Computer Science Journal of Moldova, vol.**28**, no.1(82), 2020

Process Modeling and Extraction of Patterns of Computer Crimes Using Data Mining

Abbas Karimi, Saber Abbasabadei, Javad Akbari Torkestani, Faraneh Zarafshan

Abstract

The main purpose of this research is to model the process and extract patterns of computer crime using data mining and employing MATLAB software for data modeling. The results have been simulated and presented graphically. The simulation results show that this system can be considered as one of the most effective and lowest cost ways to identify the cyber-criminal behavior, therefore, computer crime experts can run effectively this model on their systems.

Keywords: Computer Crimes, Data Mining, Neural Network.

1 Introduction

Nowadays, the cybercrime has become a global problem due to the progress of information and communication technology. Two factors have been identified as the main variables to predict the rate of cybercrime: the rate of computer use and membership in social networks. Computer and electronic crimes, depending on its nature are divided into three categories including: hacking, phishing and identity fraud. Hacking means an attempt to exploit a computer system or a private network inside a computer; phishing is the fraudulent attempt to obtain sensitive information such as usernames, passwords and credit card details by disguising oneself as a trustworthy entity in an electronic communication; identity fraud is the use by one person of another person's

personal information, without authorization and making huge profits from it [2]. Cybercrime analysis can be performed on the computer networks in different conditions. Analysis of cybercrime reports can be carried out in different conditions to identify quickly the offenders by police and preventing those by the detention of offenders and their mobile phones [3]. Abnormal traffic conditions in computer networks may mean that a computer has been hacked, and the sensitive information has been sent to an unauthorized destination, or existence of abnormalities in transactions data of credit cards may cause to hack the identifications and credit card. These problems can be solved by various methods of analyzing such crimes [4]. In the era of information and communication technology, many organizations, companies, etc. use databases for their commercial, educational and statistical affairs to make decisions and providing different reports for their managers, planners and researchers. These organizations use data mining (the quick and accurate discovery of information) for their scientific, technical and economic development. New techniques for data mining such as clustering, classification etc. can be used in various educational fields or other cases. Also, decision makers of the organizations employ the high-level data mining to introduce and analyze data. To develop the classification of computer crimes, one can use the models which are controlled by the machine learning algorithms such as the neural network, and these algorithms are applied in various processes of data mining, such as text mining [5]. Compression between the text mining process and other text processing techniques shows that the data mining algorithms are used for activities such as data collecting, preparing and extracting the knowledge and words required, improving the business value level, facilitating decision making process and also cost reduction. The data mining method can be used to identify the relationships among complicated data related to internet theft or violations of cyberspace laws using the existing databases and data mining algorithms. In the crime areas, the rate of crimes can be anticipated and prevented by more precise monitoring using the data mining algorithms. Generally, there is a functional category for implementing works in the field of crimes identification and their forecasting and

preventing in this framework that are distinguished in terms of the application of data mining techniques. The data mining algorithms of crime analysis are used for creating a model from databases related to the crimes in this way. To create a model, at first, an algorithm must analyze a set of data to find a specific pattern and processes of implementing them; then one can use the results of this analysis by defining the characteristics of the extraction models. Finally, we need to know how to extract the patterns of computer crimes' occurrence using the data mining techniques.

2 Literature

Sohrabi et al. stated that semi-supervised learning method is a proper performance that is used based on data mining algorithms such as support-vector machine for predicting the type of crime [6]. Javideh et al. expressed that it is assumed that the internet theft is lesser between the provinces; the information thefts were separately investigated in different provinces and then the outputs improved [7]. During a survey on computer crimes, Ghayom et al., found out that data mining has positive and negative aspects to explore the techniques of detecting computer crimes [8]. Caneppele et al. argue that internet development has changed the lifestyle and everyday activities of people and the violations have increased in the crime prevention strategies. The new process leads to consolidation of private security; and companies involve indirectly to collecting general data that can be used to study the computer crimes [9].

3 Materials and Methods

Systems that classify based on text mining have two input sets. The first one is a training set in which the data are in the specialized categories as a default and enter into the system with their classification structure; the system is trained based on them or some characteristics are provided by selecting and extracting them from the system.

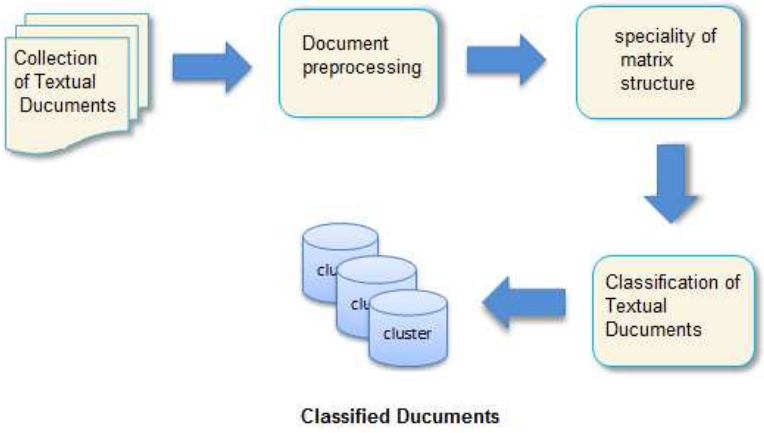

Figure 1. The process of text mining

The other set is the inputs that are entered into the system after the training stage to determine the category. Classification leads to the identification of characteristics that specify which group each case belongs to. This model can be used to understand existing words and to predict how any new model works. It is possible to classify specific words and keywords that are found through text mining to identify the extracted computer crimes by utilizing neural networks of Multilayer Perceptron (MLP) to obtain accurate conclusions from the patterns of these crimes, and deal with them. Support Vector Machine (SVM) can also provide accurate results. Finally, the extracted data or words from these methods are implemented with MATLAB software and the results are simulated and presented as a graph.

For finding the link between words and indexed documents and keywords (w) in the set of words, each word is associated with a subset, and each pair of words (W, W) is known as a rule of relationship. With this rule, for extracting the interested keywords from a set of words, first of all, the text will be read; the extra words deleted after processing and then the total number of document words will be counted one by one.

The number of the remained words is also calculated after the removal of extra words to obtain the number of repetitions of each word. Then, the frequency value of the expression is calculated as follows:

$$TF = \frac{T_w}{\sum w}. \tag{1}$$

In this equation, T_w is the number of times a word is used in the set of words and $\sum w$ is the total number of words in a document. According to the similarity rank of each word, they are put together in the same way as the following:

$$WS(w_i, w_j) = \frac{S(w_i, w_j) - \min S}{\max S - \min S}. \tag{2}$$

In this equation, $WS(w_i, w_j)$ indicates weighted similarity of w_j and w_i, $S(w_i, w_j)$ is the similarity number between the two words, $minS$ shows the smallest similarity number between all of the similar word pairs, $maxS$ is the largest similarity number between all the pair words. Therefore, this equation is computed for all of similar word pairs. At the next level, the reverse similarity rank will be computed to validation of the importance of each word. The Extra words are deleted and all computations are done again. The effect of a deleted word is defined by calculating the difference between the similarity of n and $n-1$ words; its similarity rank is as follows:

$$OS = OS(n-1) - OS(n). \tag{3}$$

In this equation, OS indicates the overall similarity of words, $n-1$ shows the previous word pair, and n is the current word pair.

4 Classification through neural network

Neural networks have a less classification error rate than decision trees, but they also require more time to learn. The main problems in this network are the classification rules that are learned by a few ways in a set of educational data because the learning or training of the

neural networks require a long time for obtaining high classification accuracy. The active amount of nodes in the hidden layer is calculated by transferring the sum of the weight of the input values in nonlinear active function. w_i^m allows the weights to connect the input node to the hidden node m. The input pattern is $x^i, i \in 1, 2, ..., k$ in which k is the number of word pairs in a set of words.

$$g_\theta(A) = diag(\theta_1\lambda_1, \ldots \theta_m\lambda_n), \theta \in R^n, \tag{4}$$

where $f(.)$ is an active function

$$f(x) : \delta(x) = \frac{e^x - e^{-x}}{e^x - e^{-x}}. \tag{5}$$

Active function of hidden nodes creates an active value range of hidden nodes. When the active value of all hidden nodes is computed, the ν_p^m output calculation of the network for the input pair x_i is defined as follows:

$$S_p^i = \sigma(\sum_{m=1}^{h} a^m \nu_p^m). \tag{6}$$

That ν_p^m is the connection weight between hidden node m and output node p, and h is the number of hidden nodes in the network.

In the neural network, the first input is multiplied by the relevant weight factor to communication line of that input. Then the same procedure is repeated for the second input and other ones. Finally, all the resulted values are added together in the target function:

$$\sum_{i=1}^{n} w_i x_i = w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4 + ... + w_nx_n. \tag{7}$$

The sum of the above values must be compared with the threshold value of the interested neuron. In comparing the threshold, if the obtained sum is more than the threshold value, then the output of neurons would be 1, and if it is less than the threshold value, it would be zero.

Table 1. The values obtained for 5 neurons

Test Error	Learning Error	Type of hidden layer membership functions	Type of Input membership functions	Layers row number	
37.605	40.504	-	tan sig	2	1
8.790	10.604	Tan sig	Tan sig	3	2*
38.489	3.210	Tan sig	Tan sig	4	3
4.100	20.918	Tan sig	Tan sig	5	4

4.1 Simulation result for tan-sig membership function

The result of tan-sig membership function is shown in Table 2.

As it is shown in Table 1, the interested result (the lowest value for test and learning error) that was expected for the number of 5-neurons with constant membership function is obtained for a number of 3-layers. Figures of network view, performance and regression of network learning are shown in Figure 2, and charts are represented in Figures 3 and 4 respectively, and the corresponding row in Table 1 is shown by (*).

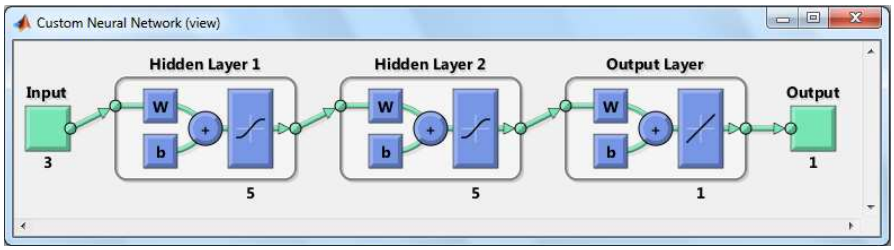

Figure 2. Network view corresponding to the second row of Table 1

As it can be seen in the chart in Figure 3, the blue line represents the network learning, the green line is related to validation and the red line shows the network test. The software uses validation performance for the early stop to avoid over- fitting of the neural network. The next step in network validation is to create a regression chart that shows

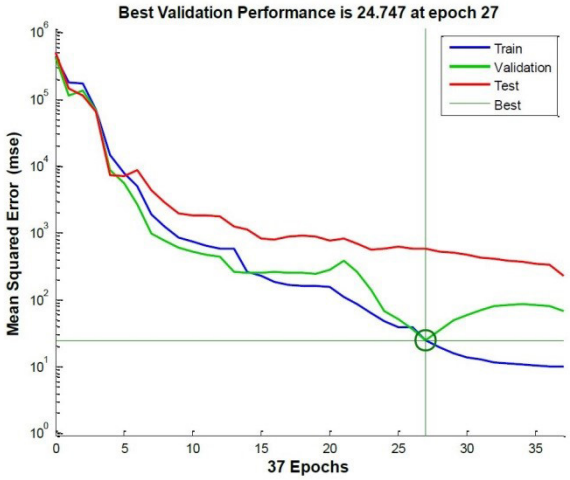

Figure 3. Network performance for the best output of Table 1 for the learning steps

the relationship between network outputs and goals. If the learning becomes complete, the network outputs and goals would be exactly equal, but the relation between these two would be rarely completed in learning phase. Three axes with charts in Figure 4 provide learning data, validation and test. The cutoff line in each axis shows the best result (output=goals). The continuous line represents the best linear fit of regression between the outputs and goals. The value of R shows the relation between the outputs and goals. If R = 1, it indicates the exact linear relationship between the outputs and goals. If R is close to zero, then there will be no linear relationship between the outputs and goals. In this figure learning data and validation are compatible. The results of the test also indicate that the values of R are greater than 0.9.

Table 2 shows the best result for the number of 7 neurons with constant membership function for the obtained number of 3 layers, and the performance chart for the network learning is shown in Figure 4.

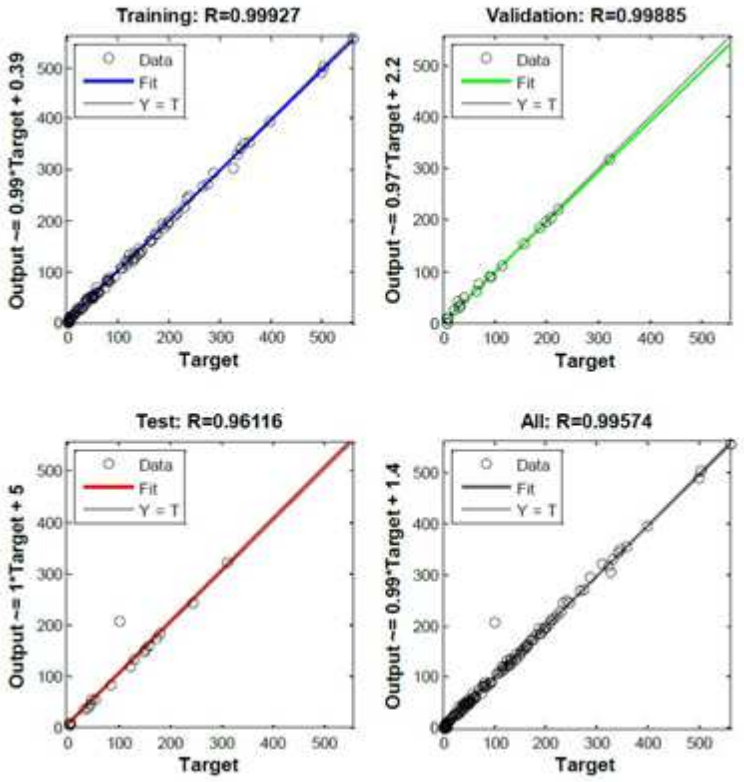

Figure 4. The network performance for the best output of Table 1 for the learning steps

Table 2. The values obtained for 7 neurons

Test Error	Learning Error	Type of hidden layer membership functions	Type of Input membership functions	Layers number	row
10.203	15.942	-	Tan sig	2	1
0.982	1.780	Tan sig	Tan sig	3	2*
137.215	61.536	Tan sig	Tan sig	4	3
50.160	119.480	Tan sig	Tan sig	5	4

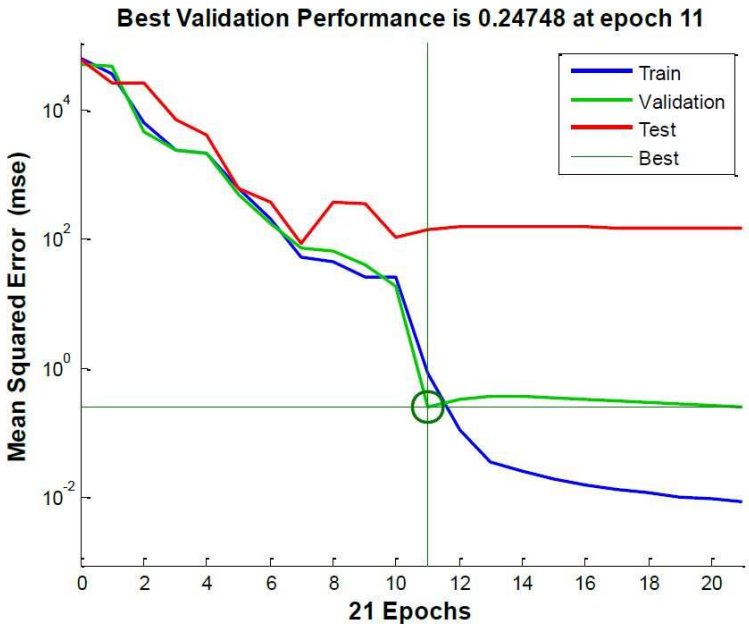

Figure 5. Network performance for the best output of Table 2 for the learning steps

Figure 3 represents the best validation performance that is equal to 0.248 in the 11th period, which is an appropriate value, and it is also clear that the training process has stopped in the 21st period. In Figure 5, R has a high value, and among the four axes, the R value of the test mode is lower than the others.

5 Results and comparing them with those of other researches

In the previous studies, various methods such as conducting a survey on learning with monitoring and learning techniques about criminal

identification and the use of machine learning methods, including stepwise regression, penalty method and random forests were used to create predictive models of violent crimes, and data mining was used to investigate and detect crimes. In the present study, in order to investigate and detect crimes, the intelligent methods of the neural network and modeling in MATLAB software have been used. The results of this study show that the gap between the theory and the implementation must be reduced, especially in the police area, in order to use artificial neural networks, where cybercrime is directly related to the increase of the crime in society. Cybercrimes lead to increase of the financial, psychological, cultural, social, political, and security damages.

This research can be very valuable and practical, since it has described the applied evidences using artificial intelligence in the field of cybercrime. The method used in this study, in addition to identifying crimes, saves lives, property, privacy, and security of human beings by identifying and detecting attacks and cybercrimes as well as reducing the possibility of committing such crimes by the offenders and consequently, reducing the damage of these violations. The used method in this research is most effective because of a low cost and has an optimal output.

6 Findings and Suggestions

1. It will be possible to implement this research as a software package for cybercrime detection systems in the future.

2. This research can be a background to design and produce online cybercrime recognition software in the future and can be used in the same way for other software that is used in cyber police centers employing artificial neural networks and special development programming techniques for detection of cybercrime in police stations.

3. The information and documentation of this research on cybercrime and analyzing it from various aspects is the basis of future

works of researchers in the law and psychology sciences on cyber criminology.

4. Using the results of this study, one can conclude a combination of which types of the cybercrime cause more cost and damage, and as a result, the legislator can punish more those who commit a combination of several specific violations.

5. Server data and systems of some organizations and corporations and cybercrime cases have been studied and documented. The results showed that:

   - Most cyber-attacks in organizations involve infiltrating office websites and intranet attacks
   - The largest cluster in terms of crime similarities is a cluster, which is used by criminals for their criminal acts using phishing techniques.

6. This system can be considered as one of the most effective and lowest cost ways to identify the cyber-criminal behavior, therefore, computer crime experts can run effectively this model on their systems.

7. The statistical and social studies of this research can be useful for scholars and researchers in this field. In another word, they can use these data in writing their future theses and articles on the psychology of the cybercrime.

References

[1] M. Delshad, B. Tudeh Za'im, and E. Rastegar, "Computer crime analysis using artificial intelligence methods and data mining for pre-trial crime," *Computer and Information Technology*, vol. 24, pp. 12, 2018. (in Farsi)

[2] M. Jantani, "Investigation and identification of the proposed algorithm in the case of electronic crimes on the website of the national court of justice computation," pp. 5–77, 2018. (in Farsi)

[3] D. K. Tayal, A. Jain, S. Arora, S. Agarwal, T. Gupta, and N. Tyagi, "Crime detection and criminal identification in india using data mining techniques," *AI & SOCIETY*, vol. 30, no. 1, pp. 117–127, apr 2014. (in Farsi)

[4] Shiju Sathyadevan, Devan M. S, and Surya Gangadharan S., "Crime analysis and prediction using data mining," in *2014 First International Conference on Networks & Soft Computing (ICNSC2014)*, (Guntur, India), IEEE, aug 2014, pp. 406–412. DOI: 10.1109/cnsc.2014.6906719. Available: https://doi.org/10.1109%2Fcnsc.2014.6906719. (in Farsi)

[5] S. M. A. M. Gadal and R. A. Mokhtar, "Anomaly detection approach using hybrid algorithm of data mining technique," in *2017 International Conference on Communication, Control, Computing and Electronics Engineering (ICCCCEE)*, (Khartoum), IEEE, jan 2017, pp. 1–6. DOI: 10.1109/iccccee.2017.7867661. Available: https://doi.org/10.1109%2Ficccccee.2017.7867661. (in Farsi)

[6] B. Sohrabi, I. R. Vanani, and E. Abedin, "Human resources management and information systems trend analysis using text clustering," *International Journal of Human Capital and Information Technology Professionals*, vol. 9, no. 3, pp. 1–24, jul 2018. DOI: 10.4018/ijhcitp.2018070101. Available: https://doi.org/10.4018%2Fijhcitp.2018070101. (in Farsi)

[7] B. M. Mustafa Javideh, Einollah Khanjari, "Suggesting models for intelligent identification of burglars using local and behavioral information," in *International Congress on Innovation in Engineering and Technology Development*, 2017, pp. 6–7. (in Farsi)

[8] M. Ghayom, B. Pes, and S. Serusi,"Data mining for detecting bitcoin ponzi schemes," in *2018 Crypto Valley Conference on*

Blockchain Technology (CVCBT). IEEE, 2018, pp. 75–84. (in Farsi)

[9] S. Caneppele and M. F. Aebi, "Crime drop or police recording flop? on the relationship between the decrease of offline crime and the increase of online and hybrid crimes," *Policing: A Journal of Policy and Practice*, vol. 13, no. 1, pp. 66–79, sep 2017. DOI: 10.1093/police/pax055. Available: https://doi.org/10.1093%2Fpolice%2Fpax055. (in Farsi)

Abbas Karimi, Saber Abbasabadi,
Javad Akbari Torkestani, Frane Zarafshan

Received June 16, 2019
Revised February 15, 2020

Abbas Karimi
Assistant Professor, Department of Computer Engineering,
Arak Branch, Islamic Azad University, University, Arak, Iran.
E–mail: saber.abbasabadi1398@gmail.com

Saber Abbasabadei
PhD student, Department of Computer Engineering,
Arak Branch, Islamic Azad University, University, Arak, Iran.
E–mail: saber.abbasabadey1@gmail.com

Javad Akbari Torkestani
Associate Professor, Department of Computer Engineering,
Arak Branch, Islamic Azad University, University, Arak, Iran.
E–mail: j-akbari@iau-arak.ac.ir

Faradeh Zarafshan
Assistant Professor, Department of Computer Engineering,
Ashtian Branch, Islamic Azad University, Ashtian, Iran.
E–mail: fzarafshan@aiau.ac.ir