Intrusion Detection with Low False Alarms using Decision Tree-based SVM Classifier

Tree-based Intrusion Detection

Aliakbar Tajari Siahmarzkooh
Department of Computer Sciences, Faculty of Sciences
Golestan University
Gorgan, Iran
a.tajari@gu.ac.ir

Received: 2021/05/03 Revised: 2021/06/03 Accepted: 2021/07/15

***Abstract*— Todays, Intrusion Detection Systems (IDS) are considered as key components of security networks. However, high false positive and false negative rates are the important problems of these systems. On the other hand, many of the existing solutions in the articles are restricted to class datasets due to the use of a specific technique, but in real applications they may have multi-variant datasets. With the impetus of the facts, this paper presents a new anomaly based intrusion detection system using J48 Decision Tree, Support Vector Classifier (SVC) and k-means clustering algorithm in order to reduce false alarm rates and enhance the system performance. J48 decision tree algorithm is used to select the best features and optimize the dataset. Also, an SVM classifier and a modified k-means clustering algorithm are used to build a profile of normal and anomalous behaviors of dataset. Simulation results on benchmark NSL-KDD, CICIDS2017 and synthetic datasets confirm that the proposed method has significant performance in comparison with previous approaches.**

Keywords—Intrusion Detection; K-Means Clustering; Decision Tree; Support Vector Classifier; NSL-KDD Dataset.

1. INTRODUCTION

With the spread of the Internet around the world, computer networks have penetrated in almost every aspect of our lives. Especially in recent years, networks have been widely used for Internet services such as online trading, internet banking, online social networks and many other online services. The popularity of the Internet has changed the computing infrastructure used by the user community. However, the growing storage and processing requirements have created many problems for most Internet-based activities [1].

In past years, there have been many attacks on Internet networks that have disrupted their workflows. For example, over 20% of companies over the world have reported at least one DDoS (Distributed Denial of Service) attack on their substructure [2]. It is said that billions of dollars has been stolen using virtual attacks in recent years [2]. In order to remain secure against the different types of attacks, it is needed to have the ability to react to known attacks and also to new security challenges. Therefore, attack detection approaches are important research topics among different issues [3, 4].

In general, the main purpose of Intrusion Detection Systems (IDS) is to detect security attacks and consequently, prevent a dangerous event in the network. IDSs are usually used as a supplement for firewalls. These security systems are used to analyze unpleasant occurrences that put security policies in computer networks under pressure [5]. Indeed, IDSs should be able to detect all abnormal traffic by analyzing activities in the network data. However, due to its large amount of data in the network, IDSs face the problem of data processing [6]. Hence, there are various approaches that can confront this problem. Some of related works done in this scope are presented in section 1.1.

1-1. Related Work

Guo et al. [7] used a combination of k-NN and k-means clustering algorithm to reduce false positive rates alarms. Mazraeh et al. [8] proposed a solution based on AdaBoost algorithm and J48 decision tree and also they used Naïve Bayes for feature selection. In [9], Al-Yaseen et al. presented a multilevel attack detection model using k-means and support vector machines algorithm. The efficiency and detection rate in this method was so high. Prasad et al. [10] used cuckoo search strategy and shark algorithm to reduce the overhead of computational cost. Singaravelan et al. [11] proposed a system based on data mining rules to increase accuracy and decrease response time. Venkatesan et al. [12] proposed a cookie-based approach to enhance the efficiency of systems.

In the survey of Chandola et al. [13], relations between proposed ways are introduced for different applications. Also, Akoglu et al. [14] studied the concept of anomaly detection in real world area such as social networks. Celik et al. [15] used a novel clustering with a noise algorithm for training and testing the data to detect the anomalies. Lv et al. [16] proposed a new algorithm based on machine learning model and classifier with local sensitive hashing. Wang et al. [17] proposed a statistical method using probability function to detect the attacks.

Moreover, many researches have been proposed in other articles to detect malicious behavior. For example, authors in [18] proposed a new attack detection method for smart grid. However, its simulation results were not enough for real-time scenarios. Furthermore, this way had less detection rate in the real scenario. In [19], a multilevel anomaly detection model was proposed which analyzed log files to identify intrusions, however, it was inefficient for large amount of network data.

1-2. Motivation

It is obvious that proposing a novel anomaly detection method on the initial data has high error rates. Therefore,

intrusion detection problem can be introduced as an important problem which can be resolved with more analysis of data features. These features should consider the large size and variety of datasets and also selection of an appropriate clustering approach. A proper feature selection is applied using J48 decision tree to reduce the dataset's dimensionality. Moreover, proposing an appropriate clustering algorithm is an important issue in creating an effective intrusion detection model. A combination of SVC and Fuzzy C-means algorithm is useful to overcome the problem of k-means clustering [20] that each node has a probability of belonging to more than one cluster.

In this paper, a feature selection strategy is proposed based on J48 decision tree. Apart from that, a dataset is generated using network simulator tool and then, the performance of the proposed method is evaluated by using the standard datasets and manual dataset. Finally, a modified k-means clustering algorithm and SVC are used to identify normal and abnormal data. Simulation results show the efficiency of our proposed solution to detect attacks in network data.

1-3. Organization

The rest of the paper is organized as follows: section 2 presents the proposed model followed by its detailed description. In section 3, results are presented based on the experimental evaluation. Finally, section 4 concludes the paper.

2. PROPOSED MODEL

The proposed intrusion detection system includes three phases: i) data preprocessing including the data normalization and discretization, ii) feature selection and iii) applying data clustering to identify normal and attack data.

Our strategy for designing the intrusion detection system is based on J48 decision tree feature selection and cluster-based classifier. The k-means clustering classifier singly provides low ratio of intrusion detection for different types of attacks. Hence, a SVC is used to enhance the accuracy of clustering. Each attack is trained with the selected features chosen based on its special feature selection technique. The proposed method includes the following steps:

- Preprocessing: The network data should be analyzed before the data could be processed. In this phase, some preprocessing actions apply to remove the redundant values. Also, the data is normalized into the desired format. Apart from this, the preprocessed data is divided into two sections for training and testing purposes.
- Feature selection: Some of the features are not appropriate for identifying the class of instances in the dataset and should be ignored. In this phase, related features are identified from the dataset using J48 decision tree. These features are more effective in identifying the label or class of dataset.
- Clustering: In this phase, optimal clusters are provided to achieve appropriate detection rates, reduce false positive and false negative rates. In fact, anomalies are detected from normal behavior of data.

Fig. 1 represents the proposed secure system to detect anomalies. The proposed solution is a secure system for different types of attacks. It uses the most effective features for each type of attack to detect attacks by using clustering algorithm. But it has some drawbacks such as the model does not change during detection. Thus, the detection rate would be improved by adding learning capability in clustering phase.

The fuzzy learning solution detects attacks using C4.5 feedbacks. The information about difference between anomalies and normal data is saved into the database to find unknown attacks. Then, the proper fuzzy action is selected and inputs are converted to fuzzy sets. Four classes of fuzzy sets are considered as four different situations of state space. Fuzzy rules are defined using fuzzy inputs. Fuzzy classes are used in modeling the attack behavior. A weight would be assigned to all states based on fuzzy logic.

In the following, we study the details of components of proposed approach.

2-1. J48 Decision Tree

J48 builds a decision tree from a training dataset. It uses the concept of information entropy. The training data is a set $S=s_1,s_2, \ldots$ of already classified instances. Each instance includes a p-dimensional vector($x_{1,i}$, $x_{2,i}$, $x_{p,i}$), where x_j reveals features values of instances. In each node of the decision tree, the feature that most effectively divides the instances into subsets is selected.

The splitting parameter is the normalized information gain. Therefore, the feature with the highest normalized information gain is selected. The algorithm repeats on the partitioned subsets. If all of list instances belong to the same class, it builds a leaf node for the decision tree and chooses a label for it. The general algorithm for building a decision tree is as Fig. 2.

2-2. SVC

A support vector machine is based on a supervised learning algorithm and is identified as a classifying concept. This algorithm consists of lines such as (1) follow form,, where w, x and b are the weight, input data and bias, respectively.

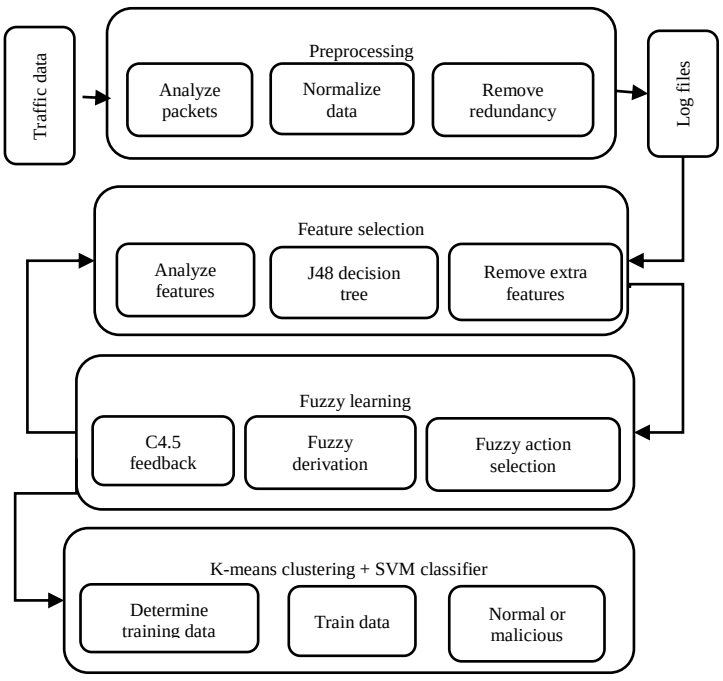

Fig. 1. The flowchart of the proposed security system

Algorithm of J48 decision tree
Input: a dataset, **output:** a decision tree
Step 1: Check the above base cases.
Step 2: For each feature a, find the normalized information gain ratio from splitting on a, based on this equation. $Info_A(D) = \sum_{j=1}^{V} \frac{|D_j|}{|D|} * Info(D_j)$, where v is the number of partitions and D is the dataset.
Step 3: Let a_best be the feature with the highest normalized information gain.
Step 4: Create a decision node that splits on a_best.
Step 5: Repeat subsets achieved by splitting on a_best and add those nodes as children of node.

Fig. 2. Pseudo-code of J48 decision tree

$$f(x) = w^T x + b = 0 \qquad (1)$$

SVC is a support vector classifier can fulfill multi-class classifications. It takes two arrays as input: an array X including the training data and an array y of class labels. After training phase, the model is proper for prediction targets. SVC functions depend on the subset of training dataset, named the support vectors. SVC implements multi-class classification. If n is the number of classes, then n * (n-1) / 2 classifiers are built and each one trains the data from two classes.

Given training vectors $x_i \in R^p$, i= 1, 2, …, n in two classes and a vector $y \in \{-1,1\}^n$, SVC solves the primal problem shown in (2) to (6):

$$\varphi = \min_{w,b,\rho} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \rho_i \qquad (2)$$

Subject to:

$$y_i(w^T \varphi(x_i) + b) \geq 1 - \rho_i \qquad (3)$$

And

$$\rho_i \geq 0, i = 1, \dots, n \qquad (4)$$

Its dual is as follows:

$$\min_{\alpha} \frac{1}{2} \alpha^T Q \alpha - e^T \alpha \qquad (5)$$

Subject to:

$$y^T \alpha = 0, 0 \leq \alpha_i \leq C\ , i = 1, \dots, n \qquad (6)$$

Where e is the vector of all ones is, C>0 is the upper bound, Q is a semi-definite matrix. The training vectors are mapped into a larger dimensional space by the function φ.

2-3. Modified K-means Clustering

The main purpose of fuzzy C-mean algorithm is to minimize the objective function shown in (7) and (8) [21], where X, V and U represent the dataset, cluster centers and membership degree.

$$J(U,V,X) = \sum_{k=1}^{c} \sum_{i=1}^{n} (u_{k,i})^m ||x_i - v_k||^2 \qquad (7)$$

$$\sum_{k=1}^{c} U_{k,i} = 1 \qquad (8)$$

Based on the research paper of Bezdek [22], parameter m is usually assumed equal 2, which is related to fuzzy membership of each instance. Finally, $\| x_i - v_k \|$ is used to reveal the Euclidean distance.

This fuzzy algorithm repeats after a termination term occurs. In this work, we use the maximum number of iterations (n=100) as the termination condition. Also, to overcome the dependency of algorithm to the centers, we consider additional hypothetical 20% centers as the set of centers of clusters. This work would be fix the local minimum problem.

3. EXPERIMENTAL RESULTS

In this section, the efficiency of the proposed intrusion detection system is evaluated on different datasets such as NSL-KDD, CICIDS2017 and synthetic dataset. The detailed explanation is mentioned as follows.

3-1. Attacks

In this section, some of the mentioned types of attacks are described.

DDoS: A distributed denial of service attack is every anomalous attempt to damage the normal traffic of targeted servers, services and networks by overwhelming targets or their surrounding infrastructure with a flood of Internet traffic. From another view, DDoS attacks are like unexpected traffic on the highway, preventing regular traffic from arriving at their destinations.

Dictionary: A dictionary attack is a brute force type of attack techniques for destroying a cipher or authentication methodologies by trying to recognize its decryption passwords by trying millions of words in a dictionary or previously used passwords, often from lists obtained from past security gaps.

SYN Flooding: A SYN flooding attack is a type of DoS attack which tries to make a server unavailable to legitimate users and normal traffic by occupying all available resources. In fact, attackers send connection requests such as SYN packets to server resources to overwhelm all available ports on a targeted machine. Therefore, the targeted server or device responds to normal traffic slowly.

Port Scan: Port scanning attack is a common method used by hackers to detect open doors and network vulnerabilities. A port scan attack helps disruptors find open ports to receive and eavesdrop on data traffic. It can also determine whether active security devices such as firewalls are used by an organization.

After the hackers send a message to the ports, they receive a response that determines whether those ports are being used and whether there are vulnerabilities that can be exploited.

R2L: Remote to user or R2L attack is a type of network attacks, in which attackers send some of packets to another device or server over a network where he/she does not have permission to access as a legitimate user.

Patator: Patator is a multi-purpose brute force attack, with a modular design and a flexible usage that supports some modules such as: ftp_login, ssh_login, pop_pasd, etc.

3-2. NSL-KDD Dataset

NSL-KDD dataset [23] is used in the evaluation process which is a modified version of KDD99 dataset. It was suggested to solve the redundancy, which KDD99 suffer from

it. This dataset consist of several types of attacks simulated in the military network. NSL-KDD dataset comprises of 23 different types of intrusions, 41 features and 4 classes including DoS, Probe, User to Root (U2R) and Remote to Local (R2L).

3-3. CICIDS2017 Dataset

CICIDS2017 dataset includes normal and some new attacks, which simulates the real network data. This dataset also consists of the traffic analysis using with labeled flows. Generating real traffic was the important principle in creating this dataset. The data capturing started at 9 AM, Monday, July 3, 2017 and ended at 5 PM on Friday July 7, 2017, in 5 days. Monday is the normal day and only includes the normal data. The existing attacks are FTP, SSH, DoS, botnet and DDoS. They have been applied both morning and afternoon on Tuesday, Wednesday, Thursday and Friday.

3-4. Synthetic Dataset

Basically, a dataset includes different properties gathered from the supervised network and is used to compare the performance of an intrusion detection system with another ones [24]. The dataset is consists of two parts including training data and testing data which are used to construct the model and evaluate the model, respectively.

In this paper, the generated dataset includes normal data, DDoS, dictionary, flooding, IP scan and R2L attack data which are generated by the following steps:

1) One minute time frame is assigned to each piece of data and then, the properties of the data are characterized by the basic network features such as protocol, source and destination IP and port, number of received packets and etc.

2) Based on the ratio of the number of received packets and sent packets, the class of data is identified; if it was more than a threshold then attack is labeled, else normal is detected.

3) It is obvious that, the normal and attack models are generated separately.

The generated dataset includes 248,000 packets which consist of 12 features are presented in Table 1.

3-5. Performance Evaluation

In this paper, J48 decision tree is used to select the effective features for clustering in dataset. Statistical classifiers are used as the main core of feature selection strategy. Then, the modified k-means clustering algorithm is applied to determine the class of the data. Hence, based on the data label, further processing is allowed or prevented. Simulation results would be available using the synthetic dataset, which is divided into two categories. The train section is used for the training purposes and the other section is used for testing the proposed solution. As it was intended from the beginning, we would like to improve the false positive and false negative rates.

The proposed approach is evaluated based on the standard datasets and synthetic dataset collected from the network. The detection rate, false positive and false negative rates and some other criteria of the simulation results on the synthetic dataset are shown in Table 2 (The mentioned criteria are shown in (9) to (15)).

$$FPR = \frac{FP}{FP+TN} \quad (9)$$

$$R = \frac{FN}{FP+TN} \quad (10)$$

$$Precision = \frac{TP}{TP+FP} \quad (11)$$

$$Recall = \frac{TP}{TP+FN} \quad (12)$$

$$f-measure = \frac{2*Precision*Recall}{Precision+Recall} \quad (13)$$

$$Sensitivity = \frac{TP}{P} \quad (14)$$

$$Specificity = \frac{TN}{N} \quad (15)$$

Apart from this, the evaluation of the proposed approach on standard datasets is an important criterion to identify the advantages of the intrusion detection system. Therefore, the simulation results of proposed method on standard datasets such as NSL-KDD and CICIDS2017 are shown in Table 3 and Table 4, respectively.

Tables 2-4 clearly reveals that false positive and false negative rates are very low values and are so good result for us. However, detection rates or f-measures are not so good, but, it is acceptable.

Next, Table 5 shows the superiority of the proposed solution over the other approaches such as SVM (Support Vector Machine), GA (Genetic Algorithm), PSO (Particle Swarm Optimization) and Bayesian network. Also, Table 6 represents the training and testing times of our approach which demonstrates that less time is needed for training and testing by using the attributes of J48 decision tree, SVC and k-means clustering algorithm. Thus, they are appropriate for this issue.

TABLE 1. FEATURES OF SYNTHETIC DATASET

Feature Number	***Feature Name***	***Feature Number***	***Feature Name***
1	Packet size	7	Destination IP
2	Received packets	8	Number of flows
3	Packet delivery ratio	9	Flow delivery ratio
4	Source port	10	Received bytes
5	Destination port	11	Byte delivery ratio
6	Source IP	12	Time

TABLE 2. SIMULATION RESULTS OF PROPOSED APPROACH ON SYNTHETIC DATASET

Attack Name	***False Positive***	***False Negative***	***F-measure***	***Sensitivity***	***Specificity***
Normal data	0.004	0.008	0.993	0.984	0.966
DDoS	0.017	0.044	0.977	0.966	0.938
Dictionary	0.012	0.009	0.955	0.933	0.914
Flooding	0.088	0.059	0.922	0.928	0.918
Port scan	0.098	0.118	0.936	0.915	0.918
R2L	0.046	0.073	0.901	0.892	0.904

TABLE 3. SIMULATION RESULTS OF PROPOSED APPROACH ON NSL-KDD DATASET

Attack Name	*False Positive*	*False Negative*	*F-measure*	*Sensitivity*	*Specificity*
Normal data	0.022	0.038	0.937	0.976	0.981
DoS	0.007	0.022	0.938	0.983	0.965
U2R	0.007	0.018	0.966	0.977	0.983
Probe	0.044	0.027	0.936	0.963	0.925
R2L	0.015	0.036	0.944	0.988	0.944

TABLE 4. SIMULATION RESULTS OF PROPOSED APPROACH ON NSL-KDD DATASET

Attack Name	*False Positive*	*False Negative*	*F-measure*	*Sensitivity*	*Specificity*
Normal data	0.024	0.026	0.966	0.975	0.982
DDoS	0.032	0.031	0.984	0.948	0.968
FTP patator	0.025	0.019	0.974	0.957	0.984
SSH patator	0.066	0.053	0.922	0.988	0.917
Port scan	0.008	0.018	0.977	0.975	0.989

TABLE 5. COMPARISON OF THE PROPOSED SOLUTION WITH OTHER APPROACHES

Attack Name	*False Positive*	*False Negative*	*F-measure*	*Sensitivity*	*Specificity*
Proposed approach (DT+SVC+ k-means)	0.009	0.012	0.983	0.978	0.981
PSO	0.089	0.079	0.912	0.922	0.903
SVM	0.055	0.063	0.933	0.966	0.976
GA	0.044	0.036	0.928	0.884	0.904
Bayesian	0.083	0.074	0.915	0.859	0.845

TABLE 6. COMPARISON OF THE TRAINING AND TESTING TIMES WITH OTHER APPROACHES

Approach	*Training Time*	*Testing Time*
Proposed approach (DT+SVC+ k-means)	0.5415	0.0013
PSO	0.5819	0.0025
SVM	0.5914	0.0025
GA	0.5618	0.0026
Bayesian	0.5928	0.0019

4. CONCLUSION

In this paper, more effective attributes of datasets are selected to detect the class of each record using the features of J48 decision tree algorithm. Therefore, in the next step, these useful features of the datasets and then using the support vector and k-means clustering yield high detection rates and low false positive and negative rates for all types of existing attacks. The efficiencies of the proposed approach in feature selection and classifying the data have been clearly shown with comparing the results with some other solutions on different intrusion detection datasets. As a final result, the comparisons represent that the proposed attack detection method using the combination of attributes of decision tree, k-means clustering and SVC led to less training and testing times, less false positive and less false negative rates, and so, it is appropriate for securing the network systems. As a future work, some preprocessing ideas could be done on the dataset to obtain high accuracy.

ACKNOWLEDGMENT

This study is extracted form a research Project No. 992031 entitled "Intrusion detection in networks using decision tree and feature reduction" at Golestan University.

REFERENCES

[1] S. Gupta, A. Garg, A. Singh, S. Batra, N. Kumar, and M. Obaidat, "ProIDS: Probabilistic Data Structures Based Intrusion Detection System for Network Traffic Monitoring", in *IEEE Global Communications Conference (GLOBECOM 17)*, 2017, pp. 1-6.

[2] *Internet Security Threat Report (ISTR)*, 2017. URL: https://docs.broadcom.com/doc/istr-22-2017-en.

[3] S. Garg, A. Singh, S. Batra, N. Kumar, and L. Yang, "UAV-Empowered Edge Computing Environment for Cyber-Threat Detection in Smart Vehicles", *IEEE Network*, Vol. 32, No. 3, pp. 42–51, 2018.

[4] S. Garg, K. Kaur, N. Kumar, S. Batra, and M. Obaidat, "HyClass: Hybrid Classification Model for Anomaly Detection in Cloud Environment", in *2018 IEEE International Conference on Communications (ICC)*, 2018, pp. 1-7.

[5] M. Raman, N. Somu, K. Kirthivasan, R. Lisano, and V. Sriram, "An efficient intrusion detection system based on hyper graph Genetic algorithm for parameter optimization and feature selection in support vector machine, *Knowledge-Based Systems*, Vol. 134, No. 4, pp. 1-12, 2017.

[6] R. Singh, H. Kumar, and R. Singla, "An intrusion detection system using network traffic profiling and online sequential extreme learning machine", *Expert Systems with Applications*, Vol. 42, No. 22, pp. 8609-8624, 2015.

[7] C. Guo, Y. Ping, N. Liu, and S. Luo, "A two level hybrid approach for intrusion detection", *Neurocomputing*, Vol. 214, No. 4, pp. 391-400, 2016.

[8] S. Mazraeh, M. Ghanavati, and S. Neysi, "Intrusion detection system with decision tree and combine method algorithm", *International Academic Journal of Science and Engineering*, Vol. 3, No. 2, pp. 21-31, 2016.

[9] W. Al-Yaseen, Z. Othman, and M. Nazri, "Multi-level hybrid support vector machine and extreme learning machine based on modifed K-means for intrusion detection system", *Expert Systems with Applications*, Vol. 67, No. 1, pp. 296-303, 2016.

[10] K. Prasad, A. Reddy, and K. Rao, "BARTD: Bioinspired anomaly based real time detection of under rated App-DDoS attack on web", *Journal of King Saud University– Computer and Information Sciences*, Vol. 32, No. 1, pp. 73-87, 2017.

[11] S. Singaravelan, R. Arun, D. Arunshunmugam, S. Joy, and D. Murugan, "Inner interruption discovery and defense system by using data mining", *Journal of King Saud University-Computer and Information Sciences*, 2017, in press.

[12] S. Venkatesan, M. Basha, C. Chellappan, A. Vaish, and P. Dhavachelvan, "Analysis of accounting models for the detection of duplicate requests in web services", *Journal of King Saud University-Computer and Information Sciences*, Vol. 25, No. 1, pp. 7-24, 2013.

[13] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection for discrete sequences: A survey", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 24, No. 5, pp. 823–839, 2012.

[14] L. Akoglu, H. Tong, and D. Koutra, "Graph based anomaly detection and description: a survey", *Data Mining and Knowledge Discovery*, Vol. 29, No. 3, pp. 626–688, 2015.

[15] M. Elik, F. Dadas¸ E. Elik, and A. Dokuz, "Anomaly detection in temperature data using dbscan algorithm", in *International Symposium on Innovations in Intelligent Systems and Applications (INISTA)*, 2011, pp. 91–95.

[16] Y. Lv, T. Ma, M. Tang, J. Cao J, Y. Tian, A. Al-Dhelaan, and M. Al-Rodhaan, "An efficient and scalable density-based clustering algorithm for datasets with complex structures", *Neurocomputing*, Vol. 171, No. 1, pp. 9–22, 2016.

[17] W. Wang, B. Zhang, D. Wang, Y. Jiang, S. Qin, and L. Xue, "Anomaly detection based on probability density function with kullback–leibler divergence", *Signal Processing*, Vol. 126, No. 1, pp. 12–17, 2016.

[18] C. Song, Y. Sun, G. Han, and J. Rodrigues, "Intrusion detection based on hybrid classifiers for smart grid", *Computers & Electrical Engineering*, Vol. 93, No. 4, pp. 285-298, 2021.

[19] J. Gu, and S. Lu, "An effective intrusion detection approach using SVM with naïve Bayes feature embedding", *Computers & Security*, Vol. 103, No. 3, pp. 315–329, 2021.

[20] S. Garg, and S. Batra, "Flexible Subspace Clustering: A Joint Feature Selection and K-Means Clustering Framework", *Big Data Research*, Vol. 23, No. 1, pp. 211-231, 2021.

[21] Y. Tao, Y. Zhang, and Q. Wang, "Fuzzy c-mean clustering-based decomposition with GA optimizer for FSM synthesis targeting to low power ", *Engineering Applications of Artificial Intelligence*, Vol. 68, No. 2, pp. 40-52, 2018.

[22] J. Bezdek, R. Ehrlich, and W. Full, "Fcm: The fuzzy c-means clustering algorithm", *Computers & Geosciences*, Vol. 10, No. 2, pp. 191–203, 1984.

[23] M. Tavallaee, E. Bagheri, W. Lu, and A. Ghorbani, "A Detailed Analysis of the KDD CUP 99 Data Set", In *Proceedings of the 2009 IEEE Symposium on Computational Intelligence*, 2009.

[24] A. Thakkar, and R. Lohiya, " A Review of the Advancement in Intrusion Detection Datasets ", *Procedia Coputer Science*, Vol. 167, No. 2, pp. 636-645, 2020.

Aliakbar Tajari Siahmarzkooh received the B.Sc. degree in Computer Engineering from Ferdowsi University of Iran in 2009, and the M.Sc. and Ph.D. degree in Computer Science from University of Tabriz, Iran in 2012 and 2017, respectively. He has been working with the Department of Computer Sciences, Golestan University, since 2017, where he is now an assistant professor. His current research interests include network security, data mining and artificial intelligence.