AMSE JOURNALS-AMSE IIETA publication-2017-Series: Advances B; Vol. 60; N°2; pp 318-330
Submitted Feb. 28, 2017; Revised May 17, 2017; Accepted Jun. 12, 2017

A IDS Model Based on HGA and Data Mining

*Lina Lin, **Dezhi Wei, ***Fuji Chen

*Jimei University Chengyi College, Xiamen 361021, China (linda_839@126.com)
**Jimei University Chengyi College, Xiamen 361021, China (weidezhi@163.com)
***School of Economics and Management, Fuzhou University, Fuzhou 350116, China (chenfuji@fzu.edu.cn)

Abstract

The paper proposes a IDS that is based on HGA and Data mining. In this model, an improved clustering algorithm is introduced to classify the normal/abnormal behaviour library from behaviour records on the network and in the system. Then it takes the HGA and data mining as a basis to dig out the the invasion rules and put them into the rule base. Finally, Hybrid Detection Module is proposed to detect the intrusion system. The experiment shows that with a high adaptability, the model has enabled to detect unknown intrusion, improve the detection rate and reduce the false detection rate, thus to protect the computer systems from exotic intrusion.

Key words

Data mining, Intrusion detection, HGA, Clustering algorithm, Information gain.

1. Introduction

With the rapid development of the Internet, the network security has become a hotspot issue in the world; Intrusion detection has then emerged to protect computer and network from malicious attack, which has been increasingly prevalent. Many scholars have conducted a lot of meaningful research in this field. A Aaptive Model Generation, as a framework for intrusion detection, is enabled to collect data and build models on the fly. This model has addressed many

challenges by data-mining based intrusion detection system, such as how to mine the rules of normal and abnormal behaviours from a great mass of initialized data and how to detect and update the mined rules automatically and efficiently.

In Columbia University, Stolfo and WenkeLee first integrated the data mining technology into the intrusion detection system. The purpose is that how to apply association rule mining and frequent pattern mining to intrusion detection [2]. Bhavani Thuraisingham discussed different threats to general security, and integrated the data mining technology to defend against these threats [3]. In the past 10 years, many scholars have concerned with the association rules, frequent pattern, clustering algorithm, classification algorithm and so on, and proposed a variety of theory models for intrusion detection. They have enjoyed some successes [4-6]. However, due to delicate adaptive ability of the intrusion detection model, it failed to detect some new or unknown intrusion variants; this implies that it shall be improved in terms of detection rate, false negative rate and false alarm rate.

Therefore, this paper proposes a data-mining based adaptive model that is used to mine potential safety information from the system and network behaviour records as acquired; the intrusion patterns are automatically extracted to update automatically the intrusion pattern base with environmental change. The improved clustering algorithm classifies the system and network behaviours as the normal and/ abnormal behaviour databases. Then the hybrid genetic algorithm excavates the normal / abnormal intrusion patterns from appropriate database. The last one is the hybrid intrusion detection framework which adopts the method with integration of misuse- and exception-based detections to improve detection rate and reduce false alarm rate.

The this paper is organized as follows. We depict the related work about our researcher in section 2, and discuss the adaptive model system for intrusion detection in detail in Section 3. Besides, we also report our experiment in section 4 and conclude our works in section 5.

2. Related Work

Intrusion detection can be divided into misuse detection and anomaly detection; many scholars engulfed themselves in their studies on anomaly detection because the miss rate remains not high in misuse detection [4].

Wang et al. (2010) proposed a neural network classification based on fuzzy clustering

analysis. The experimental results show that its detection accuracy and detection stability were higher than the decision tree and the naive Bayes and so on [7]. Ma et al (2016) also proposed an improved Dynamic Neural Network (DNN) which matched the traditional BP neural network and SVM classification algorithm. This model has been empirically proven to be very effective with higher precision [8]. Forrest et al (2007) introduced a similar artificial immune system for computer intrusion detection. The experiments results showed that it was based on Agent model to detect safety incidents [9]. Hu et al (2008) proposed a detection model based on AdaBoost algorithm that inherits the idea of decision tree classification, which had a low computational complexity and error rate [10]. Zhang et al (2008) Hachmi (2015) developed the outliers of data mining for intrusion detection [11]. Support vector machine (SVM), as a statistical model algorithm, can achieve good results with approximation of nonlinear function; therefore, many scholars have modified the support vector machine to improve its applicability. The improved SVM was used in the classification of anomaly detection, and a lot of meaningful results were obtained [13-16]. The change of network traffic timely shows the network is attacked. The unknown attack can be detected timely based on network traffic anomaly [17-19]. Aslahi-Shahri et al. (2015) proposed an intrusion detection model based on hybrid vector machine and genetic algorithm, which effectively reduced the number of features and improved the detection rate [20].

However, the recent studies show that the primary intrusion detection model, which is based on anomaly detection, exploits a variety of intelligent classification algorithm, but it has some defects to be further improved such as adaptive ability, detection rate, false negative rate and false alarm rate are lower [4-6].In this paper, we propose an adaptive IDS model which inherits the strengths of misuse detection and anomaly detection, and improves the intelligent algorithm. It is proved that the detection efficiency of this model was improved.

3. Adaptive IDS Model

3.1 Basic System Structure

The adaptive IDS model is divided into five modules, i.e. data acquisition, data mining, rule generation, and misuse / anomaly detection. The concrete structure is shown in Figure 1; for one thing, the data acquisition module oversees the target system that collects network and host activity data, and carries out data pre-processing in order to trigger a set of system and network

behaviour; Then the improved clustering algorithm is used in the data mining module to process the data set, initially distinguish abnormal / normal behaviour, and abnormal / normal behaviour database is available; the hybrid genetic algorithm (HGA) , used in the rule generation module, mines the normal / abnormal intrusion patterns from the normal / abnormal behaviour database to build a intrusion pattern base. Finally, the misuse / anomaly hybrid detection module is used to detect the intrusion in a real time.

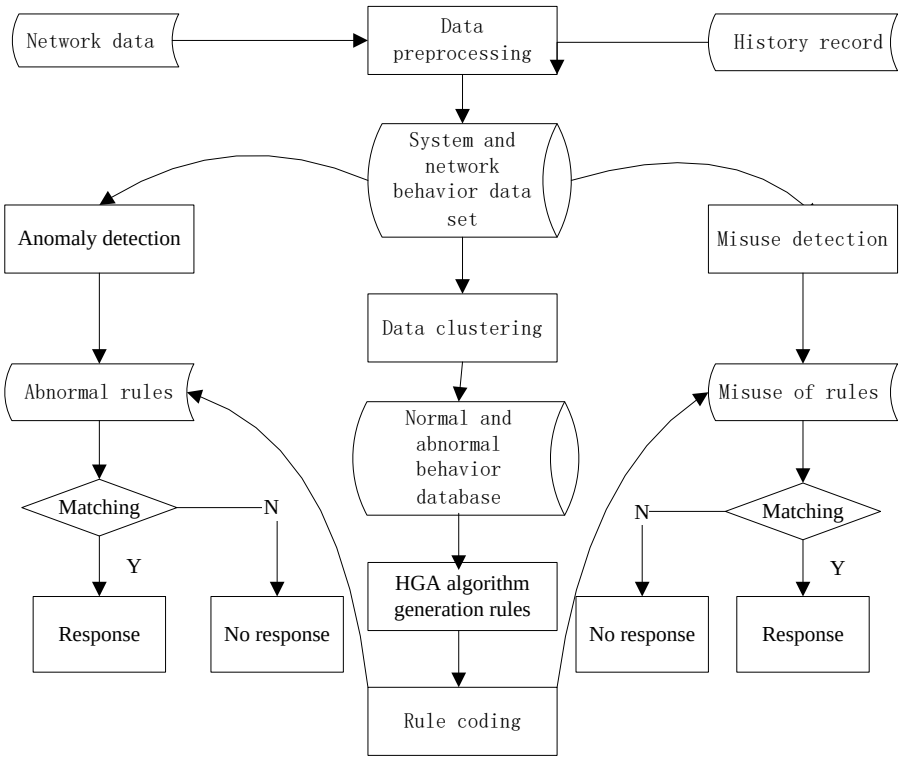

Fig.1. Structure of the Adaptive IDS Model

In the model, the intrusion pattern base is built in a process of constantly updating and improving, which involves collecting, differentiatingg and exploring behaviours as a whole. As it is an automatic process, the intrusion mode can be updated auto in the current system and under the network environment.

The hybrid detection framework is designed to integrate misuse detection and anomaly detection, in order to get a high rate in misuse detection and a high false alarm rate in anomaly detection. The model follows the method of anomaly detection after misuse detection. When the data set is detected, the misuse detection module enables to detect the attack data before the

obvious normal data starts to be detected by the anomaly detection module.

3.2 Improved Data Clustering

Both the original and current k-means algorithms using clustering criterion function feature alike. Random selection of initial clustering center causes the clustering into local optimum easily, further make each clustering result always different and unstable [4]. In this paper, an improved clustering algorithm is proposed to output the partitions of these records, based on the input of record set E with n event log objects. The algorithm does not need the final clustering number, but determined automatically by clustering. The algorithm is defined as below:

Definition 1: The record sets of system and network behaviours is labelled as E which is defined as formula 1, i.e. the sets of multiple event log objects.

$$E = \{e_1, e_2, \cdots e_n\} \tag{1}$$

Definition 2: e_i and $e_j(i \neq j)$ are any two event log objects in the Recordset E, composed of N attributes (Including Q numerical attributes), The numerical attributes of non-analog between e_i and e_j is calculated by Euclidean which is defined as formula 2.

$$\overline{sim}^{(N)}(e_i, e_j) = \sqrt{\sum_{k=1}^{q} \left| f_{ik} - f_{jk} \right|^2} \tag{2}$$

In this paper, we detect the analogs between the event log objects and the current cluster centers by scanning behaviour. If the minimum value of all non analogs is found to be greater than that among all the cluster centers, it was considered that no cluster corresponds to the current one. Then a new cluster is defined by the event log object as the clustering center and added to the clusters. Otherwise, the event log object will be aggregated to that clustering most similarly. Then the cluster center is adjusted and the next event log object continues to cluster.

We use KDD Cup99 network data set as test data, which simulates a variety of intrusions in the network environment. It includes about 4.9 million data records, each of which contains a

record of 41 dimensional features. This data set is widely used in evaluation test on IDS and data mining. Then we randomly select 12257 events records as the experimental data set from the dataset, called E1, where there are 5721 normal connection records, 6536 intrusion records, including 3560 DOS, 1032 NEPTUNE, 472 UR2L, 632 R2L, 840 PROBING.

The actual clustering effect of the improved clustering algorithm is shown in Table 1. The algorithm is proved to have a better clustering effect with high accuracy, in addition to the 3 and 4 clusters, a small number of abnormal behaviours are classified as normal ones. Other clustering number was accurate with average partition effect of 98.2%.

Tab.1. Clustering Effect of the Improved Algorithm

clustering number	Total record number	Number of normal behavior records	Number of abnormal behavior
1	1980	1980	0
2	1796	1796	0
3	1404	156	1248
4	1164	65	1099
5	1721	1721	0
6	3440	0	3440
7	532	0	532

3.3 Rules of HGA

After a set of system and network behaviours are classified into normal and abnormal behaviour databases by the improved clustering algorithm, it is necessary to extract the intrusion rule base from there. Now the algorithms used in the intrusion detection system is mainly focused on the neural network, association, sequence, classification, genetic algorithms and so on. However as the actual effect is poor, the algorithms needs to be further improved [4-6]. The HGA algorithm is proposed here; firstly the attribute with large volume of information is obtained by using the attribute correlation analysis in data mining, then the degrees of support, confidence and interest as involved in association rules are available to construct the fitness function of genetic algorithm in order to achieve an optimal solution.

The main idea of HGA algorithm is given as follows: for one thing, we get the attribute with large volume of information, generate the initial population based on these attributes, set the default parameters, and read the training data. Then the initial population evolves into several

generations, and the most appropriate rule is chosen according to the size of the fitness function in the evolution of each generation. Finally, the next generation of population is generated by GA mutation and crossover operation.

(1) Extract the attribute with bulk of information

Information gain can be used to measure mass information contained in an attribute. The greater the information gain of the attribute is, the greater the effect of the attribute is used in the classification. So the attributes with excessive gain should be preserved in the process of selecting attributes.

Suppose S is a set of training samples, the classification of a given samples presents desired information as formula 3.

$$I(s_1, s_2 \ldots s_m) = -\sum_{i=1}^{m} \frac{s_i}{s} \log_2 \frac{s_i}{s} \tag{3}$$

The attribute A with the value $\{a_1, a_2, ..., a_v\}$ can be used for the partition S $\{S_1, S_2, ..., S_v\}$. The expected information based on A is called the weighted average of A.

$$E(A) = \sum_{j=1}^{v} \frac{s_{1j} + \ldots + s_{mj}}{s} I(s_{1j} + \ldots + s_{mj}) \tag{4}$$

The information gain obtained by this partition on A is defined as formula 5.

$$Gain(A) = I(s_1 + \ldots + s_m) - E(A) \tag{5}$$

The network behaviour can be expressed by some attributes, such as source address, destination address, source port number, destination port number, protocol type and connection time length, etc. A network record often contains many attributes, many of which may play a minor role in identifying the attack. We instantiated the KDDCUP99; the data set provides 41 network attributes. If these attributes are used in intrusion detection, the complexity and difficulty of detection will be inevitably increased. The gain of each network attributes is calculated by the

attribute correlation analysis, and the attributes with mass information will be available. The complexity of the algorithm and the detection speed are all improved.

Attribute information gain value of the KDDCUP99 training data (total number of records: 494021, the number of attack records: 421110, the number of normal records: 72911) is calculated, as shown in Table 2. Sort data in the table from high to low, and only part of data is selected because of the limited space.

Tab.2. Attribute Information Gain

Serial number	Attribute name	Information gain value	Serial number	Attribute name	Information gain value
1	Count	0.384656	7	Flag	0.030943
2	Logged_in	0.330298	8	Serrorrate	0.030623
3	Srvcount	0.300408	9	Hot	0.001139
4	Service	0.258879	10	Numroot	0.001130
5	Protocol_type	0.156595	11	Numaccessfiles	0.001090
6	Duration	0.031876	12	Num_compromised	0.000685

As can be seen from the calculation results, the information content of the first 11 attributes is obviously richer than that of the other one, and the others were obviously smaller than the first 11 attributes.

(2) HGA coding

Each intrusion detection rule is coded in a "if-then" form where it concludes the conditions and results. The conditional part of the rule is constituted in conjunction of network attribute with logical symbol "AND". The “Isattack” attribute is a result that demonstrates whether it is an attack. The form of a rule is shown as below, where the string only represents the actual meaning. The value of the case will be replaced in use.

If (Duration = "ANY" and Protocol_yp e = "TCP" and Service = “telnet” and Flag = "ANY" and Logged_in = false and Count = 0 and Hot=”ANY” and Num_root= “ANY” Num_access_files = “ANY” Num_access_files = “ANY” Svr Count = 0 and Serror rate = “ANY”) then (IsAttack = “buffer_overflow”).

We allow the network attribute is represented using wildcards set to “#”. Therefore the above example is expressed as: {#, 1, 12, #, 0, 0,, #,, #,, #, 0, #, 12}. Binary encod is used to represent HGA. An binary string represents each attribute value in an proper length. Each binary string

denotes a gene. All attribute values are concatenated into a binary string as a chromosome representing an association rule.

(3) Adaptive HGA function

The task of association rule mining is to find out the association among records. These rules shall be created with a certain degrees of support, credibility and interest. We make a comprehensive evaluation on a rule from perspective of three degrees of support, credibility and interest. On this basis, the better rule can get a higher fitness value, so that they can get higher chances of survival in the competition. If a rule is expressed as: if X then Y, the fitness function of the rule is defined as.

$$S = Support(X \Rightarrow Y) = P(X \cup Y) = |X \quad and \quad Y|/N \tag{6}$$

$$C = Confidence(X \Rightarrow Y) = P(X/Y) = |X \quad and \quad Y|/|X| \tag{7}$$

The fitness function is defined as formula 8..

$$Fitness = a * S + b * c + c * I \tag{8}$$

N is the total number of training data, $|X|$ represents the number of entries in the training data that satisfiex the condition X; $|X \ and \ Y|$ represent the number of data entries according to the rule "if X then Y". "a, b, c" are used to balance the weights of the three items, and they are the constants whose ranges are: $0 \leq a, b, c \leqslant 1$. S is the degree of support, C is the degree of confidence, and I is the degree of interest. The values of a, b, and c are adjusted by the user on demand, so that the evaluation of the rules changes its core. They could evolve along the desired direction of the users. Compared to the experiment, the default value is set as follows: a=0.2, b=0.6, c=0.2.

(4) Genetic operation of HGA

Genetic operations include selection, crossover and mutation. We select the algorithm of tournament selection which allows the highest number of individuals into the next generation.

Binary uniform crossover is used for cross operation. It take the equal crossover probability for each of the two pairs of genes, thus it formed two new individuals.

4 Experimental Results

10% of KDDCup99 was used as test set and training set in the experiment. The test set consists of 494021 records, i.e. 396743 attack records and 97278 normal records. The training set consists of 311029 records, i.e. 250436 attack records and 60593 normal records. The number and distribution of each type of intrusions are shown in Table 3.

Tab.3. The Number and Distribution of the Intrusion Recodes

Intrusion method	Test set		Training set	
	Number	Proportion (%)	Number	Proportion (%)
Normal	60593	19.481	97278	19.6911
U2R	70	0.0225	52	0.0105
R2L	16347	5.2558	1126	0.2279
Probe	4166	1.3394	4107	0.8313
Dos	229853	73.9008	391458	79.2391

The detection effect of SCDNN algorithm is better than SVM, BP, RF, Bayes algorithm etc [8]. The detection effect of MSVM is better than that of SVM algorithm etc [13].Therefore, in this paper, we use the HGA, SCDNN [8] and MSVM [13] algorithms to get a set of the rules in the training data, and use the rules to test the training data and test set respectively. The results ae shown in Figure 2 and Figure 3.

TPR (True Positive Rate) represents the detection rate, as the ratio of successful intrusion detections. FPR (False Positive Rate) represents the false alarm rate, as the ratio of the intrusion data to the normal. As can be seen from the figures 2 and 3, the TPR and FPR of HGA are better than that of SCDNN and MSVM. It is proved thatHGA has a better detection effect.

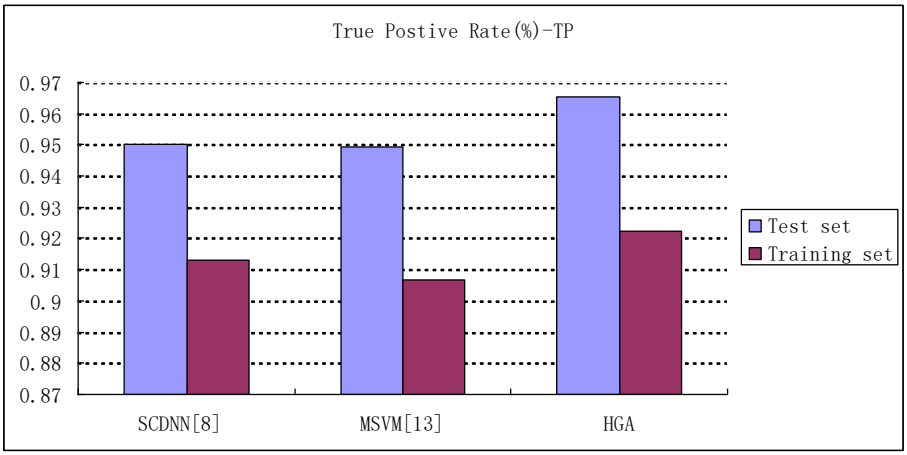

Fig.2. Column Chart of the Detection Rate

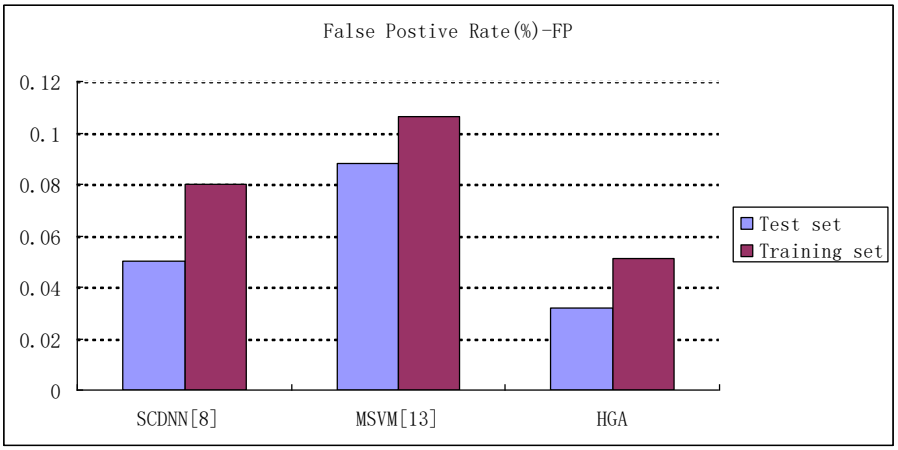

Fig.3. Column Chart of False Alarm Rate

Conclusions

Compared to the experimental results, the adaptive IDS model integrates with HGA and data mining technology with a better detection effect, where a new idea was proposed for intrusion detection. But the model currently bases itself upon the single operation. With the high-speed development of Internet, the 4G technology become increasingly popular. The increasing network traffic makes this model become the design of distributed systems, and it also leads the development direction of the next generation. Funded project from the National Natural Science Foundation of China (Grant No: 71271056), Fujian Provincial Department of education project (Grant No: C13001, JA14368).

References

1. E. Eskin, M. Miller, Z.D. Zhong, G. Yi, W.A. Lee, S. Stolfo, A daptive model generation for intrusion detection systems, 2000, In Workshop on Intrusion Detection and Prevention, 7th ACM Conference on Computer Security.
2. W. Lee, S.J. Stolfo, P.K. Chan, E. Eskin, W. Fan, M. Miller, S. Hershkop, J.X. Zhang, Real time data mining-based intrusion detection, 2001, DARPA Information Survivability Conference& Exposition, pp. 85-100.
3. B. Thuraisingham, Data mining for malicious code detection and security applications, 2009, Web Intelligence and Intelligent Agent Technologies, vol. 31, no. 2, pp. 88-100.
4. P. Garcia-Teodoro, J. Diaz-Verdejo, G. Maciá-Fernández, Anomaly-based network intrusion detection: Techniques, systems and challenges, 2009, Computers & Security, vol. 28, no. 1, pp. 18-28.
5. V. Chandola, A. Banerjee, V. Kumar, Anomaly detection: A survey, 2009, ACM computing surveys (CSUR), vol. 41, no. 3, pp. 15.
6. S.X. Wu, W. Banzhaf, The use of computational intelligence in intrusion detection systems: A review, 2010, Applied Soft Computing, vol. 10, no. 1, pp. 1-35.
7. G. Wang, J. Hao, J. Ma, A new approach to intrusion detection using Artificial Neural Networks and fuzzy clustering, 2010, Expert Systems with Applications, vol. 37, no. 9, pp. 6225-6232.
8. T. Ma, F. Wang, J. Cheng, A Hybrid Spectral Clustering and Deep Neural Network Ensemble Algorithm for Intrusion Detection in Sensor Networks, 2016, Sensors, vol. 16, no. 10, pp. 1701.
9. S. Forrest, C. Beauchemin, Computer immunology, 2007, Immunological reviews, vol. 216, no. 1, pp. 176-197.
10. W. Hu, W. Hu, S. Maybank, Adaboost-based algorithm for network intrusion detection[J]. IEEE Transactions on Systems, 2008, Man, and Cybernetics, Part B (Cybernetics), vol. 38, no. 2, pp. 577-583.
11. J. Zhang, M. Zulkernine, A. Haque, Random-forests-based network intrusion detection systems, 2008, IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 38, no. 5, pp. 649-659.

12. F. Hachmi, K. Boujenfa, M. Limam, An optimization process to identify outliers generated by intrusion detection systems, 2015, Security and Communication Networks, vol. 8, no. 18, pp. 3469-3480.
13. Y. Li, J.L. Wang, Z.H. Tian, Building lightweight intrusion detection system using wrapper-based feature selection mechanisms, 2009, Computers & Security, vol. 28, no. 6, pp. 466-475.
14. S.J. Horng, M.Y. Su, Y.H. Chen, A novel intrusion detection system based on hierarchical clustering and support vector machines, 2011, Expert systems with Applications, vol. 38, no. 1, pp. 306-313.
15. W. Feng, Q. Zhang, G. Hu, Mining network data for intrusion detection through combining SVMs with ant colony networks, 2014, Future Generation Computer Systems, vol. 37, pp. 127-140.
16. W.L. Al-Yaseen, Z.A. Othman, M.Z.A. Nazri, Multi-level hybrid support vector machine and extreme learning machine based on modified K-means for intrusion detection system, 2017, Expert Systems with Applications, vol. 67, pp. 296-303.
17. G. Giacinto, R. Perdisci, M. Del Rio, Intrusion detection in computer networks by a modular ensemble of one-class classifiers, 2008, Information Fusion, vol. 9, no. 1, pp. 69-82.
18. A. Sperotto, G. Schaffrath, R. Sadre, An overview of IP flow-based intrusion detection, 2010, IEEE communications surveys & tutorials, vol. 12, no. 3, pp. 343-356.
19. N. Paulauskas, Ą.F. Bagdonas. Local outlier factor use for the network flow anomaly detection, 2015, Security and Communication Networks, vol. 8, no. 18, pp. 4203-4212.
20. B.M. Aslahi-Shahri, R. Rahmani, M. Chizari, A hybrid method consisting of GA and SVM for intrusion detection system, 2015, Neural Computing and Applications, vol. 7, no. 27, pp. 1669–1676.