DOI: http://dx.doi.org/10.26483/ijarcs.v8i7.4453

ISSN No. 0976-5697

Volume 8, No. 7, July – August 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

ANOMALY BASED IMPROVED NETWORK INTRUSION DETECTION SYSTEM USING CLUSTERING TECHNIQUES

Sunil M. Sangve
Zeal College of Engineering and Research, SP Pune University, Pune-411041, India

Uday V. Kulkarni
SGGSIE&T, SRTMU, Nanded, India

Abstract: The detection of new threats has become a need for secured communication to provide complete data confidentiality. The network requires anomaly detection to shield from hurtful activities. There are various types of metaheuristic methods used for anomaly detection. In this paper, a new approach is proposed for network anomaly detection using multi-start metaheuristic method and enhancement in clustering algorithms. The main stages involved in the proposed approach are: preprocessing, clustering, training dataset selection and the performance evaluation based on training and testing dataset to detect anomalies. The performance of two clustering algorithms, i.e. K-means and expectation maximization (EM) is compared using detection accuracy, false positive rate, and detector generation time. The experimental results are based on NSL-KDD dataset. The results show that the EM clustering performs better than K-means clustering algorithm.

Keywords: Metaheuristic method; K-means clustering; EM clustering; Anomaly Network Intrusion detection system(ANIDS); Genetic Algorithm.

I. INTRODUCTION

The intrusion detection systems (IDSs) are security tools to strengthen the security of communication and information system like other systems such as antivirus software, firewalls and access control schemes [1]. Several approaches have been proposed for IDS. The concept of IDS introduced by Denning [2] and then relevant work was done by Stanford-Chen et al. [3]. According to National Institute of Standards and Technology (NIST), intrusion is an attempt to break confidentiality, availability and integrity (CAI) of network system [4]. The intrusion detection system is a security tool to monitor network traffic to detect unauthorized access. The intrusion prevention system (IPS) is a system which has all features of IDS and it could prevent the computer and network system from intrusion attack [5].

The IDSs are of two types: Host-based (HIDS) and Network-based (NIDS). The NIDS monitors all computer networks through analyzing network traffic. The HIDS used to monitor only individual computer system or host. It analyzes information available on a host like log files and system calls. The IDS is further classified into two types misuse-based (MIDS) and anomaly-based (AIDS). The MIDS depends on the number of rules or patterns or signatures which are written by domain experts. It uses the Snort tool for open source implementation [6]. Snort uses a rule-based language combining signature, protocol and anomaly inspection methods. The anomaly detection based on normal behavior of the subject. It considers any action that deviates from normal predefined threshold value as the intrusion. The MIDSs are used to detect only known attacks and AIDSs are used to detect unknown attacks.

In this paper, to detect novel attacks, an integrated technique using metaheuristic method and enhancement in clustering technique is proposed to improve the performance of ANIDS.

The clustering algorithms are used to partition a large dataset so that processing complexity is reduced. The main challenging task is to reduce the false positive rate and detector generation time to improve detection accuracy as compared to MIDS. Based on search heuristic, the Genetic algorithm (GA) plays important role in the field of Artificial Intelligence. This heuristic is also called as metaheuristic used for making useful solutions to optimization and search problems. The GA belong to evolutionary algorithms (EA), producing solution to optimization problems using techniques inspired by natural evolution like inheritance, mutation, crossover, and selection [7]. The population size is given in initialization step. The population size depends on the nature of problem and contains hundreds or thousands of possible solutions. The initial population generated randomly by taking all ranges of possible solution. In selection step, the proportion of existing population selected during each successive iteration. The individual solutions are selected through fitness function. The fitness function is used to select best individuals. The crossover means recombination to generate new chromosomes and mutation operator. The anomaly detection uses a fitness function to determine time to generate anomaly detectors that gives number of elements covered by detectors in training dataset. The number of detector generations are repeated until it reaches the individual that meets the desired condition [7].

Metaheuristics are strategies that give the search process having aim to explore search area to find the optimal solution [8]. The metaheuristic algorithms range from simple local search procedure to complex learning processes. There are two types of metaheuristic algorithms, single solution based and population based search. The single solution metaheuristic modifies and improves only single candidate solutions. The examples of single solution metaheuristics are simulated annealing, iterated local search, variable neighborhood search, and guided local search. The population based metaheuristic maintains and improves the multiple candidates. The examples of population based metaheuristics are evolutionary computation, genetic algorithms, and particle swarm optimization [9]. Thus, population based metaheuristic are more suitable in anomaly detection than single solution metaheuristic.

The Negative Selection Algorithm (NSA) is a classification algorithm which was first developed by Stephanie Forrest et al. The unlabeled data samples trained from a certain sub-region of the problem domain. These samples are used to check whether or not new unknown data points belong to the same sub-region.

The new variations of NSA are used to improve the algorithmic performance by developing new detector generation scheme [10]. The algorithms that use negative selection based detector generations are generally swarm intelligence and evolutionary computation [9] [11].

The paper is structured as follows. Related work is discussed in section two. The proposed method and implementation is given in section three. The results using proposed method is discussed in section four. The section five concludes the work.

II. RELATED WORK

Machine learning corresponds to optimization and Artificial Intelligence works in a computer task where programming and designing is explicit. It also works when the rule based algorithm is infeasible. A machine learning algorithm is used in modeling and evaluation (for example, fault diagnosis to determine the system whether it is in the normal state or contain several fault state). Thus, in case of ANIDS, it has the ability to modify the way of execution when it finds new information [12]. There are several machine learning techniques applied to anomaly based detection system like Bayesian networks, neural networks, fuzzy logic, outlier detection and genetic algorithm. Several times, machine learning techniques are combined with statistical techniques. Machine learning algorithm provides flexibility, adaptability and capture inter-dependencies.

The various approaches are proposed for anomaly network intrusion detection like statistical based, rule based, state based, and heuristic based approaches. The statistical-based anomaly detection approaches recognize intrusion using the values of predefined threshold, mean, standard deviation, and probabilities [13]. The state- based approaches make use of finite state machine that are derived from network behavior to detect the attack [12]. The rule-based approaches use, the number of rules which are derived like if-then or if-then-else for detecting known intrusions [14] [15]. The heuristic-based approaches are motivated from biological concepts [11].

Adaniya MH AC et al. [16], presented an algorithm named as Digital Signature of Network Segment (DSNS). For detection and characterization of an anomaly, it is very important to know about behavior patterns of the network. Anomalies are responsible to break the network security or decrease the performance of network. Thus, the DSNS algorithm is used to observe network traffic behavior pattern. They [16] also proposed clustering algorithm, K-Harmonic mean (KHM) combined with heuristic approach, called as firefly algorithm. The Firefly Harmonic Clustering Algorithm (FHCA) is used to detect anomalies in network. The experimental results that they have achieved are 80% true positive rate and 20% false positive rate. K-Harmonic mean clustering is an unsupervised classification of patterns. In clustering; the data objects in one cluster represent the similarity.

K-means clustering algorithm is one of the most popular algorithms due to its simplicity and ability to handle large volumes of data. It uses the Euclidean distance to calculate the similarity between objects. But K-means is susceptible to noise and outliers. The KHM is proposed by Zhang et al. [17], the goal is to calculate harmonic mean i. e. the distance between a data object to all the centers.

Gong M. et al. [18], proposed the improved negative selection algorithm, i.e. further training negative selection algorithm (FtNSA). The experimental results are based on synthetic dataset and KDD CUP 99 dataset. The goal of further training is to generate self detectors to occupy own region. The main objective of further training is to decrease the redundant detectors to reduce the computational cost in testing phase and also to improve the self region coverage. The parameter α is used in FtNSA which provides greater flexibility. Instead of using only NSA they used the FtNSA to improve the detector generation rate. However, they need to focus on distribution of detector generation for optimization and reduction in detector overlapping.

The finite state machine approach is introduced in [12], based on Hidden Markov Model (HMM). The HMM used to detect network attacks by observing attacker behavior with the help of a network alert correlation module. The experimental results are based on Lincoln Laboratory 2000 and the DARPA 2000 dataset. The main steps to detect anomalies are data gathering, detection component, alerts optimization, prediction component, and response. Results on DARPA 2000 dataset predict perfectly distributed denial of service attacks and multistep attacks missed by detection component.

Wang SS et al. [14], proposed rule-based approach by combining anomaly and misuse detection in one module to increase detection accuracy and lower the false positive rate. They also used a decision-making model to combine detected results and report the type of attack. The separate modules are designed for separate network devices based on their capabilities and probabilities of attacks they suffer from. M. Saniee Abadeh et al. [19], presented the Neural fuzzy systems and genetic fuzzy systems integrates reasoning method of fuzzy systems with the learning capabilities of neural networks and evolutionary algorithms. They describe a fuzzy genetic-based learning algorithm for detection of intrusions in computer network.

Aziz ASA et al. [20], presented an approach for detecting network traffic anomalies using detectors generated by a genetic algorithm with deterministic crowding Niching technique to improve hyper-sphere detector generations. The experimental results are performed on NSL-KDD dataset with the scope of NSA [21]. The evolutionary algorithm is used to generate detectors. The intrusion data classification is proposed in [9].

Table I: Statistical anomaly detection

References	***Processing Strategy***	***Detection methodology***	***Data set used***	***Network traffic***
Rousseeu P.J et al.[23]	Centralized	Unmasking multivariate outliers	-	-
E. Eskin[24]	Hybrid	Anomaly Detection by using learned probability distribution.	DARPA 99	Packet based
C. Manikopoulos and S.Papavassiliou [25]	Distributed	Network anomaly intrusion detection and fault detection	Real time data	Packet based
Ye N et al. [26]	Centralized	Multivariate statistical analysis for Host based intrusion detection	-	Packet based

The rough set used for feature selection with standard particle swarm intelligence named as a simplified swarm optimization for intrusion data classification, improvement of hyper-spheres detectors, the hyper-ellipsoid detectors and detectors are generated from evolutionary algorithms [22]. As compared to hyper-sphere detectors, the hyper-ellipsoid detectors are more suitable because of stretchiness and re-orientation which is helpful to reduce the redundant non-self-space.

Table I describes the statistical anomaly detection methods and gives processing strategy, detection methodology, dataset used and network traffic. The statistical anomaly detection methods like multivariate statistical analysis and probability distribution use DARPA or real-time dataset.

Table II: Machine Learning Methods

References	*Detection methodology*	*Dataset used*	*Network traffic*
S. C. Lee and D. V. Heinbuch [27]	Neural network based intrusion detection.	Simulated data	Packet based
M. Amini et al. [28]	Unsupervised neural network intrusion detection	KDD Cup 99, real life	Packet based
Liu et al. [29]	Neural network based intrusion detection.	KDD Cup 99	Packet based
R. C.Chen et al. [30]	Rough set and support vector machine is used for	DARPA 98	Packet based

Table II describes the machine learning methods using detection methodology, dataset used and network traffic. The machine learning methods like neural network, support vector machine uses the standard datasets to detect the intrusions.

III. IMPLEMENTATION DETAILS

The novel approach used for anomaly network intrusion detection using clustering algorithms to reduce detector generation time, to increase intrusion detection accuracy and to reduce false positive rate is as shown in Fig. 1. The anomaly network intrusion detection system uses clustering algorithms, GA, and multi-start metaheuristic algorithm to improve the anomaly detection accuracy.

The aim of this paper is to classify anomaly or normal class from input training and testing dataset. The proposed ANIDS having the modules: training dataset as input, pre-processing, clustering algorithms, training dataset selection, detector generation and optimization, performance evaluation on training and testing dataset, and output as a normal or anomaly.

The anomaly detection is measured using two classes i. e. normal and abnormal. The system gives detection accuracy, false positive rate and detector generation time. The ANIDS takes input as training dataset. Preprocessing is performed on training dataset to decrease the processing overhead and overcome the problems like classifier confusion, training overhead, detection and false rate ratios. Followed by preprocessing, clustering technique is used to divide the training dataset into the cluster and reduce the processing and time complexity.

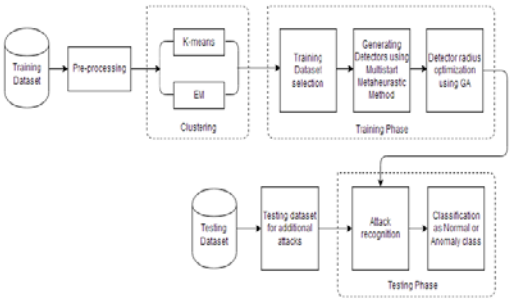

Figure 1. Proposed Anomaly Network Intrusion Detection System.

Clustering is deployed by K-means and EM technique. The multi-start metaheuristic framework is used to select multiple initial start points and to generate detectors which used for detecting anomalies. Later, GA is used to optimize detector radius. The shape of detector is hyper-sphere. At the end of detector radius optimization, the testing dataset is used with additional types of attacks. The performance evaluation of ANIDS is based on training and testing dataset. The output is normal or anomaly. The results are calculated separately using K-means and EM-clustering.

A. Input Training Dataset

In this paper, NSL-KDD dataset is used [31]. This dataset has five classes Normal, Probe, U2R (user to root), R2L (Remote to Local), DoS (Denial of Service). The number of attacks in training dataset is twenty three and additional fourteen types of attacks are included in testing file. The training dataset is denoted as 'T'.

B. Preprocessing

The preprocessing applied on training dataset (T). The data preprocessing is required to remove insignificant data or words which are not useful for extracting the features. Another main advantage of data preprocessing is that the processing time will decrease when unwanted features are removed. The following example describes how preprocessing applied on the training dataset:

Let, consider the one vector from T, {0,tcp,ftp_data,SF,491,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,2,2, 0.00,0.00,0.00,0.00,1.00,0.00,0.00,150,25,0.17,0.03,0.17,0.00,0 .00,0.00,0.05,0.00,normal}

When we apply the preprocessing on above single vector, the insignificant words like tcp, ftp_data, SF are removed to decrease the processing time further. After preprocessing, we get a vector contained only numeric value that is our interest. The last word in vector denotes the class normal or anomaly. Therefore, now we obtain the vector which has two important features i. e pattern for different types of class in numeric form and class name 'normal'. The obtained vector after preprocessing is:

{491,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,2,2,0.00,0.00,0.00,0.0 0,1.00,0.00,0.00,150,25,0.17, 0.03, 0.17, 0.00, 0.00, 0.00, 0.05, 0.00}

For normalization of training dataset, following steps are applied [32]:

$$T_{norm} = \begin{cases} \dfrac{T - \mu_T}{\sigma_T}, \sigma_T \neq 0 \\ T - \mu_T, \sigma_T = 0 \end{cases} \quad (1)$$

Where, $T = \{ x_{i,j} \mid i = 1,2,3,..., m \text{ and } j = 1,2,3,..., n \}$

$\mu_T = \{ \mu_j \mid j = 1,2,3,....., n \}$

$\sigma_T = \{ \sigma_j \mid j = 1,2,3,....., n \}$

T is of having attributes. These attributes consists of n column attributes with m samples. $x_{i,j}$ is the j[th] column attribute in *i*[th] sample, μ_T and σ_T are $1 \times x$ matrix which are mean and standard deviation respectively for each of the n attributes. Testing dataset (TS) is used to measure detection accuracy that is normalized using the μ_T and σ_T as follows [32]:

$$TS_{norm} = \begin{cases} \dfrac{TS - \mu_T}{\sigma_T}, \sigma_T \neq 0 \\ TS - \mu_T, \sigma_T = 0 \end{cases} \quad (2)$$

C. *Clustering*

Clustering algorithm is used to reduce the training dataset, decreasing processing complexity and time complexity. The processed training dataset divided into a number of clusters. For clustering, we have used two techniques K-means and EM-clustering and compared their results.

1. *The K-Means Algorithm*

In this paper, we are grouping the data instances using an unsupervised learning algorithm called K-means which are the simplest centroid based clustering algorithm [33].

The K-means begins with the initialization of cluster centroids. The iterative nature of K-means works in two phases 1. Cluster assignment 2. Centroid shift. In cluster assignment, each instance of the dataset is considered and compared with nearest cluster. In Centroid shift, K-means algorithm calculates the average of data instances in the cluster and shift the centroid to the average value calculated. The two-step process of K-means is repeated until there is no change in the cluster.

The algorithm aims to minimize squared function which is an objective function in this paper. The function is represented as:

$$J(v) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} (\| x_i - y_j \|)^2 \quad (3)$$

Where, $\| x_i - y_j \|$ is the Euclidean distance between x_i and y_j.

c_i is the number of data instances in the cluster i.

c is the total cluster centers.

The identified clusters help to identify the attribute that belongs to normal or anomaly categories

2. *The Expectation Maximization Algorithm*

1. The EM can be divided into two important steps which are Expectation (E-step) and Maximization (M-step).
2. The goal of E-step is to calculate the expectation of the likelihood (the cluster probabilities) for each instance in the dataset and then re-label the instances based on their probability estimations.
3. The M-step is used to re-estimate the parameters values from the E-step results.
4. The outputs of M-step (the parameters values) are then used as inputs for the following E-step.
5. These two processes are performed iteratively until the results convergence

In this clustering technique, we randomly initialize the parameter values. The E- step is used to assign the values to hidden variables and M-step used to compute parameters based on fully observed data.

D. *Detector Generation using Metaheuristic Algorithm*

Training dataset is divided into number of clusters. The multi-start strategy is applied to select multiple initial start points from clustered dataset. The training dataset is selected with good representative of original dataset. The selected training dataset samples are used to initialize multiple start points. The initial start points are denoted as 'isp'. The isp is selected randomly from T samples and distributed over clusters. The multi-start framework is used for generating detectors. Thus, there are two boundaries in solution space upper and lower. Here we are considering the hyper-sphere shape for calculation of detector radius. The detector radius is denoted as R and given as follows [32]:

The detector radius $R = \{ r \in R \mid 0 < r \leq hpu \}$ where hpu is the hyper-sphere radius upper bound. Thus,

$U_j = \max(x_{ij}) \; where \; i = 1,2,3,..., m.$

$L_j = \min(x_{ij}) \; where \; i = 1,2,3,..., m.$

UB –Upper bound and LB - Lower bound are used for solution space.

$UB = (u_1, u_2, u_3, ...u_n, hpu)$

$LB = (l_1, l_2, l_3, ..., 0)$

The detectors $D = \{ d_1, d_2, d_3, ..., d_{isp} \}$

The solution space obtained by multi-start framework is calculated as:

$D_i = (u_{i1}, u_{i2}, u_{i3}, ..., u_{in}, r_i)$ where hyper-sphere center is at $D_{center} = (u_{i1}, u_{i2}, ..., u_{in})$ and hyper sphere radius is r_i.

The objective function to control the detector generation process is:

$$F(D_i) = N_{abnormal}(d_i) - N_{normal}(d_i) \quad (4)$$

Where,

$N_{abnormal}(d_i)$ – is number of abnormal samples covered by detector d_i.

$N_{normal}(d_i)$ – is number of normal samples covered by detector d_i.

Anomaly detection is done from the generated detectors. The following rule is used,

$If \; (dist(D_{center}, x) \leq r) \; then \; \{normal\} \; else \; \{abnormal\}$

Where r is the detector hyper-sphere radius and $(dist(D_{center}, x))$ is the Euclidean distance between detector hyper sphere center D_{center} and test samples x.

E. Detector Radius Optimization using Genetic Algorithm(GA)

In the training data set selection we divide training dataset into the different cluster and from each cluster select the training dataset sample. A GA is used for detector generation which is focused on the non-overlapping of hyper-sphere detectors to gain the maximal non-self-space coverage by using a fitness function which is based on detector radius. In detector generation the normal behavior of the patterns are called 'self'. This algorithm defines 'self' as normal behavior patterns of a monitored system. It generates a number of random patterns that compared to each self-defined pattern. If any randomly generated pattern matches with the self-pattern, this pattern fails to become a detector and thus it is removed. Otherwise, it becomes a 'detector' pattern.

The initial population each detector radius is initialized to its value generated by the multi-start algorithm. The fitness function used to optimize detector radius is

$$F(r_i) = N_{abnormal}(r_i) - N_{normal}(r_i) \quad (5)$$

Where, normal and abnormal sample covered by detector using ri as its radius

F. Detector Reduction

The number of detector is reduced to improve effectiveness and speed of anomaly detection. There are two levels of reduction step:

1. If $N_{abnormal}(D_i) > thr_{\max abnormal}$ *or* $N_{normal}(D_i) < thr_{\min normal}$ then remove detectors D_i, $D_i \in D$.
2. The second level of reduction to remove any detector D_i, if it N_{normal} is covered by one or more detectors with a percent equal or more than $thr_{intersect}$.

G. Testing Data Set

The Testing dataset is given as input after completion of training stage. In training stage we got the preprocessed dataset. The testing dataset is normalized by using training dataset.

H. Performance Evaluation

The performance evaluation of system is calculated by using both training and testing dataset.

I. Output

This is the final step where we obtain normal or anomaly class.

IV. RESULTS AND DISCUSSION

For experimental setup, we use the Windows 7 operating system, Intel i5 processor, 512MB RAM, 80GB Hard disk, Net Beans IDE 8 + JDK tool. To calculate the results, NSL KDD data set is used [31]. In training dataset, there are 23 types of attack and in testing phase additional 14 attacks are included. Using this dataset, we look for detection accuracy, false positive rate, detector generation time. The detection accuracy is calculated as

$$DetectionAccuracy = \frac{TP + TN}{(TP + TN + FP + FN)} \quad (6)$$

Where,

TP –True Positive

TN –True Negative

FP –False Positive

FN – False Negative

The False Positive Rate (FPR) is calculated as follows:

$$FalsePositiveRate = \frac{FalsePositive}{(FP + TN)} \quad (7)$$

Where, true negative means anomaly samples are classified correctly as anomalous and false positive means normal samples are detected as anomaly.

The training dataset samples are 3000, 5000, 8000, 10000, 15000, 20000. These training dataset samples are used to calculate detection accuracy (DA), false positive rate (FPR) and detector generation time (DGT). For example, here we have taken training dataset size 5000 and calculated results for DA, FPR, and DGT which are shown in Fig. 2, 3, and 4 respectively.

Using EM clustering approach, we attained the DA 98.28%, FPR 0.004 and DGT is 109 seconds. The result shows that EM approach gives better result than K-means algorithm.

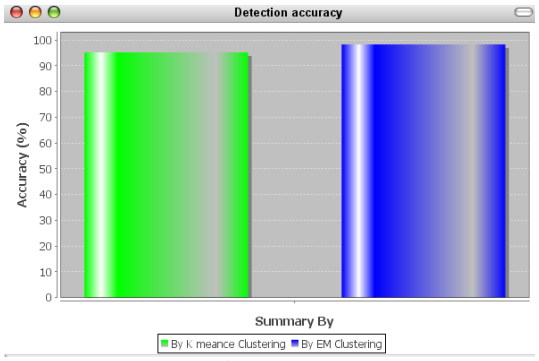

Figure 2. Detection Accuracy for Dataset 5000 using K-means and EM clustering.

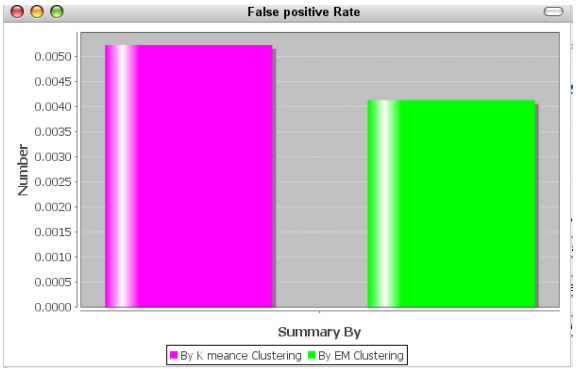

Figure 3. False Positive Rate for Dataset 5000 using K-means and EM clustering.

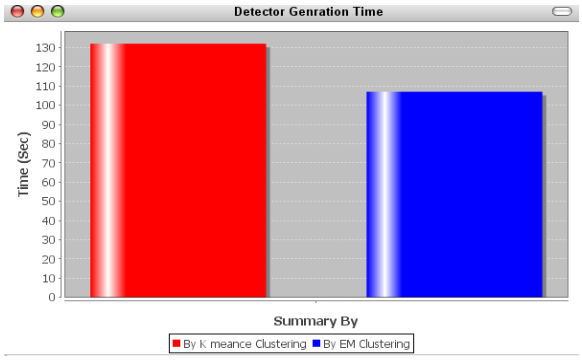

Figure 4. Detector Generation Time for Dataset 5000 using K-means and EM clustering

A. *Scenario 1*

Dataset size = 3000, 5000, 8000, 10000, 15000, 20000.
The training dataset size is 3000. The result for DA is calculated using K-means and EM clustering. Table III gives the comparative result for detection accuracy using K -means and EM clustering.

Table III: Detection Accuracy (DA)

Dataset Size	*Number Of Clusters*	*DA Using K-means (%)*	*DA Using EM-clustering (%)*
3000	100	94.97	95.70
5000	100	95.18	98.28
8000	100	95.18	98.18
10000	100	94.97	98.17
15000	100	95.18	98.27
20000	100	92.70	96.73

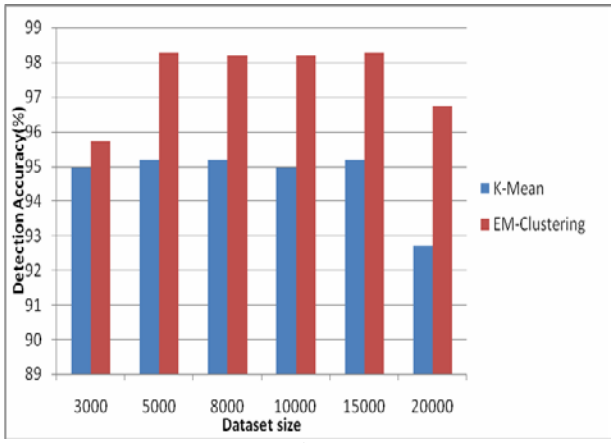

Figure 5. Comparative graph for Detection Accuracy (DA) using K-means and EM-clustering technique.

Fig. 5 shows comparative graph for DA using K-means and EM-clustering technique, which shows that detection accuracy is higher when using EM-clustering than K-means clustering. The number of detectors required to cover the normal samples increases with the training dataset size resulting in a higher processing time. While creating more number of clusters, the number of detectors is also increased and there is small effect on detection accuracy.

B. *Scenario 2*

Dataset size = 3000, 5000, 8000, 10000, 15000, 20000.
Table IV gives the FPR calculated using K-means and EM clustering for various dataset size.

Table IV: False Positive Rate (FPR)

Dataset Size	*FPR Using K-means*	*FPR Using EM-clustering*
3000	0.0080	0.005
5000	0.0052	0.0041
8000	0.0055	0.0043
10000	0.0080	0.0106
15000	0.0052	0.0041
20000	0.0057	0.0106

The comparative graph for FPR using K-means and EM-clustering technique is shown in fig. 6. The FPR is minimum in EM-clustering for training dataset size 5000 and 15000 i.e. 0.0041. The maximum FPR in EM-clustering for training dataset size 10000 and 20000 i.e 0.0106. The FPR calculated using K-means is minimum for training dataset 5000 and 15000 i.e. 0.0052. The maximum FPR using K-means is 0.0080 at training dataset size 3000 and 10000. As the training dataset increased, there have some effects on FPR also. EM clustering gives more FPR for10000 and 20000 training dataset as compared to K-means clustering.

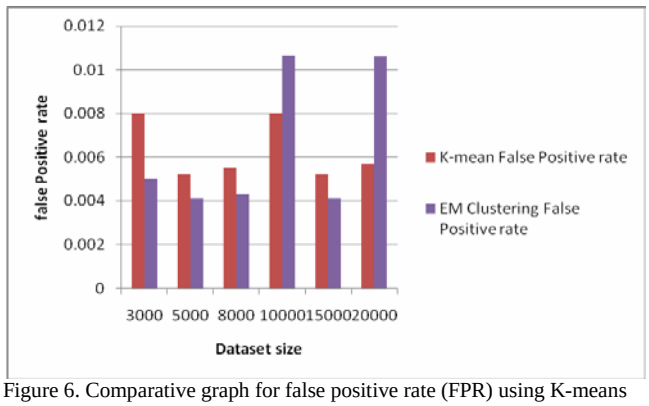

Figure 6. Comparative graph for false positive rate (FPR) using K-means and EM-clustering techniques.

C. *Scenario 3*

Dataset size = 3000, 5000, 8000, 10000, 15000, 20000.
DGT using K-means and EM clustering technique is presented in Table V.

Table V: Detector Generation Time (DGT)

Dataset Size	*DGT using K-means (second)*	*DGT using EM-clustering (second)*
3000	131	102
5000	134	109
8000	135	110
10000	135	112
15000	136	127
20000	137	130

The comparative graph for DGT using K-means and EM-clustering technique is shown in Fig. 7. Our ANIDS system gives less detector generation time as compared to K-means

technique. The values of DGT increases as increase in training dataset size. Thus, DGT is directly proportional to the training dataset size. The maximum DGT is required for 20000 training dataset size. If the training dataset size is large, the number of rules are increased and time required to generate detector is also more. The DGT will increase to cover maximum normal training samples and because of this processing time will also increase.

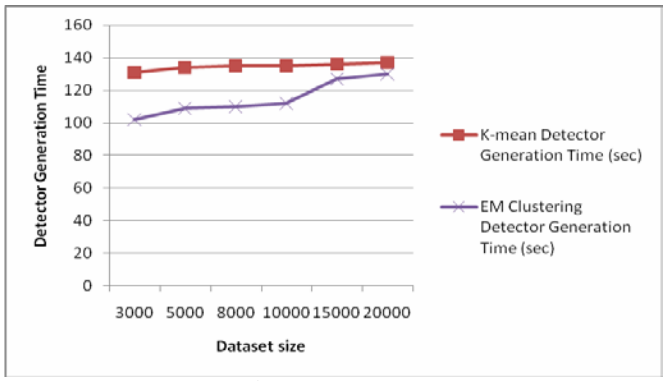

Figure 7. Comparative graph for detector generation time (DGT) using K-means and EM- clustering technique.

V. CONLCUSION

The Anomaly Network Intrusion detection System with metaheuristic method using K-means and EM clustering is proposed. The clustering technique is used to divide the training dataset to decrease the processing and time complexity. The ANIDS with multi-start metaheuristic method and a GA used to remove redundant detectors. It minimizes the number of generated detectors and thus reduces the time needed later for anomaly detection. The proposed approach is used to decrease the DGT, FPR and to increase the DA. The parameters like number of clusters, training dataset size, detector radius limit plays a vital role in anomaly detection. The experimental results are based on NSL KDD dataset which is a large scale dataset. The system compares the K-means and EM-clustering to improve the DA and to minimize the FPR and DGT. The result shows that:

1. EM clustering gives better DA over K-means clustering technique. The maximum DA obtained by EM clustering at 5000 training dataset size is 98.28% whereas K-means clustering gives 95.18%.
2. The minimum FPR obtained using K-means clustering at 5000 and 15000 training dataset size is 0.0052 whereas EM clustering gives 0.0041.
3. The DGT is increased as training dataset size increases. The EM-clustering requires less DGT as compared to K-means clustering.

In future, the system will be implemented using another standard dataset either real- time or non-real time. By reducing the offline processing time overhead, the online processing time will be minimized.

VI. REFERENCES

[1] Garcıa-Teodoro P, Dıaz-Verdejo J, Macia-Fernandez G, "Anomaly-based network intrusion detection: techniques, systems and challenges," Computer Security, 2009;28(1–2):pp.18–28.

[2] Denning ED, "An intrusion-detection model," IEEE Transactions on Software Engineering, 1987; 13(2):pp. 222–32.

[3] Staniford-Chen S., Tung B., Porrar P., Kahn C., Schnackenberg D., Feiertag R., "The common intrusion detection framework data formats," 1998, Internet draft 'draft-Stanford-cidf-dataformat00.txt'.

[4] R Bace, P Mell, "Intrusion detection systems," National Institute of Standards and Technology (NIST), Technical Report 800-31, 2001.

[5] Stavroulakis P, Stamp M, "Handbook of information and communication security," New York: Springer-Verlag, 2010.

[6] Roesch M, "Snort-lightweight intrusion detection for networks," In: Proceedings of the 143th USENIX Conference on System Administration, Seattle, Washington; 1999. pp. 229-238.

[7] Genetic algorithm. [Online] 2013.http://en.wikipedia.org/wiki/Genetic_algorithm

[8] The metaheuristic method available at http://en.wikipedia.org/wiki/Metaheuristic#/media/File:Metaheuristics_classification.svg

[9] Chung YY, Wahid N, "A hybrid network intrusion detection system using simplified swarm optimization (SSO)," Applied Soft Computing, 2012;12(9): pp. 3014–22.

[10] Dasgupta D, Yu S, Nino F, "Recent advances in artificial immune systems: modelsandapplications," Applied Soft Computing, 2011; 11 (2): pp.1574–87.

[11] Abadeh MS, Mohamadi H, Habibi J, "Design and analysis of genetic fuzzy systems for intrusion detection in computer networks," Expert System Applications, 2011; 38(6): pp. 7067–75.

[12] Shameli Sendi A, Dagenais M, Jabbarifar M, Couture M, "Real time intrusion prediction based on optimized alerts with hidden Markov model ," JNW, 2012;7(2): pp.311–21.

[13] Xu X, "Sequential anomaly detection based on temporal difference learning: principles, models and case studies," Applied Soft Computing, 2010; 10(3): pp. 859–67.

[14] Wang SS, Yan KQ, Wang SC, Liu CW, "An integrated intrusion detection system for cluster- based wireless sensor networks," Expert System Applications, 2011; 38(12): pp. 15234–43.

[15] KartitA, Saidi A, Bezzazi F, El MarrakiM, Radi A, "A new approach to intrusion detection system,"JATIT, 2012; 36(2): pp.284–90.

[16] Adaniya MH AC, Lima MF, Rodrigues JJPC, Abraa˜o T, Jr. MLP, "Anomaly detection using DSNS and firefly harmonic clustering algorithm," In: IEEE international conference on communications (IEEE ICC 2012), Ottawa, Canada; 2012. pp.10–5.

[17] B. Zhang, M. Hsu, and U. Dayal, "K-harmonic means - a data clustering algorithm," Hewlett- Packard Laboratories, Palo Alto, Tech. Rep. HPL-1999-124, Outubro 1999.

[18] Gong M, Zhang J, Ma J, Jiao L., "An efficient negative selection algorithm with further training for anomaly detection," Knowledge-Based System, 2012; 30: pp.185–91.

[19] M. Saniee Abadeh, J. Habibi, C. Lucas, "Intrusion detection using a fuzzy genetics-based learning algorithm," Journal of Network and Computer Applications, 30 (2007) pp.414–428.

[20] Aziz ASA, Salama M, ellaHassanien A, El-Ola Hanafi S., "Detectors generation using genetic algorithm for a negative selection inspired anomaly network intrusion detection system," In: FedCSIS proceedings of federated conference on computer science and information systems; Wroclaw: IEEE, 2012. pp. 597–602.

[21] Wang D, Zhang F, Xi L, "Evolving boundary detector for anomaly detection," Expert System Applications, 2011; 38(3): pp. 2412–20.

[22] Shapiro JM, Lamont GB, PetersonGL, "Anevolutionary algorithm to generate hyper- ellipsoid detectors for negative selection," In: Beyer HG, editor. GECCO '05.Proceedings of the 2005 conference on Genetic and evolutionary computation. NewYork, NY, USA: ACM; 2005. pp. 337–44.

[23] Rousseau P.J., Van Zomeren B.C., "Unmasking multivariate outliers and leverage points,"Journal of the American Statistical Association, Vol. 85 (411), 1990, pp. 633-651

[24] E. Eskin, "Anomaly detection over noisy data using learned probability distributions," in Proc. 7th International Conference on Machine Learning, Morgan Kaufmann, 2000, pp. 255– 262.

[25] C. Manikopoulos and S.Papavassiliou, "Network Intrusion and Fault Detection: Statistical roach," IEEE Commun. Mag., vol. 40, no. 10, October 2002, pp.76–82.

[26] Ye N, Emran SM, Chen Q, Vilbert S, "Multivariate statistical analysis of audit trails for host- based intrusion detection," IEEE Transactions on Computers2002.

[27] S. C. Lee and D. V. Heinbuch, "Training a neural-network based intrusion detector to recognize novel attacks," IEEE Trans. Syst. Man Cybern. A, vol. 31, no. 4, 2001, pp.294–299.

[28] M. Amini, R. Jalili, and H. R. Shahriari, "RT-UNNID: A practical solution to real-time network-based intrusion detection using unsupervised neural networks," Computers & Security, vol. 25, no. 6, 2006, pp. 459–468.

[29] [29] G. Liu, Z. Yi, and S.Yang, "A hierarchical intrusion detection model based on the PCA neural networks," Neurocomputing, vol. 70, no. 7-9, 2007, pp.1561–1568.

[30] [30] R. C. Chen, K. F. Cheng, Y. H. Chen, and C. F. Hsieh, "Using Rough Set and Support Vector Machine for Network Intrusion Detection System," In Proc. First Asian Conference on Intelligent Information and Database Systems.Washington, DC, USA: IEEE Computer Society, 2009, pp. 465–470.

[31] The NSL-KDD dataset. The available World Wide Web is http://nsl.cs.unb.ca/NSL-KDD/

[32] Tamer F.Ghanem, Wail S. Elkilani, Hatem, "A hybrid approach for efficient anomaly detection using metaheuristic methods," Journal of advanced research,2014.

[33] Tapas Kanungo, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman and Angela Y.Wu. "An Efficient k-means Clustering Algorithm: Analysis and Implementation," IEEE Transactions on pattern analysis and machine intelligence, vol.24, No.7, July 2002, pp.881-892.

[34] The EMclustering available at https://en.wikipedia.org/wiki/Expectation%E2%80%93maximization_algorithm

