The discovery of the credit card transactions suspicious of fraud using unsupervised data-mining methods (single-link hierarchical clustering)

Zohreh Darbandian
Department of Computer Engineering, Yazd Branch, Islamic Azad University, Yazd, Iran

Alimohammad Latif *
Department of Electrical and Computer Engineering, Yazd University, Yazd, Iran

Sima Emadi
Department of Computer Engineering, Yazd Branch, Islamic Azad University, Yazd, Iran

*Corresponding Author: Alatif@yazd.ac.ir

Abstract

Fraud, the intentional misuse of the resources for personal interest, is an unlawful act. The discovery of the fraudulent actions is an operation which takes place in respect to concrete decision making regarding a suspicious behavior. The method proposed in the present dissertation is the discovery of fraudulent acts by making use of the unsupervised data mining methods (single-linkage hierarchical clustering method) by the application of which the amount of the data suspicious of cheating can be identified in credit cards. The collection of the data gathered for this purpose pertained to the transactions carried out by some individuals in 2013, and the data were extracted from a state bank in Tehran. The data analysis and methods implementation stages have been conducted by taking advantage of R-software. The (unsupervised) data-mining technique provides for better results in respect to the other algorithmic data-extraction methods. Due to the confidential nature of the bank data in Iran, the available data for the identification of the data suspicious of being fraudulent are unlabeled (fraudulent and non-fraudulent together) data. The results obtained from the present study indicated meaningful and sensible findings and the method used by the current thesis article was applied to reveal the suspicious cases from among the great many of the performed transactions and then in the end it enabled the bank supervisors to only scrutinize the suspicious cases with a higher precision and it was not necessary to survey the entire volume of the bank transactions which were too many.

Keywords: fraud, data-mining, clustering, suspicious bank data.
1. Introduction

The world today is a technological one and the atmosphere dominating the contemporary life propels the human beings towards mechanizing it and making it much easier. Nowadays, it has become compulsory to coordinate everything, even the tasks which do not seem to be important and necessary, to a mechanized life. Through the increasingly high development in technology and replacement of the traditional and old-fashioned structures by modern devices in all of the fields and areas, following this upgrade, the banking system is seeking to provide its customers and clients with the highest favorability possible regarding the technological progresses and advances and to be able to make use of electronic banking system. The credit cards which are considered as one of the most substantial reasons behind the electronic banking growth, have now been transformed into one of the most widely used banking tools, thus a great deal of the fraudulent activities have been directed to the transactions carried out by such cards. Credit cards and cash cards are known as bank cards or cheque cards which enable the possessors to have access to their bank accounts in the nearest financial institute. Some of these cards have been deposited with a sum of money by the use of which the payment operation can be fulfilled, in such a manner that a message containing the cash transfer to the receiver account is sent to the card owner’s bank. This card can be applied as an alternative to cash when purchasing things. In some of the cases a main account number is devoted to the exclusive internet use and there would be no more of a physical card. In the majority of the countries, the use of cash cards is very common, in such a manner that the transactions volume via such cash cards is more than cheques and/or cash. The development of the cash cards unlike the credit cards is different from country to country and this has resulted in the emergence of various systems worldwide which mostly do not correspond to one another.

However, unlike the credit cards, the sum paid by the cash card is withdrawn from the account of the card owner and there is no possibility for the card owner to pay it later. The cash cards usually provide for cash withdrawal like ATM cards, there has not been devised a comprehensive mechanism for the identification and prevention from frauds related to the card-based transactions up to date [1]. And it is for the lack of such an appropriate system that many of the fraudulent acts remain unrecognized. In the other countries, also the performed researches and studies have focused on credit cards due to the extensive use of them, while the use of such cards has not yet become so prevalent in our country and the most important transactions are more or less carried out via the (prepaid) cash cards. Also, according to the security concerns, the studies undertaken cannot be thoroughly published and they cannot be taken into proper exploitation. Therefore, taking advantage of the designed models in the other countries research literature is not that much affordable. According to the vast volume of the daily banking transactions and the need to recognize the frauds in a timely manner and prevent them from occurring it is not possible to identify them manually and it entails labor and time-intensive measures. The discovery of the frauds is an operation which is conducted to make a concrete decision regarding a suspicious act or behavior. Therefore, the main objective of the present study is the proposition of the unsupervised data-mining (single-linkage hierarchical clustering) method for the identification of the amount of the data suspicious of fraud in credit cards.
2. Theoretical principles and study hypotheses statement:

2.1. Electronic banking:
Electronic banking is an integrated system which offers the entire array of the banking services and products and the navigation and management operations through the use of electronic equipments to a centralized database in the format of a system. The different kinds of the frauds and the attack-recognition approaches can be divided into two broad categories based on the attacks types:

**Misuse detection:** this method attempts to recognize the previously observed attacks in the format of a signature pattern. When a transaction is carried out, it would be compared with the previously existing signature specimens and previously recognized attacks and in case a similarity is recognized the transaction would be identified as an attack.

**Anomaly detection:** in this method it is tried to create a diagnostic feature for each of the user’s performance history and after a big enough deviation has been extracted in the user’s diagnostic feature an attack is detected. Due to the unconstrained nature of this method, the ability to recognize the new attacks is considered as an advantage of this method. This method is, in fact, the recognition of the unauthorized efforts for having access to the system. The anomaly detection method, contrary to the misuse detection, is not based on previously recognized solutions and signatures; rather its mechanism is based on the clients’ behavior analysis. According to the anomaly detection method and misuse detection, there have been numerous techniques designed and implemented to identify fraud which are going to be pointed out in the sections to come.

2.2. Data-mining
The novel data-mining knowledge is one of the ten developing knowledge realms which is predicted to face the next decade with a technological revolution and for the same reason it has undergone a remarkably rapid expansion during the recent years in a global level. The data-mining knowledge is the process of knowledge discovery in a database. Data-mining includes the process of extracting valid previously unrecognized information in an understandable and reliable manner from the large databases and taking advantage of it in making decision for critical business activities and the analysis of the observable data for figuring out the inter-data confident relationships. In all of the definitions proposed for data-mining there has been made references to concepts such as knowledge discovery, finding and analyzing the inter-data patterns. Seemingly, knowledge discovery and data-mining are taken as two synonyms but in fact data mining is a part of knowledge discovery process [20].

2.2.1. The knowledge discovery process steps from the database:
Knowledge discovery is a repetitive combination which includes the following stages [20]:

- Data storage
- Data cleaning
- Data selection
- Data transformation
- Data mining
- results evaluation

As it is observed, data-mining is one step of the process of knowledge discovery which plays a critical role as part of the data knowledge discovery.

2.3. Clustering:
Clustering is the process of dividing a series of data or objects into subclasses called clusters by the use of which procedure it would become more easier to understand the data structure and data groupings. Each cluster includes a series of similar data which act in the form of a group. A clustering is a collection of the data which are similar to one another in comparison to the other data of the same cluster but different from the other clusters samples of data.

2.3.1. The main approach in clustering
The clustering techniques can be categorized from various viewpoints to different sets:
- Partitioning method
- Hierarchical method
- Density-based method
- Grid-based method
- Model-based method

2.3.1.1. Hierarchical clustering and flat clustering:
In the hierarchical clustering technique we are faced with one of the two agglomeration and/or divisive actions.
The type-I algorithm: at first we assign each of the samples to a cluster and the algorithm type II: all of the samples are assigned to a single cluster. In hierarchical clustering method, the final cluster is assigned with a hierarchical structure based on its genericity rate; the same as single-linkage method. But in the flat clustering method all of the final clusters have a similar genericity rate, such as k-means. The hierarchical structure resultant from the flat clustering method is called a Dendogram. The clustering methods by the use of dendograms are usually divided into two sets based on the generative hierarchical structure as stated below:
1. Top-down or divisive: in this method, firstly, the entire data are considered as one single cluster and then during an iterative process the data having less similarity are cleaved into separate clusters and this continues until clusters are reached with one member.
2. Bottom-up or agglomerative: in this method, firstly, the data are considered as separate clusters and then during an iterative process the clusters with more similarities are merged in every step to finally reach to a single cluster or a certain number of clusters. Among the different types of common agglomerative hierarchical clustering algorithms, single-link, average-link, complete link algorithms can be enumerated. The chief difference in all of these methods pertains to the way the similarities between the clusters are calculated.
Single-link clustering method:
This is one of the oldest and simplest methods of clustering and it is considered as a member of the hierarchical and exclusive clustering method. This clustering method is also called nearest-neighbor technique. In this method, the following measure is taken advantage of to compute the similarity between two clusters, naming A and B:

\[ d_{AB} = \min_{i \in A, j \in B} d_{ij} \]

Where, i is a sample data belonging to the cluster A and j is a sample data belonging to cluster B. In fact, in this method, the similarity between two clusters is defined as the smallest distance between one member from one cluster to another member from another cluster. This idea has been better conceptualized in the following figure [24].

Figure 1: the similarity between two clusters in single-link method equal to the smallest distance between the two clusters’ data

According to the fundamental principles posed till now, the following hypotheses have been proposed and analyzed to reach to the main study objective:

H1: the fraudulent data patterns in the POS-bank can be extracted by taking advantage of the unsupervised technique.
H2: solutions can be offered to reduce the frauds and fraudulent transactions by making use of the presented pattern.

2.4. A review of the previously undertaken studies:
In 2008, Srivastava et al made use of Markov’s hidden model in which the credit cards transactions have been investigated by the use of such a model, in such a manner that if a transaction cannot be accepted by a great likelihood margin it can be named to be a fraud. This method is consisted of two stages of learning and recognition. In the first stage the clustering algorithm is taken into practical use to cluster the extant transactions. In the second stage which is the fraud recognition step when a new transaction enters the system the related strand is formed and it will be compared with the card-owner previous strand. If the difference of the two strands is found in excess of a certain threshold then the new transaction is recognized as the fraudulent one and the strand related to the card owner is updated by the new transaction,
otherwise [3]. In 1994, Ghosh et al presented a three-layer feed-forward radius-limited perceptron network model. The model has been used in pattern identification. In this model, the test data are read only twice. In the exit layer a numeral amount is created as the transaction rank and if it is found out to be lower than a threshold value the transaction is recognized as a fraudulent one [4].

In 1994, Brause offered a fraud management system, Falcon, which is a very strong tool for preventing from the debt cards and credit cards misuse by the fraudulent cheaters. There has been made use of perceptron networks algorithm in this method, this system estimates the fraud likelihood on an account by taking advantage of comparing the current and the past transactions by the card owner. The system is capable of being trained with the card owner expenditure pattern by making use of perceptron networks and recognizing any changes in the way and the quality the sums of money are paid and declaring it as a fraud. In 2002, Bolton et al categorized the fraud discovery models in credit cards areas based on two approaches: supervised and unsupervised and they made use of clustering methods for discovering the frauds in bank cards. By the help of such a method, the accounts which reflect a certain different behavioral pattern in a time interval are identified [6]. In 2004, Tue et al presented a framework based on security and case-based reasoning systems to recognize fraudulent transactions. When a new transaction is carried out its similarity is calculated by the probers and if the similarity is higher than a certain threshold an alarming signal is activated. After confirmation by an expert, the transaction is exploited to be used for improving the model based on it being fraudulent or not [5].

In 2005, Vatsa et al presented a two-layer system in the first layer of which the general client-specific rules and regulations are applied to determine the suspiciousness degree of a transaction; but, since these rules and regulations are fixed and because the yearly fraudulent transactions which escape detection and the non-fraudulent transactions which are introduced as fraud can be very high, a second layer is applied here which makes use of the game theory technique for the recognition of the fraud. The fraudulent individual and the fraud-detection system are two ends of this game both of which try to maximize their own advantages [7]. In 2006, Kundu et al proposed an idea that because the fraudulent person is not familiar with the original owner of the card’s behavioral pattern and at the same time tries to make the maximum use in the shortest time possible, thus, there are usually found deviations from the card owner’s normal behaviors which can be used for the discovery of the fraud. Most of the existing systems work either based on false behaviors or deviations from the normal behavior. The first set cannot recognize the new fraudulent patterns and, quite contrarily, the second set introduce many of the normal behaviors as the false ones [8]. In 2007, Quah et al made use of clustering competencies in perceptron networks. The first stage includes the preliminary transactions validation which evaluates the PIN address. The second stage is clustered corresponding to the input data organization map and the hidden patterns can be discovered. In fact, such an application filters the entrance to the next stage. Then the output resulting from this stage is sent to a perceptron network to finally assign each transaction with a number as the amount of suspiciousness. In the last stage the transactions with scores higher than a threshold are reviewed [9].

In 2007, Shen et al used three methods for the classification f the data to recognize the fraud: decision tree, logistic regression, and perceptron network. The comparison between these three
methods has indicated that logistic regression and perceptron outperform decision tree. Every transaction is found to have specific characteristics based on which the transaction is assigned to a class, so the objective of the classification is to construct a function which maps very transaction onto one of the several predetermined groups based on the values given to its characteristics [10]. In 2008, Gadi et al applied five clustering methods for fraud detection: neural networks, Naïve Bayes, Bayesian network, artificial immune system, decision tree. Except the artificial immunity system for which there has been written separate programs, each of these methods have been evaluated in two sensitive to cost and simple methods. Also, as for the parametrical methods, these methods have been evaluated once with Weka default parameters (Weka is the name of a type of free software which is comprised of a collection of machine learning and data-mining algorithms) and once with the optimized parameters. The results of the comparison between these two states indicate that the Weka default parameters have not been optimum in none of the methods [11].

In 2008, Guo et al used the neural networks to recognize the fraudulent transactions and to build the neural network there was made use of a concept called confidence rate. After the neural network was built, upon entering a new transaction to the network, if the transaction output, which was its confidence rate, was found lower than a threshold the transaction was considered as a fraudulent one. The determination of the threshold is a key issue in algorithm efficiency. This method has been used for comparing the imaging discrepancies in medical physics. This curve is illustrative of the ration between FPR (correctly recognized fraudulent transactions to all fraudulent transactions ratio) and TPR (correctly recognized normal transactions to all normal transactions ratio) [12]. In 2008, the fuzzy logic method was used to identify the suspicious behaviors in internet banking. In this study, to take advantage of the fuzzy system competencies in the identification of the fraudulent data, firstly the entire collection of the clients’ past behaviors were classified as the knowledge base in several different levels and the system was trained and then the fuzzy expert system was designed to infer the outputs. Although the system incorporates a vast array of the factors for identifying the user’s behavior and performance it spends a lot of time to process the factors and the input factors [13]. In 2010, Pulina et al proposed statistical methods based on this fundamental presupposition that the occurrence likelihood of the normal sample data in a randomized model is higher than the occurrence likelihood of the abnormal sample data. Most of the fraudulent statistical methods construct a data distribution likelihood model and evaluate it for every transaction therefore the transactions with low occurrence likelihood are classified as abnormal [18].

In 2011, Duman et al made use of a tail adjustment-based method to recognize the frauds in credit cards. In processing the credit cards transactions the tail used (the client’s past behaviors) including information regarding the sum of money, the temporal distance from the last purchase, day, week and so forth are available to the card issuer. Any deviations from the current existing tail (the studied transaction) can be calculated from the tails juxtapositions. The processing operation is very time-consuming due to the individual tails comparison and correspondence. This is meanwhile there has not been paid attention to the imitation of the original card owner’s behavior by the fraudulent individual at the time of performing the fraud which is considered as one of the tricks by the fraudulent individuals [15]. In 2011, Ogwueleka et al designed an applied
artificial neural network program for clustering, and such a program can make use of a large volume of the transactional data. In the aforementioned study, four clusters with high, moderate, low and mild risks have been used and it is in this way that the processed transactions will be placed in one such clusters, and if the transaction is a suspicious one it is returned to the database. The superiority of the neural networks in contrast to the other methods is that it can learn from the past transactions and it produces more improved results through the passage of time [19].

3. Study findings
Data mining method has been made up of various sciences such as machine learning, database and statistics. Machine-learning methods have been divided into two methods of supervised and unsupervised and the supervised method cannot be taken into practical use because it makes use of unlabeled data (the fraudulent and non-fraudulent data have not been separated and distinguished), so here we make use of unsupervised (hierarchical clustering). In the present study, database consisting of different individuals’ bank cards transactional information in the city of Tehran have been taken into consideration ad there has been made use of the single-linkage hierarchical clustering method for the columns which are in the form of a sum of money to discover the data suspicious of fraudulent activities (in this method the outlier data are per se put in a single cluster). It is worth mentioning that in the present dissertation we made use of the unsupervised data, but in the other researches performed (outside Iran and foreign articles) they have made use of supervised techniques (such as neural networks, decision trees) due to having access to the labeled data (the fraudulent data and normal data have been distinguished). The data information bank has been used to identify the data suspicious of fraud and this was made possible through the selection of the best and the most important feature in the data. The main study data have been extracted from the bank cards transactions recorded in the database of an internal state bank for the acquisition of which all of the ethical concerns was observed and a permit was also obtained and such data were employed to design a framework for the identification of the data suspicious of fraud in the bank cards. The collection of the data used in the present study were extracted and collected in an approximately one-year period in 2012 which reached a number of 17000 bank card transactions per individual in various hours during day and night which belonged to a state bank located in Tehran. To undertake data-mining and mount the algorithms we made use of R-software. R-software is an object-oriented mathematical programming language. At the present time, this language is protected and maintained by an international group and they voluntarily develop it.

4. Study findings:
4.1. Clustering:
In this section it has been attempted to better explain the single-link clustering implementation method in an example assuming the availability of 6 sample data and their distance matrix which has been given in the following table.

Table 1: the distance matrix between the 6 sample data
At first, each data has been considered as a cluster and finding the nearest cluster is in fact finding the smallest distance between the above data. According to the above table it was indicated that the data 3 and 5 have the smallest distance; as a result they have been merged and a new cluster emerges the distance of which from the other clusters is equal to the smallest distance between 3 and/or 5 from the other clusters. The results have been tabulated as the table below.

Table 2: the distance matrix between 5 clusters resulting from the first iteration

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>4</td>
<td>13</td>
<td>24</td>
<td>12</td>
<td>8</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>22</td>
<td>11</td>
<td>10</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>7</td>
<td>3</td>
<td>9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>6</td>
<td>8.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>8.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

According to the above table, it is evident that the data 1 and 2 possess the smallest distance; and therefore they are merged together and a new cluster has been obtained the distance of which from the other clusters is equal to the smallest distance between 1 and 2 from the other clusters. The result has been shown in the table below.

Table 3: the distance matrix between 4 clusters obtained from the second iteration

<table>
<thead>
<tr>
<th></th>
<th>1 and 2</th>
<th>(3 and 5)</th>
<th>4</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1 and 2)</td>
<td>0</td>
<td>10</td>
<td>22</td>
<td>8</td>
</tr>
<tr>
<td>(3 and 5)</td>
<td>0</td>
<td>6</td>
<td>8.5</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>18</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

According to the above table, it is clear that the clusters (3 and 5) and 4 have the smallest distance; and therefore they are merged together and a new cluster is obtained the distance of which from the other clusters is equal to the smallest distance between (3 and 5) and/or 4 from the other clusters. The results have been indicated in the table below.

Table 4: the distance matrix between the 3 clusters obtained from the third iteration
According to the table (4), it has been made clear that the clusters (1 and 2) and (6) have the lowest distance; and therefore they are merged with one another and a new cluster has been obtained the distance of which from the other clusters is equal to the smallest distance between (1 and 2) and/or 6 from the rest of the clusters. The results have been given as illustrated in table 5.

<table>
<thead>
<tr>
<th></th>
<th>(1 and 2)</th>
<th>(3, 4 and 5)</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1 and 2)</td>
<td>0</td>
<td>10</td>
<td>8</td>
</tr>
<tr>
<td>(3, 4 and 5)</td>
<td>0</td>
<td>8.5</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td>0</td>
</tr>
</tbody>
</table>

Table 5: the distance matrix between 2 clusters obtained from the fourth iteration

Finally, these two obtained clusters are being merged and the results have been illustrated as in the dendogram presented in figure (2).

Figure 2: dendogram example

<table>
<thead>
<tr>
<th></th>
<th>(1, 2 and 6)</th>
<th>(3, 4 and 5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1, 2 and 6)</td>
<td>0</td>
<td>8.5</td>
</tr>
<tr>
<td>(3, 4 and 5)</td>
<td></td>
<td>0</td>
</tr>
</tbody>
</table>

Regarding the bank data, due to their confidentiality, the banks do not provide the fraudulent and non-fraudulent data in separation and the data which have been made available are a combination of the fraudulent and non-fraudulent data and due to the same reason we have made use of unsupervised method for the identification of the data suspicious of fraud. Here, we take advantage of the columns which are not void and are therefore efficient to reach a conclusion. Until there is new data (fraudulent and non-fraudulent) as input we are enabled to identify the
data suspicious of fraud by making use of the unsupervised (clustering) method. In such cases the supervised (neural network) cannot be used to identify the data suspicious of fraud since the neural network can be trained. The following table presents the recorded columns of the bank card transactions which have been used in the current study.

Table 6: the general description of the data bank in one file

<table>
<thead>
<tr>
<th>Row</th>
<th>Field (column)</th>
<th>Explanations</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>AMOUNT</td>
<td>Sum</td>
</tr>
<tr>
<td>2</td>
<td>BALANCE</td>
<td>Remainder</td>
</tr>
<tr>
<td>3</td>
<td>TERMINAL_KEY</td>
<td>ATM no. or the card reader</td>
</tr>
<tr>
<td>4</td>
<td>ACTION_TYPE</td>
<td>Transaction type</td>
</tr>
<tr>
<td>5</td>
<td>SRC_CARD_NUM</td>
<td>Source card no.</td>
</tr>
<tr>
<td>6</td>
<td>DST_CARD_NUM</td>
<td>Destination card no.</td>
</tr>
<tr>
<td>7</td>
<td>SRC_CARD_KEY</td>
<td>Source card key</td>
</tr>
<tr>
<td>8</td>
<td>DST_CARD_KEY</td>
<td>Destination card key</td>
</tr>
<tr>
<td>9</td>
<td>SRC_DPST_KEY</td>
<td>Source deposit key</td>
</tr>
<tr>
<td>10</td>
<td>DST_DPST_KEY</td>
<td>Destination deposit key</td>
</tr>
<tr>
<td>11</td>
<td>SRC_ACNT_KEY</td>
<td>Source card key</td>
</tr>
<tr>
<td>12</td>
<td>DST_ACNT_KEY</td>
<td>Destination card key</td>
</tr>
<tr>
<td>13</td>
<td>VOUCHER_NUM</td>
<td>Voucher no.</td>
</tr>
<tr>
<td>14</td>
<td>VOUCHER_DATE</td>
<td>Voucher date</td>
</tr>
<tr>
<td>15</td>
<td>BRANCH_COD</td>
<td>Branch code</td>
</tr>
<tr>
<td>16</td>
<td>REVERSE_TYPE</td>
<td>Type</td>
</tr>
<tr>
<td>17</td>
<td>RESPONSE_COD</td>
<td>Response code</td>
</tr>
<tr>
<td>18</td>
<td>IS_SWIFT</td>
<td>Swift</td>
</tr>
<tr>
<td>19</td>
<td>EFFECTIVE_DATE</td>
<td>Transaction duration</td>
</tr>
<tr>
<td>20</td>
<td>TIME_KEY</td>
<td>Time</td>
</tr>
<tr>
<td>21</td>
<td>PREV_TR_DISTANCE</td>
<td>Transaction previous distance</td>
</tr>
<tr>
<td>22</td>
<td>FEE_AMOUNT</td>
<td>Amount</td>
</tr>
<tr>
<td>23</td>
<td>TERM_TYPE</td>
<td>Terminal type (ATM, Card reader, internet)</td>
</tr>
</tbody>
</table>

Table 7: the numeral sample of the transactions carried out in card number 22.

| AMOUNT | BALANCE | TERMINAL_KEY | ACTION_TYPE | SRC_CARD_NUM | DST_CARD_NUM | SRC_CARD_KEY | DST_CARD_KEY | SRC_DPST_KEY | DST_DPST_KEY | SRC_ACNT_KEY | DST_ACNT_KEY | VOUCHER_NUM | VOUCHER_DATE | BRANCH_COD | REVERSE_TYPE | RESPONSE_COD | IS_SWIFT | EFFECTIVE_DATE | TIME_KEY | PREV_TR_DISTANCE | FEE_AMOUNT |
|--------|---------|--------------|-------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|-------------|-------------|------------|-------------|-------------|-----------|----------------|----------|----------------|------------|----------------|-----------|

http://www.ijhcs.com/index.php/ijhcs/index
4.2. The presentation of the proposed method:

The general implementation methods based on which the present study procedure has been progressed are as follows:

1. Inputting the data to the software in a separate Excel file and reading the amount
2. Running the data-mining algorithms (by the use of R-software)

The input data are the real data belonging to one of the state banks in the city of Tehran and each card contains several records (performed transactions) and 23 fields which include amount, remainder, ATM No., source Card no., destination card no., transaction time, transaction type. The two methods of histogram and hierarchical clustering (single-link) have been taken...
advantage of to identify the line suspicious of fraud. To identify the data suspicious of fraud the columns 4, 23 and 3 have been used respectively to define the transaction type, terminal type and ATM no. or card reader no in histogram method and columns 1, 2 and 14 have been used to identify amount, remainder and voucher data, respectively, in hierarchical clustering (single link) method.

For example, the above algorithms have been run on card no. 22 and on the aforementioned columns and the card no.22 has been found to have 200 bank transactions

4.2.1. The use of hierarchical clustering algorithms (single-link):
At first, we implement it on the first column of the above algorithm:
Figure 3: the output resulting from the hierarchical clustering algorithm run on card no.22 for the first column

According to the specificity of the hierarchical clustering algorithm (single link) in which the outlier data are per se put in one cluster, we can figure out from the above dendogram that the line 97 contains false or suspicious data and if it is found that the other columns also have suspicious data in the same line then it can be claimed that the line no. 97 has suspicious data and its number is 23000000. The hierarchical clustering algorithm now is run on the second column.
Figure 4: the output resulting from the hierarchical clustering algorithm run on card no.22 for the second line
According to the specificity of the hierarchical clustering algorithm (single link) in which the outlier data are themselves put in one single cluster, it can be perceived from the above dendogram diagram that the lines 76, 77, 78 and 82 are suspicious of having false data and if it is found that the other columns also have the same suspicious data on the same lines the these lines are classified as containing fraudulent data. Then, the above algorithm is run on column 14.

Figure 4: the output resulting from the hierarchical clustering algorithm (single-link) run on card no.22 for column 14
According to the specificity of the hierarchical clustering (single-link) algorithm in which the outlier data are themselves put in a single cluster, it can be understood from the above dendogram diagram that the lines nos. 32, 42 and 47 contain suspicious data and if the other lines are found to have suspicious data on the same lines these lines are classified as the lines with fraudulent data.

**Conclusion**

According to the results obtained from the implementation of the aforementioned algorithm on the various columns it can be stated that the card no.22 contains data suspicious of fraud on the line no. 82 for the columns 4 and 2.

For the other cards the same algorithm will be run on them and if there is not found any commonalities in the lines containing suspicious data then it can be said that the card does not seem to have data suspicious of fraud and if the above-mentioned algorithm is run and then shared and common lines are found they are to be named as the lines having data suspicious of fraud. In the supervised techniques method the precision and exactness is an issue, since it is clear that which data are fraudulent and which are not, so the data which have been correctly distinguished as fraudulent and non-fraudulent can be taken into consideration for the calculation of the precision and the method precision can be computed but in the method employed in the current thesis paper because the data used are not appropriate to be used with supervised method (data with fraud or normal labels), the banks in Iran do not provide the labeled data even for the companies which write fraud-detection software and programs for them, therefore we were compelled to make use of unsupervised clustering method. The comparison of the works performed outside Iran due to having access to the labeled data (the fraudulent and normal data have been distinguished) supervised techniques (such as neural network, decision tree) have been taken advantage of, but the implementation method in the current dissertation due to having unlabeled data (fraudulent and normal data have not been distinguished) we were forced to make use of the unsupervised method.

**Discussion**

The result obtained from the current study has been the identification of the data suspicious of fraud in bank cards. Because in Iran the available data for the identification of fraudulent data are not clear as to whether they are fraudulent or non-fraudulent, therefore we made use of unsupervised data-mining methods (single-link hierarchical clustering). The collection of the data taken into use for this purpose pertained to the transactions carried out by credit cards for a number of individuals in 2013 in one of the state banks in the city of Tehran. Every card was found to have several records (the transactions conducted) and 23 fields which are: amount, remainder, ATM no., source card no., destination card no., transaction time, transaction type and so forth. To recognize the data suspicious of fraud we made use of histogram method on the columns 4, 23 and 3 which denoted transaction type, terminal type and ATM or card reader no., respectively, and to identify the data suspicious of fraud we made use of the single-link hierarchical clustering method on the columns 1, 2 and 14 which denoted amount, remainder and
voucher data, respectively. The steps of analysis and implementation were performed by taking advantage of R-software.

And the data suspicious of fraud in the credit cards were identified by making use of these two methods. Unsupervised data-mining technique provided better results respective to the other algorithms, and due to the bank data confidentiality, the data which were made available to be used in the identification of the data suspicious of fraud were unlabeled data (fraudulent and non-fraudulent together). This model will easily identify the fraudulent data in bank transactions through connecting to the electronic banking system. The objective of the current study was to figure out the suspicious cases among the great many of the available data which was successfully accomplished and then the bank supervisors are required to only scrutinize the suspicious cases with a higher precision and they are not to survey the entire performed bank transactions which are too many. The fraud-detection software act more or less the same way and in this manner the bank supervisors’ space has been limited to the investigation of the suspicious cases the number of which is very low in comparison to the entire myriads of the transactions.

Suggestions
With the data similar to the ones used in the current dissertation the followings can be suggested as future works:

1. Clustering the clients and the comparison of the suspicious change of behavior in a client in respect to the other customers in a cluster (whether such a change in the behavior is deemed as normal by the other client, if so then it is not suspicious).
2. The use of outlier data recognition models based on several simultaneous fields (that is, the change in the behavior based on several features all at once). In the method proposed by the present dissertation paper the models work with one field (column) at a time.
3. We can make use of iterative patterns techniques, such as the iterative orders such as “this client mostly performs purchasing transaction on Wednesday mornings and by means of a card reader with the code no. 100”.
4. In line with the identification of the online fraudulent transactions there is a lot of research work that can be performed and in such a case the techniques which are applied should be very fast and the data-mining models memory usage should be low.
References


