Intelligent Decision Technologies 15 (2021) 231–237
DOI 10.3233/IDT-200036
IOS Press

Intrusion detection system using a new fuzzy rule-based classification system based on genetic algorithm

Zahra Asghari Varzaneh and Marjan Kuchaki Rafsanjani*
Department of Computer Science, Faculty of Mathematics and Computer, Shahid Bahonar University of Kerman, Kerman, Iran

Abstract. Intrusion can compromise the integrity, confidentiality, or availability of a computer system. Intrusion Detection System (IDS) is a type of security software designed to monitor network traffic and identify network intrusions. In this paper, A Fuzzy Rule – Based classification system is used to detect intrusion in a computer network. In order to improve the classification rate, a new method is proposed based on Genetic Algorithm (GA) for rule weights specification. The proposed method is tested on KDD99 dataset. Experimental results show the proposed method improves the performance of the fuzzy rule-based classification systems in terms of detection rate and false alarm rate.

Keywords: Intrusion detection, fuzzy rule-based, rule weighting, genetic algorithm

1. Introduction

With the rapid growth of the Internet, various attacks on the network can pose a major threat to network and information security. Attackers attempt to attack networks in order to gain access to information resources. So intrusion detection [1,2] is a practical mechanism to handle the hackers from exploiting the data. An intrusion detection system (IDS) is a type of security software designed to automatically alert administrators when someone or something is trying to compromise information system through malicious activities or through security policy violations. There are multiple ways that detection is performed by IDS. In signature-based detection, a pattern or signature is compared to previous events to discover current threats. This is useful for finding already known threats, but does not help in finding unknown threats, variants of threats or hidden threats. Another type of detection is anomaly-based detection, which identifies malicious traffic based on deviations from recognized normal network traffic patterns. The problem of intrusion detection has received a lot of attention in machine learning and data mining [3,4]. Zamini and Hasheminejad [5] had an overview of the research on anomaly detection. They classified anomaly detection according to their application and then categorized their techniques. Furthermore, they discussed on differences among existing techniques in each specific category and described the advantages and disadvantages of each technique.

Data mining generally refers to the process of extracting useful rules from large stores of data. That is one of the technologies applied to intrusion detection. Rafsanjani and Varzaneh [6] introduced various data mining techniques used to implement an intrusion detection system. They reviewed some of the related studies focusing on data mining algorithms. Fuzzy rule-based systems have been successfully applied to various application areas. Fuzzy if-then rules are traditionally gained from human experts. Recently, various methods have been suggested for automatically generating and

*Corresponding author: Marjan Kuchaki Rafsanjani, Department of Computer Science, Faculty of Mathematics and Computer, Shahid Bahonar University of Kerman, Kerman, Iran. E-mail: kuchaki@uk.ac.ir.

ISSN 1872-4981/$35.00 © 2021 – IOS Press. All rights reserved.

adjusting fuzzy if-then rules without using the aid of human experts [7,8].

The remainder of the paper is organized as follows. Related works on intrusion detection is introduced in Section 2. In Section 3, the method used for designing FRBCS from numerical data is presented. In Section 4, the proposed method is explained. Simulation results on database KDD99Cup are given in Section 5. Finally, Section 6 gives concluding remarks.

1.1. Contributions of our paper

In recent years, finding an effective and suitable model in the field of intrusion detection has been considered as an important issue. Due to the importance of intrusion detection and considering the fact that data mining is one of the practical technologies which proposes a new pattern from data of mass networks, researchers have been focused on hybrid algorithms, fuzzy techniques, neural networks, genetic algorithm, and etc. Even though, the preceding approaches are able to produce acceptable models of intrusion detection, they could not reach the ideal result.

In this paper, the technique which has been used to detect intrusion in the computer network is based on Fuzzy Rule-Based Classification Systems (FRBCS) [9,26]. In overall, finding a compressed set of if-then classification rules that are able to model the behavior of a system is the main obligation of a FRBCS. The first step of the proposed model is producing the rules that will be followed by selecting the best rules at the second step to reduce the complexity. In order to improve the classification rate, in this paper, an evolutionary approach is introduced for learning rule weights, which is evolved by genetic algorithm method [10,11]. Afterwards, genetic algorithm is used to find the optimum weights which are able to increase the level of accuracy for classification.

According to the results of the proposed method, classifiers based on fuzzy rules have a lower false alarm rate in comparison with the other classifiers. Therefore, classifiers based on fuzzy rules are the best choice for systems with the goal of having the lowest number of false alarms.

2. Related works

Researchers have proposed various methods for detecting intrusions. Su [12] proposed a method to identify flooding attacks in real-time, based on anomaly detection by genetic weighted KNN (K-nearest-neighbor) classifiers. A genetic algorithm is used to train an optimal weight vector for features. An SVM-based intrusion detection system is presented in [13], one which combined a hierarchical clustering algorithm (BIRCH), a simple feature selection procedure, and the SVM (Support Vector Machines) method. This method is also evaluated using on the KDD 1999 datasets. Mabu et al. [14] proposed a novel fuzzy class-association-rule mining method based on genetic network programming (GNP) for detecting network intrusions which can be flexibly applied to both misuse and anomaly detection in network-intrusion-detection problems. A PSO-based optimized clustering method (IDCPSO) proposed in [15] to optimize the clustering results and obtain the optimal detection result. Boughaci et al. [16] proposed a fuzzy particle optimization algorithm (FPSO) for intrusion detection that works on a knowledge base modeled as a fuzzy rule if-then and improved by a PSO algorithm. Abadeh et al. [17] proposed a technique base on fuzzy genetic learning. Moreover, they suggested a new fitness function calls SRPP. A method to cascade k-Means clustering and the C4.5 decision tree methods proposed in [18] to classify anomalous and normal activities in a computer network. At the first stage, k-Means clustering is performed on training instances to obtain k disjoint clusters. In the second stage, the k-Means method is cascaded with the C4.5 by building decision trees using the instances in each k-Means cluster. Nadiammai and Hemalatha [19] proposed EDADT algorithm to reduce the space occupied by the dataset. So, it would be useful for the network administrator/manager to avoid the delay between the arrival and the detection time of the attacks respectively. Yang et al. [20] proposed a data-driven network intrusion detection system using fuzzy interpolation in an effort to address the aforementioned limitations. The experiment results demonstrated that the proposed method detect the unknown types of treats.

In [21], A novel strategy for intrusion detection in wireless sensor networks based on accurate neural models of specific attacks learned from network traffic data is proposed and evaluated. In 2020, Zhang et al. [22] presented a new network intrusion detection method based on Auto-Encoder network (AN) and long-term memory neural network (LSTM). AN is constructed by superimposing multiple auto-encoder networks to map high-dimensional data to low-dimensional space. Then the LSTM model optimizes the cell structure is used to extract features, train data and predict intrusion detection types. Alazzam et al. [23] proposed a wrapper

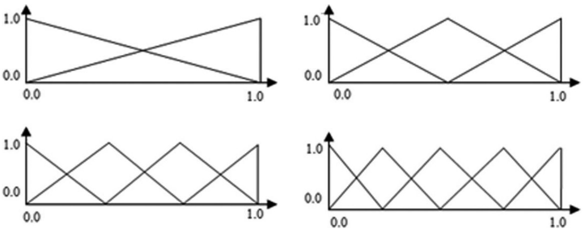

Fig. 1. Fuzzy partitions of the domain interval [0, 1].

feature selection algorithm for IDS based on pigeon inspired optimizer to utilize the selection process. They proposed a new method to binarize a continuous pigeon inspired optimizer.

Karthikeyan et al. [24] implemented a classifier ensemble based intrusion detection systems (CEBIDS) by combining feature level and data level techniques in WEKA tool with KDD cup'99 dataset. In [25], a clustering-based outlier detection (CBOD) approach is proposed for classifying normal and intrusive patterns. The proposed scheme extracts the most relevant features, then learns the normal pattern in the training data by forming clusters and identifies outliers in the testing data.

3. Generate of fuzzy If-Then rules

This paper uses a fuzzy rule-based classifier proposed in [26] which first fuzzy if-then rules are generated from numerical data. Then the generated rules are used as candidate rules. For an M-class problem in an n-dimensional feature space it is supposed that m real vectors $x_p = \{x_{p1}, x_{p2}, \ldots, x_{pn}\}$, $p = 1, 2, \ldots, m$, are given as training patterns. It is also assumed that each attribute of x_p is normalized to a unit interval [0, 1]. In the presented fuzzy classifier system; fuzzy if-then rules are used as the following form.

$$\text{Rule } R_q\text{: If } x_1 \text{ is } A_{q1} \text{ and } \ldots \text{ and } x_n \text{ is } A_{qn}$$
$$\text{then Class } C_q \text{ with CF }_q; \tag{1}$$

where R_q is the label of the q-th fuzzy if-then rule, $x = (x_1, \ldots, x_n)$ is an n-dimensional pattern vector, $A_q = (A_{q1}, \ldots, A_{qn})$ represents a set of antecedent fuzzy sets, C_q is the consequent class, CF_q is the confidence of the rule R_q, and N is the total number of generated fuzzy if-then rules. Also, Triangular membership functions are used as antecedent fuzzy sets. Figure 1 shows the domain interval of each attribute x_i which is divided into 14 fuzzy sets.

We define the compatibility grade of each training pattern x_p with the rule R_q by the product operation as

$$\mu_q(x_p) = \prod_{j=1}^{n} \mu_{qi}(x_{pi}) \tag{2}$$

where $\mu_{qi}(\cdot)$ is the membership function of the antecedent fuzzy set A_{qi}. On the other hand, the support of $(Aq \Rightarrow Cq)$ is defined as follows

$$S(A_q \Rightarrow \text{Class}C_q) = \frac{\sum_{x_p \in \text{class}C_q} \mu_q(x_p)}{m} \tag{3}$$

The support S indicates the grade of the coverage by $(A_q \Rightarrow C_q)$. Confidence (denoted by C) of a fuzzy rule R_q is defined as [27]

$$C(A_q \Rightarrow \text{Class}C_q) = \frac{\sum_{x_p \in \text{class}C_q} \mu_q(x_p)}{\sum_{p=1}^{m} \mu_q(x_p)} \tag{4}$$

The most common reasoning methods are single winner reasoning method and weighted vote reasoning method [28]. For classifying an input pattern $x_p = (x_{p1}, \ldots, x_{pn})$ using single winner reasoning method, the single winner rule R_w is determined as follows:

$$\mu_w(x_p) \cdot CF_w = \max\{\mu_q(x_p) \cdot CF_q :$$
$$q = 1, 2, \ldots, R, R_q \in S\} \tag{5}$$
$$w = \arg\max\{\mu_q(x_p) \cdot CF_q :$$
$$q = 1, 2, \ldots, R, R_q \in S\} \tag{6}$$

where R_w is the single winner rule and the new pattern X_p is classified as class C_w, which is the consequent class of the winner rule R_w. If there is not any X_p that is covered by the fuzzy if-then rules or in the case where multiple classes have the maximum value in Eq. (5), the consequent class is set as empty. When the 14 linguistic values are employed for each axis of the n-dimensional pattern space, the total number of possible fuzzy if-then rules is 14^n. In Ref [27] Ishibuchi and Yamamoto added the fuzzy set "don't care" to each attribute. The membership function of this fuzzy set is defined as don't care $(x) = 1$ for all values of x. In this paper only fuzzy if-then rules are generated whose length is less than or equal to two. Since the number of generated candidate rules can be quite large for the problem, therefore they are sorted in descending order according to the evaluation criterion in Eq. (7) and Q candidate rules will be selected from each class.

$$E(Rj) = \sum\nolimits_{x_p \in \text{class}C_j} \mu_j(x_p) - \sum\nolimits_{x_p \notin \text{class}C_j} \mu_j(x_p). \tag{7}$$

4. Proposed method for rule weighting

In analogy with the preceding approaches like [26], the proposed method stands for finding proper weights for rules. For the objective of improving the accuracy of fuzzy rule based classification systems, weights are considered for fuzzy rules. This paper uses genetic algorithm to find the most suitable weights for the rules. The process starts with generating initial random weights for initial fuzzy rules. Afterwards, suitable weights are found by genetic algorithm. The genetic algorithm considers the weight of each rule as an optimization parameter (gene). Regarding the definition of each gene, each individual in a population is composed of weights of all the rules. Also, an individual with the highest fitness will be selected as the final answer. Moreover, in this research, the accuracy of classification based on the fuzzy rules is considered as the fitness function. In the other words, the best class is identified for each input data and classes are classified accordingly. Finally, the accuracy of classification is calculated for both attack and normal data. This section details the proposed method.

4.1. *Genetic algorithm*

Genetic algorithms [10,11,29] by inheriting the process of natural evolution, such as inheritance, mutation, selection and genetic crossover, perform heuristic search and optimization practices and provide solutions to optimization problems, such as inheritance, mutation, selection and Genetic intersection that occurs during the marriage of parents to produce offspring. It is used to guide a method of evaluating a list of parameters that provide possible solutions to the problem (also called chromosomes or genomes) in this list of parameters. Evolution is an iterative process that usually starts from a random population. The population of each repeat is called a *generation*. These chromosomes are evaluated and a value of *goodness* or *fitness* is returned. Usually, the algorithm terminates either by producing the maximum number of generations, or by achieving acceptable fitness levels for the population. The genetic algorithm is sketched in Fig. 2.

Initially, many individual solutions (chromosomes) are generated randomly to form an initial population. Individual solutions are selected using Roulette Wheel method through a *fitness-based* process, where fitter solutions (as measured by a fitness function) are typically more likely to be selected. The fitness value of the individual is the best accuracy. Next, a population of second-generation solutions is generated from the solutions selected through genetic operators: crossover and mutation. By producing a “child” solution using the above methods of *One-point crossover* and *Generational replacement* (mutation), a new solution is created which typically shares many of the characteristics of its “parents". New parents are selected for each new child, and the process continues until a new population of solutions of appropriate size is generated.

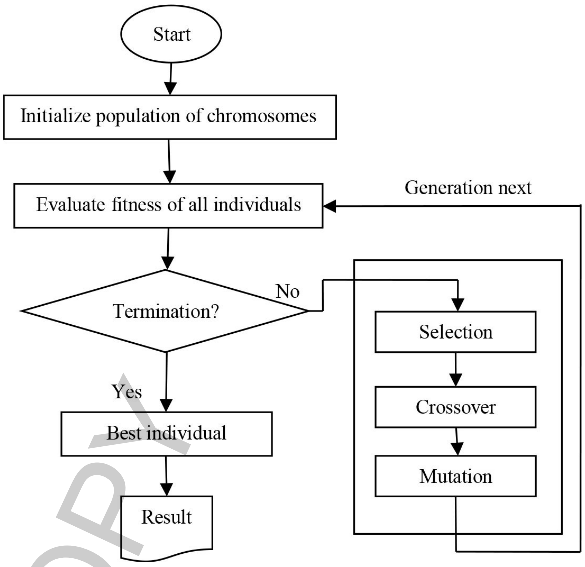

Fig. 2. The operation of a GA.

4.2. *Proposed algorithm*

Initially, each feature of the data set is normalized. The normalization formula given in Eq. (8) is applied in order to set attribute numerical values in the range [0.0, 1.0].

$$x = \frac{(x - \min)}{(\max - \min)} \tag{8}$$

where x is the numerical attribute value, min is the minimum value that the attribute x can get and max is the maximum one. All fuzzy rules of length one and two are generated, according to the method of the previous section. Then the genetic algorithm is used to select the best weight fuzzy rules. Figure 3 shows an overview of Fuzzy Rule – Based classification system and proposed method for rule weight specification. In this method, after initial populations of individuals are generated, the fitness value is calculated for each individual and the best individual (rule weighting) with the highest fitness is selected.

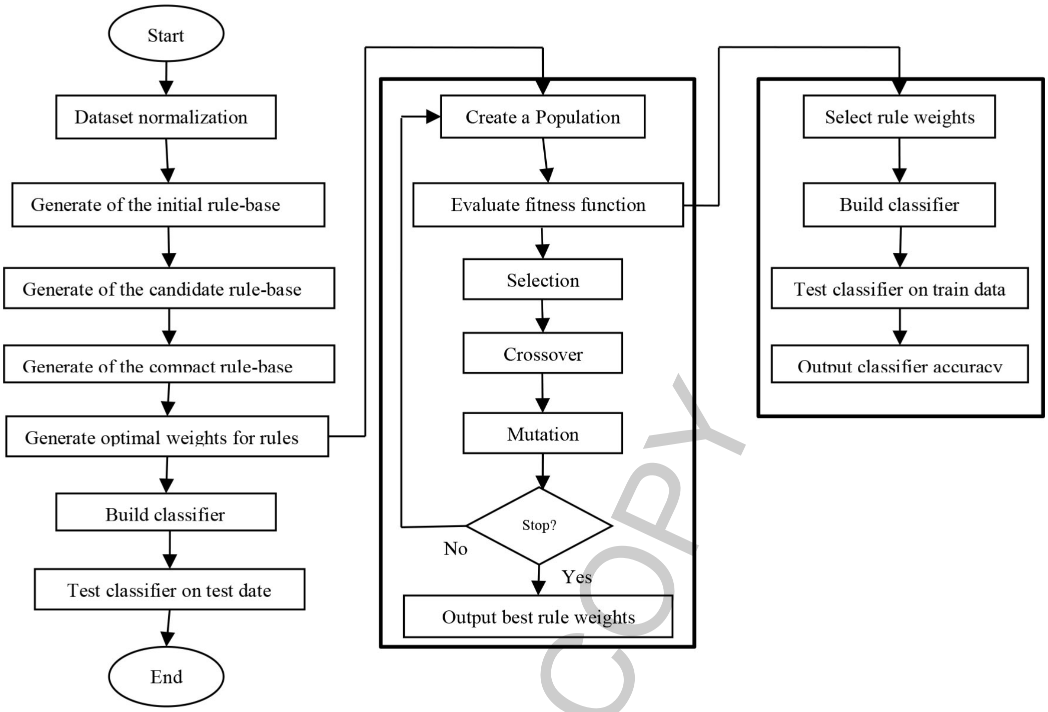
Start
Dataset normalization
Generate of the initial rule-base
Generate of the candidate rule-base
Generate of the compact rule-base
Generate optimal weights for rules
Build classifier
Test classifier on test date
End
Create a Population
Evaluate fitness function
Selection
Crossover
Mutation
Stop?
No
Yes
Output best rule weights
Select rule weights
Build classifier
Test classifier on train data
Output classifier accuracy

Fig. 3. An overview of the proposed method.

5. Experiments and results

To evaluate the performance of the proposed method, a series of experiments on a subset of the KDD CUP 1999 dataset are conducted. In these experiments, the proposed method is implemented and is evaluated and is made comparisons in order to validate the performance analysis of the proposed algorithm.

5.1. Dataset

In this section, the experimental dataset KDD99 [30] is discussed for intrusion detection. This data set prepared and managed by MIT Lincoln Labs. The dataset has 41 different attributes (32 continuous attributes and 9 discrete attributes) and 1 attack type label. The dataset contains about five million connection records as training data and about two million connection records as test data. Each data point represents either an attack or a normal connection. There are four categories of attacks, namely Denial of Service (DoS): making some computing or memory resources too busy to accept legitimate users access these resources, Probe (PRB): host and port scans to gather information or find known vulnerabilities, Remote to Local (R2L): unauthorized access from a remote machine in order to exploit machine's vulnerabilities and User to Root (U2R): unauthorized access to local super user (root) privileges using system's susceptibility [15].

In order to implement the proposed method, a subset of this large dataset is used as train and test datasets; hence 10,000 generated samples are selected randomly. Table 1 presents the distribution of classes in the dataset.

5.2. Results

In experiments, each item is described by 41 features which form a vector and normalized the train and test data sets, where each numerical value in the data set is normalized between 0 and 1. To construct a compact rule-base, 40 rules ($Q = 40$) from each class in candidate rule base are selected. Table 2 shows the parameter specifications used in computer simulations for the genetic algorithm. Leave one out technique (which is a special case of n-fold cross validation) is used to assess the generalization ability of the proposed method. To evaluate the accuracy of a system, two indicators used in [19] are used: Detection Rate (DR) and False Alarm Rate (FAR). DR equals the number of intrusions divided by the total number of intrusions in the data set; FAR equals the total of normal data that are mistakenly taken as an attack. The classification performance of

Table 1
Distribution of different classes in dataset

Classes	Samples
Normal	2432
DOS	5000
PR	1520
U2R	50
R2L	1000

Table 2
Parameters setting for the genetic algorithm

Parameters	Values
Population size	500
Crossover probability (Pc)	0.7
Mutation probability (Pm)	0.005
Maximum number of generations	50
Rand function	Uniform distribution

Table 3
Detection rate and False alarm rate obtained with different techniques

Technique	DR (%)	FAR (%)
C4.5	93.23	1.56
MLP	94.50	1.00
SVM	87.18	3.2
Naïve Bayes	93.20	4.2
Y-Means clustering	89.89	1.00
K-means clustering	89.40	5.7
K-NN	92.00	1.00
Neural network + PCA	92.22	–
C4.5 + PCA	92.16	–
Hidden Markov Model	79.00	–
SVM + PSO	91.57	1.94
Proposed method	*95.33*	*0.18*

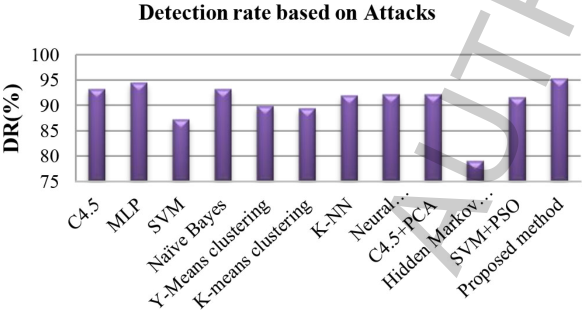

Fig. 4. Comparison between detection rate obtained by different techniques.

the proposed method is measured and compared with that of different baseline algorithms (i.e., C4.5, Naïve Bayes (NB), k-Nearest Neighbor (k-NN) and Support Vector Machine (SVM). Note that in k-NN classifier parameter k is set to 5). Table 3 shows the comparative results of different methods. Figures 4 and 5 specify the corresponding chart for the result obtained in Table 3.

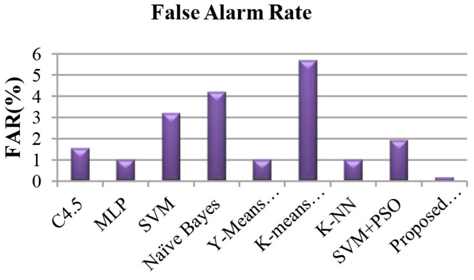

Fig. 5. Comparison between false alarm rate obtained by different techniques.

6. Conclusion and future work

In this paper, a fuzzy rule-base classification system is used to find a compact set of fuzzy if-then classification rules. Then, a new method is proposed for rule weights specification witch this method is based on genetic algorithm. It is desirable for anomaly rule-based IDS to achieve high classification accuracy, the proposed method is an accurate and interpretable fuzzy system for intrusion detection. Moreover, KDD99 data set is used for conducting the experiments. Performance analysis is measured by using DR and FAR which are two important criteria for security systems. The simulation experiments compared with other algorithms prove that the proposed method obtains higher detection rate and lower FAR. We must be mentioned that no method is able to achieve the best performance on all criteria. Thus, it is necessary to use more than one performance measure to evaluate performance of intrusion detection systems. Our future work will focus on introducing a feature selection technique on intrusion detection dataset for identifying the most suitable feature subsets which may provide better results in the shortest time.

References

[1] Axelsson S. Intrusion detection systems: A survey and taxonomy. Dept of Computer Engineering. Chalmers University Technical Report. 2000; 99-15.

[2] Amoroso E. Intrusion detection: An introduction to internet surveillance, correlation, traps, trace back, and response. Sparta, NJ: Intrusion Net Books; 1999.

[3] Jain A, Sharma S, Sisodia MS. Network intrusion detection by using supervised and unsupervised machine learning, techniques: A survey. International Journal of Computer Technology and Electronics Engineering. 2011; 1(4): 14-20.

[4] Lee W, Stolfo S, Mok K. Mining audit data to build intrusion detection models. Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining. 1998; 66-72.

[5] Zamini M, Hasheminejad MH. A comprehensive survey of anomaly detection in banking. Wireless sensor networks, social networks, and healthcare. Intelligent Decision Technologies. 2019; 13(2): 229-270.

[6] Kuchaki Rafsanjani M, Asghari Varzaneh Z. Intrusion detection by data mining algorithms: A review. Journal of New Results in Science. 2013; (2): 76-91.

[7] Ishibuchi H, Nozaki K, Tanaka H. Distributed representation of fuzzy rules and its application to pattern classification. Fuzzy Sets and Systems. 1992; 52(1): 21-32.

[8] Wangm LX, Mendel JM. Generating fuzzy rules by learning from examples. IEEE Transactions on Systems, Man, and Cybernetics. 1992; 22(6): 1414-1427.

[9] Zolghadri M, Taheri M. A proposed method for learning rule weights in fuzzy rule-based classification systems. Fuzzy Sets and Systems. 2008; 159: 449-459.

[10] Holland JH. Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control and artificial intelligence. 2nd ed. MIT Press; 1992.

[11] Horn J, Nafpliotis N, Goldberg DE. A niched pareto genetic algorithm for multi-objective optimization. Proceeding of the first IEEE Conference on Evolutionary Computation, IEEE World Congress on Computational Intelligence. 1994; 1: 82-87.

[12] Su MY. Using clustering to improve the KNN-based classifiers for online anomaly network traffic identification. Journal of Network and Computer Applications. 2011; 34: 722-730.

[13] Horng SJ, Su MY, Chen YH, Kao TW, Chen RJ, Lai JL, Perkasa CD. A novel intrusion detection system based on hierarchical clustering and support vector machines. Expert Systems with Applications. 2011; 38: 306-313.

[14] Mabu SH, Chen C, Lu N, Shimada K, Hirasawa K. An intrusion-detection model based on fuzzy class-association-rule mining using genetic network programming. IEEE Transactions On Systems, Man, and Cybernetics. 2011; 41(1).

[15] Zheng H, Hou M, Wang Y. An efficient hybrid clustering-PSO algorithm for anomaly intrusion detection. Journal of Software. 2011; 6(12).

[16] Boughaci D, Kadi MDE, Kada M. Fuzzy particle swarm optimization for intrusion detection. Springer-Verlag. 2012; 541-548.

[17] Saniee Abadeha M, Habibi J, Lucas C. Intrusion detection using a fuzzy genetics-based learning algorithm. Journal of Network and Computer Applications. 2007; 30: 414-428.

[18] Muniyandi AP, Rajeswari R, Rajaram R. Network anomaly detection by cascading K-means clustering and C4.5 decision tree algorithm. Procedia Engineering. 2012; 30: 174-182.

[19] Nadiammai GV, Hemalatha M. An enhanced rule approach for network intrusion detection using efficient data adapted decision tree algorithm. Journal of Theoretical and Applied Information Technology. 2013; 47(2).

[20] Yang L, Li J, Fehringer G, Barraclough Ph, Sexton G, Cao V. Intrusion detection system by fuzzy interpolation. Proceeding of the IEEE International Conference on Fuzzy Systems. 2017.

[21] Batiha T, Prauzek M, Krömer P. Intrusion detection in wireless sensor networks by an ensemble of artificial neural networks. Intelligent Decision Technologies. 2019; 323-333.

[22] Zhang Y, Zhang Y, Zhang N, Xiao M. A network intrusion detection method based on deep learning with higher accuracy. Procedia Computer Science. 2020; 174: 50-54.

[23] Alazzam H, Sharieh A, Sabri KE. A feature selection algorithm for intrusion detection system based on Pigeon Inspired Optimizer. Expert Systems with Applications. 2020; 148.

[24] Karthikeyan D, Mohanraj V, Suresh Y, Senthilkumar J. Hybrid intrusion detection system security enrichment using classifier ensemble. Journal of Computational and Theoretical Nanoscience. 2020; 17(1): 434-438.

[25] Rene Beulah J, Shalini D. An efficient mixed attribute outlier detection method for identifying network intrusions. International Journal of Information Security and Privacy (IJISP). 2020; 14(3).

[26] Ishibuchi H, Yamamoto T. Rule weight specification in fuzzy rule-based classification systems. IEEE Transactions on Fuzzy Systems. 2005; 13(4): 428-435.

[27] Ishibuchi H, Yamamoto T. Fuzzy rule selection by multi-objective genetic local search algorithms and rule evaluation measures in data mining. Fuzzy Sets and Systems. 2004; 141(1): 59-88.

[28] Ishibuchi H, Nakashima T, Morisawa T. Voting in fuzzy rule-based systems for pattern classification problems. Fuzzy Sets and Systems. 1999; 103(2): 223-238.

[29] Sivanandam SN, Deepa SN. Introduction to Genetic algorithms. Springer Berlin Heidelberg New York. 2008.

[30] Index of /databases/kddcup99 [EB/OL]. http://kdd.ics.uci.edu/databases/kddcup99. 2009.