions from account i to account j. Equations (1), (2), and (3) are used to weigh the edge between account i to account j.

$$RNT_{ij} = NT_{ij} / TNT \tag{1}$$

where, RNT_{ij} is the relative number from account i to account j, NT_{ij} is the number of transactions from account i to account j, and TNT is the total number of transactions.

$$RSTA_{ij} = STA_{ij} / TSTA \tag{2}$$

where, RSTA_{ij} is the relative sum of the amounts of transactions from account i to account j, STA_{ij} is sum of amounts of transactions from account i to account j, and TSTA is the total amounts of transactions.

$$weight_{ij} = \alpha.RNT_{ij} + (1-\alpha).RSTA_{ij}\; 0 \le \alpha \le 1 \tag{3}$$

where, α shows the relative importance of RNT, and RSTA affect calculating the weight of the edge from account i to account j.

As it is clear in (1), (2), and (3), factors affecting the weights of edges are the relative number (RNT) and relative sum of the amounts of transactions (RSTA) from account i to account j.

Applying these two suggestions to the financial interaction social network affects the accuracy of fraud detection. The pseudo-code for the financial interactions social network phase in the proposed CAA method is presented in figure 3.

3.2. Analyzing social network and extracting new features

In analyzing the social network phase, using special criteria, the social network created is analyzed, and new features for each account is extracted. Generally, algorithms are used to analyze social network focus on entities and nodes of the social network [27], while what seems to be more important is the information hidden in relationships between accounts in fraud detection in bank accounts. For example, Hits [31] and PageRank [32], which are basic algorithms for social network analysis, focus on determining the centrality of web pages. Other examples from this kind of algorithms are BadRank [12] and gspan [14] that pay attention to the nodes of the potential network, and do not consider relationships and their complexity and conditions involved in financial interactions. In front of these algorithms, the algorithms have been proposed that have paid attention to relationships in the simple social network [27]. Hence, in this paper, the directed and weighted network [30] is analyzed to propose new features.

```
Algorithm: creating financial interactions
           social network
Input: transactionData, accountData
Output: relationship matrix

Assign accounts as nodes and transactions
  as directed edges between nodes in social
  network
For each two nodes
  Unite all of edges and save number and
  amount of transactions between them
  Calculate weight of edge using equals (1),
  (2) and (3)
End
Create matrix including edges (two nodes)
   and their weights
```

Figure 3. Pseudo-code for creating financial interactions social network phase in the proposed CAA method.

A new feature called Fake_score, proposed in this paper, shows the fraudulence score. Thus a higher score of an account means that the account has a stronger relationship with fraudulence accounts. According to the criterion proposed in [8, 27], this criterion, in general, depends on 3 factors:

- Distance from fraudulence nodes
- Sum of the degrees of the nodes existing in the paths
- Number of paths ended to fraudulence nodes

Equation (4) shows how to calculate the Fake_score.

$$\begin{aligned}&Fake\_score(i) = \\ &A * PathElement(i) + B * DegreeElement(i) \\ &+(1-(A+B)) * EndPo\text{int}\,Element(i) \\ &0 \le A \le 1,\ 0 \le B \le 1-A\end{aligned} \quad (4)$$

where A and B show the relative importance of the three effective factors PathElement, DegreeElement, and EndPointElement.

As shown in (4), to calculate the Fake_score for each account, we use the weighted mean of the three factors PathElement, DegreeElement, and EndPointElement that define distance from fraudulence nodes, sum of the degrees of the nodes in the paths, and number of paths ending in fraudulence nodes. The degree of importance of each factor can change in various situations to calculate the Fake_score.

Since the main purpose of this work was to propose a new feature extraction method to increase the accuracy and speed of fraud detection, it must be noted that each relationship is not necessarily important. Considering a node with a number of weighted indegrees or outdegrees [30] that are more than a special threshold is not useful, and investigating the paths including this node is not necessary. In this paper, the average number of indegrees and outdegrees of all nodes in the network are special thresholds for a number of weighted indegrees and outdegrees of each node, respectively.

Another point investigated here is that if the distance between the examined node and the fraudulence nodes is higher than a threshold, that node does not seem to be dangerous. This is because fraudsters always try to show a normal behavior not to be detected quickly to achieve their goal. The maximum of the path [30] can also be changed from 2 up to the network diameter [30], and as mentioned in [8], the suitable value to achieve the highest accuracy is the average length of all paths in the network because whenever the search space depth in the network becomes more than this value, the possibility of fraud through relationships with fraudulent accounts becomes lower, and this search uses more time. In contrast, whenever the search space depth becomes less than the usual value, accuracy becomes lower; hence, we have to instate a trade-off between accuracy and speed. Thus, the length of paths [28] investigated here has been set to 4 for the approved dataset. In equations, the length of this path has been shown by Ψ.

3.2.1. Distance from fraudulence node factor

In the proposed criteria, in order to apply the effect of distance from fraudulent nodes, the PathElement component is used. Equations (5), (6), and (7) show how to calculate this factor.

$$PathElement(i,F) = \sum_{Allpath\_With\_Lenght \le \psi\_Between(i,F)} y \quad (5)$$

$$y = \begin{Bmatrix} 1 & SOWOP > AOWOP * \psi \\ SOWOP / (AOWOP * \psi) & else \end{Bmatrix} \quad (6)$$

$$PathElement(i) = \frac{\sum_{AllFraudAccs} PathElement\,(i,\, FraudAcc.)}{NOP\_With\_Lenght \le \psi\_between(i,\, AllFraudAccs)} \quad (7)$$

where ψ is a constant for the considered length of the path, SOWOP is the sum of the weights of the edges existing in the path from i to F, AOWOP is the average of the weights belonging to all paths in the network, and NOP is the number of paths.

This distance is influenced by the number of aligned edges between investigated node (i), fraudulent node (F), and weight of the mentioned edges. These aligned edges together are the paths in the directed graphs.

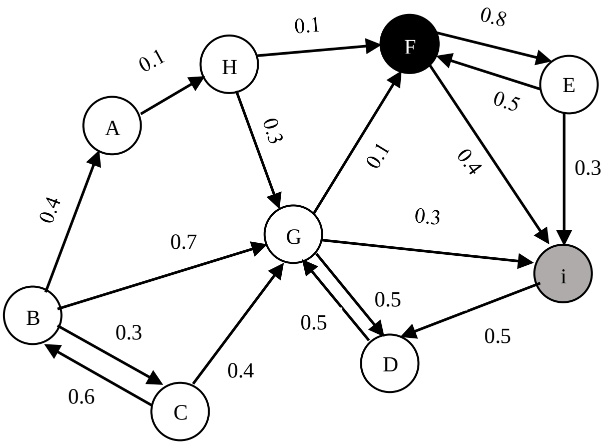


Figure 4. An example for existence of paths between an investigated node (i) and a fraudulent node (F).

An example for the existence of paths between an investigated node (i) and a fraudulent node (F) is shown in figure 4. For each path between i and F, (6) is calculated, and then according to (5), these amounts are calculated together. As mentioned in (7), according to the proposed method, PathElement of node i will be obtained by means of the calculated PathElement between i and each fraudulent node F. The more this value is, the more its effect is; and the less this value is, the less its effect is on Fake_score of the investigated account.

According to (6), it is clear that this factor has the highest effect on the Fake_score if the sum of the weights of existing edges in the path is more than the sum of the weights of edges in a path with length ψ with the weight that is the average of all of the weights in the network.

3.2.2. Sum of degree of nodes existing in path factor

Degrees of the existing nodes influence the Fake_score by DegreeElement component. Sum of the indegrees and outdegrees [33] in the path is examined separately to calculate the value of this component, and based on their weights, determine the sum of indegrees and outdegrees separately in a normal path. In this work, a normal path is considered but the weighted indegree and outdegree of each node in that path [30] are the mean of the weighted indegrees and the mean of weighted outdegrees in all paths of that network with length ψ. To calculate DegreeElement, (8), (9), (10), (11), and (12) are used.

$$DegreeElement(i,F) = \sum_{Allpath\_With\_Lenght \;\leq\; \psi\_Between(i,F)} x \qquad (8)$$

$$x = \left\{ \begin{array}{ll} 0 & z \geq 1 and\ s \geq 1 \\ \max\left(\frac{1}{z},\frac{1}{s}\right) & z < 1 and\ s < 1 \\ \min\left(\frac{1}{z},\frac{1}{s}\right) & else \end{array} \right\} \qquad (9)$$

$$z = \frac{\sum_{AllNodesInPath} WOIE}{AOWOE * NID * (PathLenght - 1)} \qquad (10)$$

$$s = \frac{\sum_{AllNodesInPath} WOOE}{AOWOE * NOD * (PathLenght - 1)} \qquad (11)$$

$$DegreeElement(i) = \frac{\sum_{1}^{AllFraudAccs} DegreeElement(i,\ FraudAcc.)}{NOP\_With\_Lenght \leq \psi\_between(i,\ AllFraudAccs)} \qquad (12)$$

where WOIE is the sum of the weights of the input edges for the node, AOWOE is the average of the weights of all edges in the network, NID is the normal indegrees for a node, WOOE is the sum of the weights of output edges for the node, and NOD is the normal outdegrees for a node.

According to (9), (10), and (11), if this relation in the path between the investigated node and a fraudulent node for all indegrees and outdegrees is at least equal to 1, the DegreeElement value for that node will be equal 0. This means that this component will have the least effect on the Fake_score of that node. Otherwise, if both of these ratios are less than 1, the value for this component is the maximum of these two ratios for the path. Finally, if only one of these ratios is at least equal to 1, the DegreeElement value for that node will be equal the minimum of the ratio of indegrees in the path to indegrees in a normal path and the ratio of outdegrees in the investigated path to outdegrees in a normal path. After calculating the DegreeElement for all paths between the investigated node and a fraudulent node, as mentioned in (8), with regard to a fraudulent node, the DegreeElement for that node will be calculated by sum of the obtained DegreeElement from all paths between the investigated node and that fraudulent node. Consequently, in order to calculate the total DegreeElement for each node, the DegreeElement values obtained from all paths between the investigated node and all fraudulent nodes are averaged (12).

3.2.3. Number paths ending in fraudulence nodes factor

As mentioned earlier, the number of fraudulent nodes that have relationships [29] with other nodes is important to calculate the Fake_score. According to the method proposed in this paper, when two nodes have a relationship with each other, the distance [30] between those two nodes in the network is utmost ψ. Thus the third component affecting the Fake_score called the EndPointElement is defined. According to (13), if all nodes related to the investigated node are frauds, the maximum value for this factor that equals 1 will be obtained. In contrast, if none of the related nodes are frauds, this component has the least effect on the Fake_score feature.

$$EndPo\text{int}\,Element(i) = \frac{\left|\left\{node\ w \middle| w \in \text{Re}\,latedNodes\_With\_i\ and\ Label(w) = 'fraud'\right\}\right|}{\left|\text{Re}\,latedNodes\_With\_i\right|} \qquad (13)$$

Pseudo-code of analyzing the social network and extracting new feature phase in the proposed CAA method is presented in figure 5.

4. Experiments

4.1. 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 [34]. This dataset is used to evaluate many methods in different fields [35-38]. Due to the availability of financial transaction data, demographic information, and validity of this dataset, it has been used here to test our proposed method. We have used the transaction table to form social network and the account table to extract simple data. We have also applied some changes to the transaction table like eliminating transactions that are not transactions for transferring money and those accounts whose information does not exist in accounts table.

```
Algorithm: Analyzing social network and
           extracting new feature
Input: relationship matrix
Output: Fake_score

For each node- fraudulent node
  For each fraudulent node
     Calculate PathElement (node- fraudulent
          node, fraudulent node) using equal (5)
     Calculate DegreeElement (node- fraudulent
          node, fraudulent node) using equal (8)
  End
  Calculate PathElement (node- fraudulent
          node) using equal (7)
  Calculate DegreeElement (node- fraudulent
          node) using equal (10)
  Calculate EndPointElement (node- fraudulent
          node) using equal (11)
  Fake_score (node-fraudulent node) using
          equal (4)
End
For each fraudulent node
    Fake_score (fraudulent node) = 1;
end
```

Figure 5. Pseudo-code of analyzing social network and extracting new feature phase in the proposed CAA method.

Table 1. 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

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 5 features: trans-id, source_account-id, destination_account-id, amount, and date. Each account also has 3 features: account-id, district-id, and date.

Transaction data is used to calculate the Fake_score for accounts through social relations using the proposed method. Based on relations, a score that represents the probability of a fraud activity as a new feature is assigned to unknown accounts.

4.2. Evaluation criteria

For performance evaluation of the proposed feature extraction methods for fraud detection, popular criteria are used: True Negative (TN) rate, False Positive (FP) rate, False Negative (FN) rate, precision, recall (also called True Positive (TP) rate), F_1score, accuracy, and runtime.

- TNrate: as (14) shows, it is the proportion of negatives that are correctly identified as such.

$$TNrate = \frac{TN}{TN+FP} \tag{14}$$

- FPrate: as stated in (15), it is the proportion of negatives that are wrongly identified as positives.

$$F\Pr ate = \frac{FP}{FP+TN} \tag{15}$$

- FNrate: the proportion of positives that are wrongly identified as negatives (16).

$$FNrate = \frac{FN}{FN+TP} \tag{16}$$

- Precision: as shown in (17), it is the number of items correctly labeled as belonging to the positive class (TP) divided by the total number of elements labeled as belonging to the positive class (i.e. the sum of true positives and false positives, which are items incorrectly labeled as belonging to the class).

$$precision = \frac{TP}{TP+FP} \tag{17}$$

- Recall: the number of true positives divided by the total number of elements that actually belong to the positive class (i.e. the sum of true positives and false negatives, which are items not labeled as belonging to the positive class but should have been) (18).

$$recall = T\text{P}rate = \frac{TP}{TP+FN} \tag{18}$$

- F_1score: as stated in (19), it is the harmonic mean of precision and recall.

$$F_1score = \frac{2*precision*recall}{precision+recall} = \frac{2TP}{2TP+FP+FN} \tag{19}$$

- Accuracy: the proportion of positives and negatives that are correctly identified as such (20).

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad (20)$$

- Runtime: the time used to perform the method completely, obtain the results, and label the data.

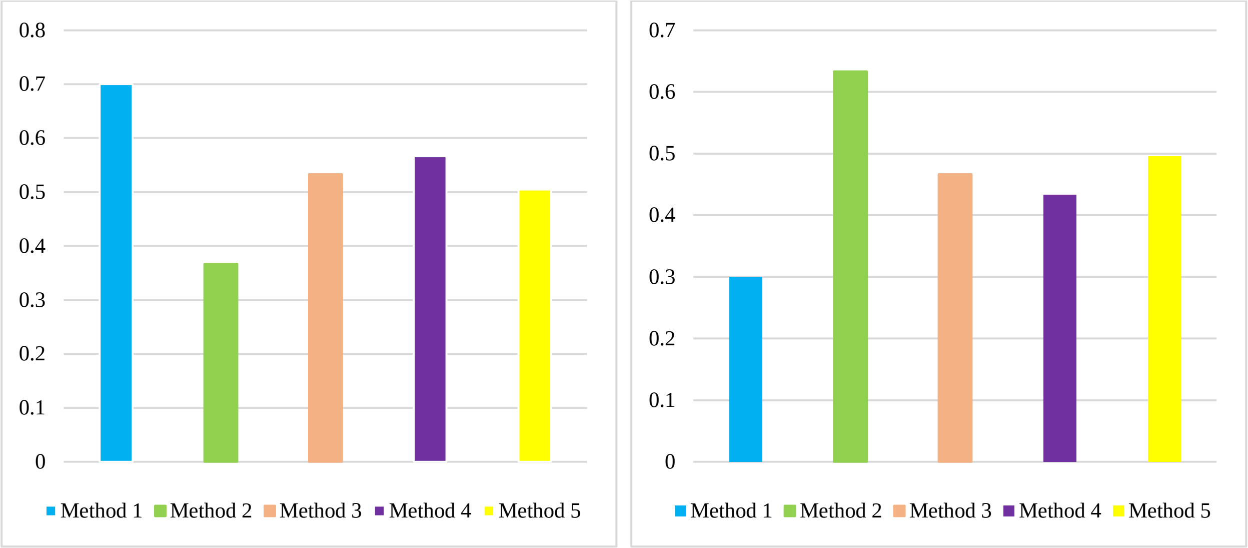
4.3. Experimental results

For evaluating MEFUASN, we have compared PCKmeans with Fake_score feature obtained from MEFUASN and called Method 1 in figures 6, 7, and 8 with other four methods based on the eight criteria expressed in the section on evaluation criteria. These methods include bad-score feature proposed in [27], Fake_score obtained from undirected social network, Fake_score obtained from unweighted social network and without feature extraction phase that have called Methods 2, 3, 4, and 5, respectively.

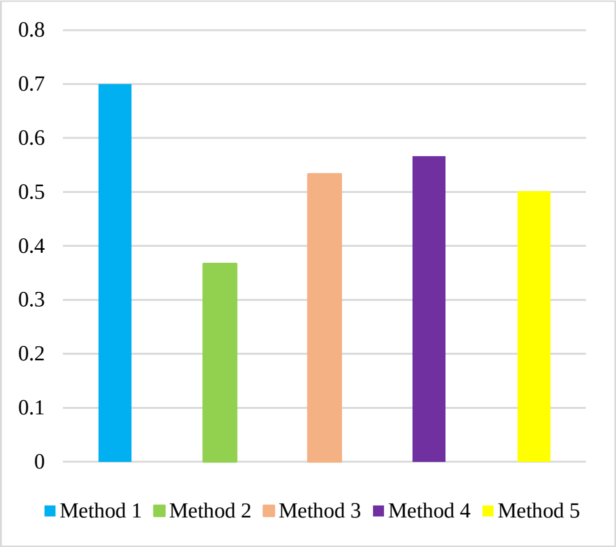
Our method is compared with what is proposed in [27] because the feature extracted in this paper has also been obtained from social network using relationships among the nodes (i.e. network level feature) but from a simple, undirected, and unweighted network.

As shown in figures 6(a) and 6(b), TNrate and FPrate corresponding to PCKmeans with the Fake_score feature obtained from MEFUASN is higher than the other methods. This means that the proposed MEFUASN has had a major effect on increasing TNrate, decreasing FPrate, detecting non-fraud accounts correctly, and reducing wrong alarms. This is because of correct and appropriate restrictions applied on Fake_score calculation such as considering weighted and directed network and ignoring non-effective factors in detecting frauds correctly.

The length of the paths (ψ) has been set to 4 in order to control runtime and as some useful information hidden in longer paths may not be considered and relationships among some nodes may be ignored. Furthermore, in the proposed method, to detect fraud, there are other limits. For example, relational loops in social network have not been investigated, and to extract network features, it has been supposed accounts with the same owner do not exist, while the existing accounts with the same owner can include some information that helps us detect fraud. For example, the existing accounts with the same owner that have relationship with each other can have useful information to detect fraud. It seems that accounts with the same owners must be studied more. Therefore, as shown in figures 6(c) and 7(a), the rate of detection of fraud accounts (recall) is less, and thus the amount of FNrate is more in the proposed method than the other methods. Lack of feature extraction phase in learning has led to labelling as positive (fraud) the data that is not fraud, and so its TNrate is lower and its recall is more than other. However, in PCKmeans with bad-score using the method proposed in [28], the recall of this method and TNrate amounts are modest and lie between our method, and PCKmeans method without feature extraction phase because it has used simple network and it is possible to calculate higher scores for a node because of finding relationships between that node and a fraud node, while this is not distinguished in our methods.

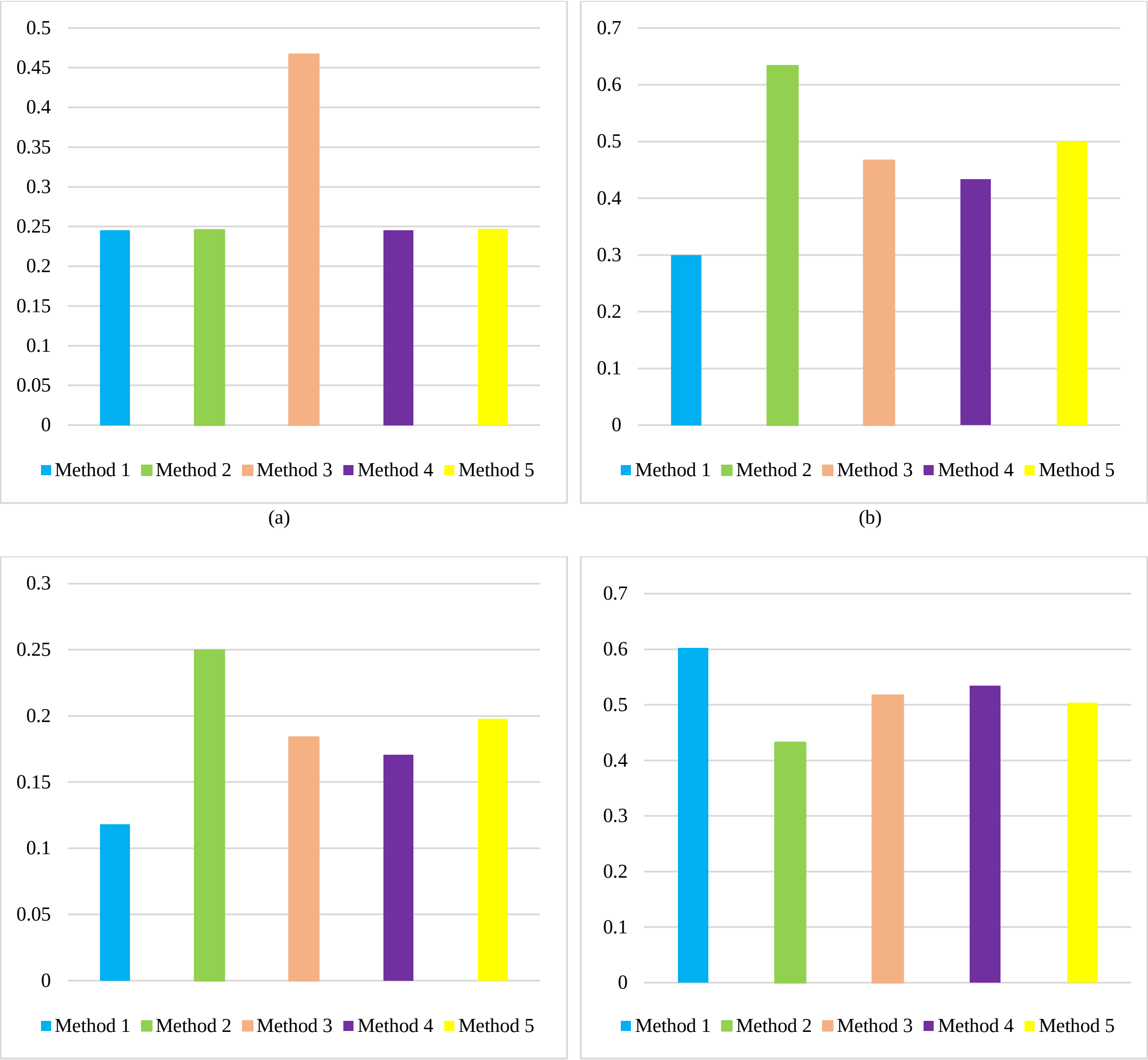


(a) (b)



(c)

Figure 6. Comparison between the proposed method and other methods based on (a) TNrate, (b) FPrate, (c) FNrate.



(a)

(b)

(c)

(d)

Figure 7. Comparison between the proposed method and other methods based on (a) recall, (b) precision, (c) F_1score, (d) accuracy.

As shown in figure 7(b) and with regard to (17), since the difference the amounts for TP and FP is more in PCKmeans with bad-score than others, the precision criterion for this method is better.

All the studied feature extraction methods in paper have tried to reduce wrong detections in addition to increasing correct detections and also adding a feature showing probability of being fraud causes fraud detection rate (whether TP or FP) is lower than PCKmeans without feature extraction phase. While TP is the only influencing parameter on F_1score shown in figure 7(c) and by paying attention to (19), PCKmeans without feature extraction phase is better based on the F_1score criterion.

As shown in figure 7(d), PCKmeans with MEFUASN is better than others based on accuracy. Since in this paper, the proposed method aimed to detect fraud and non-fraud correctly and simultaneously, we used the feature based on network level inside features based on the user level. We also used weighted and directed network; using weighted network improved TNrate (Figure 6(a)) and using directed network improved recall (Figure 7(a)).

As shown in figure 8, it is clear that the runtime of PCKmeans without feature extraction phase is very low. However, among the other methods, the study in [27] to obtain bad-score feature has used a simple network with less processing but the runtime of PCKmeans with MEFUASN method is similar. Runtime of PCKmeans with Fake_score using unweighted network and undirected network is much higher because of the higher complexity of the network and the large volume of the calculations.

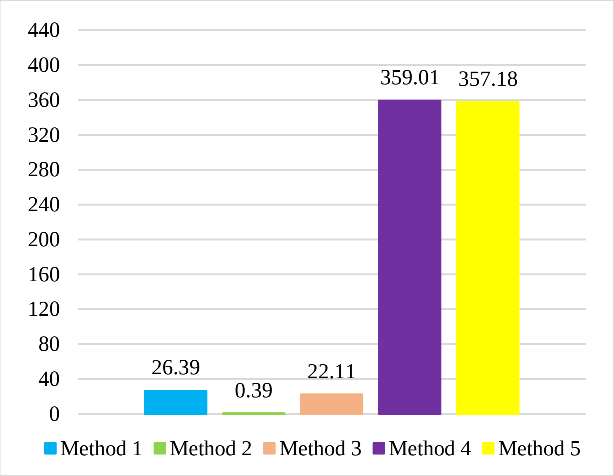


Figure 8. Comparison between the proposed method and other methods based on runtime (s).

5. Conclusion

In fraud detection area, there are two important challenges: accuracy and speed of detection [39, 40]. In this paper, a novel feature extraction method called MEFUASN as a pre-processing step has been proposed. In this method, both the user level features and network level features [21] are used. Based on the proposed method, financial interaction social network is first created and analyzed, and a new feature is extracted, and then this feature combines with the user level features. This network is weighted and directed. It was shown in the experimental results that the use of this method as the pre-processing step for fraud detection improves the accuracy of detection remarkably, while the runtime of fraud detection method is controlled and kept within an acceptable level compared to other methods.

References

[1] Petrucelli, J. (2013). Detecting fraud in organizations: Techniques, tools, and resources. USA, New Jersey. John Wiley & Sons.

[2] Cendrowski, H., Petro, L., Martin, J. & Wadecki, A. (2007). The handbook of fraud deterrence. USA, New Jersey. John Wiley & Sons.

[3] Foschi, P. G., Kolippakkam, D., Liu, H. & Mandvikar, A. (2002). Feature Extraction for Image Mining, In: Multimedia Information Systems, pp. 103-109.

[4] Karimi Zandian, Z., Keyvanpour M. (2017). 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.

[5] Mosavi, A. (2014). Data mining for decision making in engineering optimal design, Journal of AI and Data Mining, vol. 2, no. 1, pp. 7-14.

[6] Guyon, I. & Elisseeff, A. (2006). An introduction to feature extraction. In: Feature extraction. Berlin, Heidelberg: Springer, vol. 207, pp. 1-25.

[7] Kirchner, K., Zec, J. & Delibašić, B. (2016). Facilitating data preprocessing by a generic framework: a proposal for clustering, Artificial Intelligence Review, vol. 45, no. 3, pp. 271-297.

[8] Jamshidi, S. & Hashemi, M. R. (2012). An efficient data enrichment scheme for fraud detection using social network analysis. Sixth International Symposium on Telecommunications (IST), Tehran, Iran, 2012.

[9] Carneiro, N., Figueira, G. & Costa, M. (2017). A data mining based system for credit-card fraud detection in e-tail, Decision Support Systems, vol. 95, pp. 91-101.

[10] Save, P., Tiwarekar, P., Jain, K. N. & Mahyavanshi, N. (2017). A Novel Idea for Credit Card Fraud Detection using Decision Tree, International Journal of Computer Applications, vol. 161, no. 13, pp. 6-9.

[11] Behera, T. K. & Panigrahi, S. (2017). Credit Card Fraud Detection Using a Neuro-Fuzzy Expert System, In: Computational Intelligence in Data Mining. Springer, Singapore, vol. 556, pp. 835-843.

[12] Botelho, J. & Antunes, C. (2011). Combining Social Network Analysis with Semi-supervised Clustering: a case study on fraud detection. Proceeding of Mining Data Semantics (MDS'2011) in Conjunction with SIGKDD,2011.

[13] Chiu, C., Ku, Y., Lie, T. & Chen, Y. (2011). Internet auction fraud detection using social network analysis and classification tree approaches, International Journal of Electronic Commerce, vol. 15, no. 3, pp. 123-147.

[14] Almeida, M. P. (2009). Classification for fraud detection with social network analysis, Master Degree Dissertation, Engenharia Informática e de Computadores, University of Lisbon, Portugal, Lissabon.

[15] Šubelj, L., Furlan, S. & Bajec, M. (2011). An expert system for detecting automobile insurance fraud using social network analysis, Expert Systems with Applications, vol. 38, no. 1, pp. 1039-1052.

[16] Panigrahi, S., Kundu, A., Sural, S. & Majumdar, A. K. (2009). Credit card fraud detection: A fusion approach using Dempster–Shafer theory and Bayesian learning, Information Fusion, vol. 10, no. 4, pp. 354-363.

[17] Sadaoui, S., Wang, X. & Qi, D. (2015). A Real-Time Monitoring Framework for Online Auctions Frauds. International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, Seoul, Korea (Republic of), 2015.

[18] Zaslavsky, V. & Strizhak, A. (2006). Credit card fraud detection using self-organizing maps, Information and Security, vol. 18, pp. 48-63.

[19] Krivko, M. (2010). A hybrid model for plastic card fraud detection systems, Expert Systems with Applications, vol. 37, no. 8, pp. 6070-6076.

[20] Chang, W. H. & Chang, J. S. (2010). Using clustering techniques to analyze fraudulent behavior changes in online auctions, International Conference on Networking and Information Technology, Manila, Philippines, 2010.

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

[22] Gaol, F. L., Kadry, S., Taylor, M. & Li, P. S. (2013). Recent trends in social and behaviour sciences, Proceeding of the 2nd International Congress on Interdisciplinary Behaviour and Social Sciences, (ICIBSoS 2013), Jakarta, Indonezia, 2013.

[23] Lin, J. L. & Khomnotai, L. (2014). Using Neighbor Diversity to Detect Fraudsters in Online Auctions, Entropy, vol. 16, no. 5, pp. 2629-2641.

[24] Lin, S. J., Jheng, Y. Y. & Yu, C. H. (2012). 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.

[25] Yu, C. H. & Lin, S. J. (2013). Fuzzy rule optimization for online auction frauds detection based on genetic algorithm, Electronic Commerce Research, vol. 13, no. 2, pp. 169-182.

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

[27] Jamshidi, S. (2014). Developing a Dynamic Multi-Level Model for creating a Behavioral Profile to Detect Fraud in Electronic Payments, Master Degree Dissertation, School of Electrical and Computer Engineering, Tehran University, Iran, Tehran.

[28] Kosorukoff, A. & Passmore, D. L. (2011). Social Network Analysis: Theory and Applications. Passmore, D. L.

[29] Aggarwal, C. C. (2011). An introduction to social network data analytics. In: Social network data analytics. Springer, Boston, MA, Springer, pp. 1-15.

[30] Wasserman, S. & Faust, K. (1994). Social network analysis: Methods and applications. United Kingdom, Cambridge. Cambridge university press.

[31] Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment, Journal of the ACM (JACM), vol. 46, no. 5, pp. 604-632.

[32] Haveliwala, T. (1999). Efficient computation of PageRank, Technical report, Stanford University. Stanford.

[33] Trudeau, R. J. (2013). Introduction to graph theory. Kent, Ohio. New York: Courier Corporation.

[34] Berka, P. & Sochorova, M. (1999) .Discovery challenge guide to the financial data set, PKDD-99, Available: http://lisp.vse.cz/pkdd99.

[35] Zall, R. (2015). A Semi-supervised learning based method for Classification of Multi-Relational Data, Master Degree Dissertation, Faculty of computer Engineering, Alzahra University, Iran, Tehran.

[36] Frank, R., Moser, F. & Ester, M. (2007). A method for multi-relational classification using single and multi- feature aggregation functions. European Conference on Principles of Data Mining and Knowledge Discovery, Warsaw, Poland, 2007.

[37] Buda, T. S., Cerqueus, T., Grava, C. & Murphy, J. (2017). ReX: Representative extrapolating relational databases, Information Systems, vol. 67, pp. 83-99.

[38] Zhang, J. & Tay, Y. (2016). Dscaler: Synthetically scaling a given relational database, VLDB Endowment, vol. 9, no. 14, pp. 1671-1682.

[39] Seeja, K. & Zareapoor, M. (2014). FraudMiner: A Novel Credit Card Fraud Detection Model Based on Frequent Itemset Mining, The Scientific World Journal, vol. 2014.

[40] Raj, S. B. E. & Portia, A. A. (2011). Analysis on credit card fraud detection methods, International Conference on Computer, Communication and Electrical Technology (ICCCET), Tamilnadu, India, 2011.

دانشگاه صنعتی شاهرود

نشریه هوش مصنوعی و داده کاوی

IAIDM

MEFUASN: یک روش مفید برای استخراج ویژگی‌ها با استفاده از تحلیل شبکه اجتماعی جهت کشف تقلب

زهرا کریمی زندیان [1] و محمدرضا کیوانپور[2,*]

[1] آزمایشگاه داده‌کاوی، دانشکده مهندسی کامپیوتر، دانشگاه الزهرا (س)، ونک، تهران، ایران.

[2] دانشکده مهندسی کامپیوتر، دانشگاه الزهرا (س)، ونک، تهران، ایران.

ارسال ۲۰۱۷/۱۱/۱۰؛ بازنگری ۲۰۱۸/۰۴/۰۹؛ پذیرش ۲۰۱۸/۰۵/۲۲

چکیده:

کشف تقلب یکی از راه‌های مقابله با خسارت‌های ناشی از فعالیت‌های تقلبی است که به دلیل توسعه سریع اینترنت و تجارت الکترونیک رایج شده است. در نتیجه نیاز به ارائه متدهایی برای کشف دقیق و سریع تقلب وجود دارد. برای رسیدن به دقت، متدهای کشف تقلبی مورد نیاز هستند که هر دو نوع ویژگی مبتنی بر سطح کاربر و مبتنی بر سطح شبکه را مورد بررسی قرار دهند. بنابراین در این مقاله، متدی تحت عنوان MEFUASN برای استخراج ویژگی‌ها ارائه می‌شود که مبتنی بر تحلیل شبکه اجتماعی است. پس از استخراج این ویژگی‌ها، ویژگی‌های به دست آمده و ویژگی‌های مبتنی بر سطح کاربر با هم ترکیب می‌شوند تا تقلب با استفاده از یادگیری نیمه نظارتی کشف شود. نتایج ارزیابی نشان می‌دهد که استفاده از استخراج ویژگی پیشنهاد شده به عنوان یک گام پیش پردازش در کشف تقلب دقت کشف تقلب را به طور قابل توجهی بهبود می‌بخشد، در حالی که زمان اجرای آن را نیز در مقایسه با دیگر متدها کنترل می‌کند.

کلمات کلیدی: استخراج ویژگی، کشف تقلب، تحلیل شبکه اجتماعی، یادگیری نیمه نظارتی، ویژگی‌های سطح شبکه، ویژگی‌های سطح کاربر.