A Hybrid Clustering Algorithm (KM-GSA) Based on Anomaly Intrusion Detection Method

Innovation: International Journal of Applied Research; ISSN: 2347 9272 (Volume-5, Issue-3)

Research Article

June 2017

A Hybrid Clustering Algorithm (KM-GSA) Based on Anomaly Intrusion Detection Method

Salima Omar

Department of Computer Science, University of Benghazi, Benghazi, Libya

omqsalima@gmail.com

Abstract— Detection effectiveness of an IDS is characterized by high detection accuracy, high detection rate and low false positive rate. Many existing Anomaly-based Intrusion Detection Systems (AIDS) are ineffective and fail to distinguish between normal and abnormal data. This affects the detection accuracy and generates a high false alarm rate. Various clustering techniques have been used for Intrusion Detection for identifying anomalous events. The most important advantage of the clustering method is the ability to find unknown attacks that have not been previously detected. In this paper, a hybrid clustering algorithm based on the combination of the k-Means and Gravitational Search Algorithms (KM-GSA) is designed. In the KM-GSA, the GSA is used to solve the clustering problem by refining the clusters formed by the KM algorithm. The KDD 99 data set consisting of five traffic classes was used as the training and testing dataset. The result of the KM-GSA was compared against the results of the KM algorithm and GSA.

Keywords— Network Intrusion Detection, Gravitational Search Algorithm (GSA), K-Means (KM)

I. INTRODUCTION

Protecting information in an organization is crucial due to the continuous increase of network attacks. In effect, the field of information assurance and security has become an important research field in networked and distributed information sharing environments. Information assurance and security involves all the efforts and methods made to protect and secure information whether in memory, processing or in network transactions. Studies covering the prevention, detection and forensic aspect of computer network attacks have long been conducted. Prevention techniques such as encryption, the virtual private network and firewalls alone seem to be inadequate. These measures reduce exposure rather than monitor or eliminate vulnerabilities in computer systems. It is important to have a detecting and monitoring system to protect important data. The importance of safeguarding networks against confidentiality, integrity and availability breaches is an important issue, and intrusion detection plays a vital role in ensuring a secure network. An Intrusion Detection System (IDS) is very important to safeguard computer networks against confidentiality, integrity and availability breaches. Detection effectiveness of an IDS is characterized by high detection accuracy, high detection rate and low false positive rate. Many existing Anomaly-based Intrusion Detection Systems (AIDS) are ineffective and fail to distinguish between normal and abnormal data. This affects the detection accuracy and generates a high false alarm rate. Unsupervised methods are promising approaches to detecting new attacks. The clustering method is used in an unsupervised scheme as a machine learning mechanism. It is used to discover patterns that deal with unlabelled data with many different dimensions. In the clustering process, the members belonging to the same cluster are similar and are different to the members of a different cluster. The most significant advantage of using the clustering method is the capability to discover new attacks which have not been previously detected. Clustering algorithms can cluster new data examples into intelligible groups which can be used to improve the effectiveness of existing classifiers [1].

In this paper, Hybrid KM-GSA is designed to enhance the quality of clustering results. The system implements the KM-GSA by incorporating the k means algorithm to generate the initial population for the GSA and uses GSA as the clustering algorithm to refine the clusters formed by the k-Means algorithm (KM).

The rest of the paper is organized as follows: Section 2 discusses the related works on the hybrid approach in IDS. In section 3 present a brief overview of the gravitational search algorithm to provide a proper background. Section 4 and 5 present proposed approach and data used. Section 6 describes the flow of the experiment. The results and discussion of findings are presented in Section 7. Finally, Section 8 concludes the paper.

II. RELATED WORK

Due to the importance of clustering many researchers have devoted time to design new algorithms as well as to improve the existing algorithm's performance and clustering quality by new meta-heuristic approaches.

Van der Merwe and Engelbrecht [2] proposed a new approach based on the KM and PSO algorithms using two approaches. The approach consists of two steps. In the first step, the PSO is executed to cluster the dataset to find the near optimal centroids for the desired clusters. Then, KM clustering is used, and the result of the KM algorithm is used for seeding the initial swarm, which will be applied by the PSO algorithm. In the second step, PSO is used as the clustering algorithm to refine the clusters formed by the KM.

Hatamlou et al. [3] proposed the GSA as the clustering algorithm to cluster the dataset. The GSA was tested on several standard datasets. The performance of the GSA was compared with three other clustering algorithms, namely, the KM, GA and PSO algorithm. The experimental results indicated that the GSA found higher quality clusters compared to the other algorithms.

Saini and Kaur [4] proposed a new hybrid clustering approach which is a combination of two PSO and KM algorithms to achieve better clustering results. The PSO clustering algorithm is executed to search for the location of the clusters centroid. Then, the result of PSO algorithm is used Due to the importance of clustering many researchers have devoted time to design new algorithms as well as to improve the existing algorithm's performance and clustering quality by new meta-heuristic approaches.

Van der Merwe and Engelbrecht [2] proposed a new approach based on the KM and PSO algorithms using two approaches. The approach consists of two steps. In the first step, the PSO is executed to cluster the dataset to find the near optimal centroids for the desired clusters. Then, KM clustering is used, and the result of the KM algorithm is used for seeding the initial swarm, which will be applied by the PSO algorithm. In the second step, PSO is used as the clustering algorithm to refine the clusters formed by the KM.

Hatamlou et al. [3] proposed the GSA as the clustering algorithm to cluster the dataset. The GSA was tested on several standard datasets. The performance of the GSA was compared with three other clustering algorithms, namely, the KM, GA and PSO algorithm. The experimental results indicated that the GSA found higher quality clusters compared to the other algorithms.

Saini and Kaur [4] proposed a new hybrid clustering approach which is a combination of two PSO and KM algorithms to achieve better clustering results. The PSO clustering algorithm is executed to search for the location of the clusters centroid. Then, the result of PSO algorithm is used.

III. PROPOSED APPROACH

KM is a classical clustering algorithm. It is a simple and efficient algorithm that is commonly used for data clustering. However, its performance depends on the selection of the initial centroids and may lead to empty clusters, which are ineffective for the classification process [6]. The GSA is an effective method for searching problem space to find a near optimal solution. The GSA is suitable for IDSs and has the capability to improve the clustering process [2, 3, 4]. In this research, a hybrid clustering algorithm based on the combination of the KM algorithm and GSA is designed to enhance the capability of the ensemble clusters. In KM-GSA, the GSA is used to solve the clustering problem by refining the clusters formed by the KM algorithm. The main purpose of the KM-GSA is to enhance the quality of the clustering result by incorporating the KM algorithm in generating the initial population for the GSA. The KM-GSA consists of three steps. The KM algorithm is executed on the selected dataset to find the near optimal centroids for the desired clusters. Then, the result of the KM algorithm is used for seeding the initial swarm, which will be applied by the GSA. Finally, the GSA is used as the clustering algorithm to refine the clusters formed by the KM algorithm. A number of cluster centroids for KM-GSA needs to be determined before the algorithm is applied. The KM-GSA hybrid algorithm is illustrated in Fig 1, and the steps of the KM-GSA algorithm are explained in the Algorithm 1.

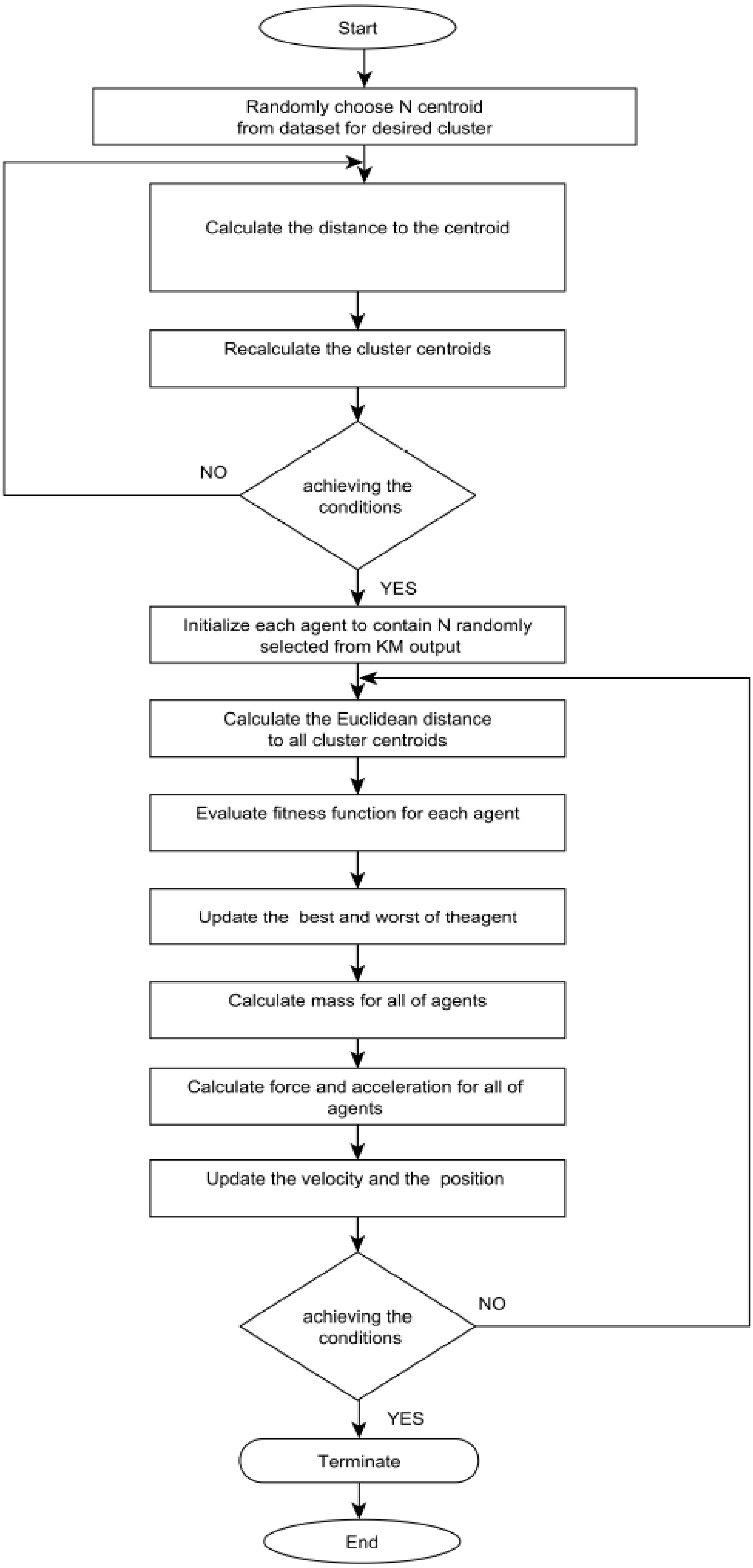

Start
Randomly choose N centroid from dataset for desired cluster
Calculate the distance to the centroid
Recalculate the cluster centroids
NO
achieving the conditions
YES
Initialize each agent to contain N randomly selected from KM output
Calculate the Euclidean distance to all cluster centroids
Evaluate fitness function for each agent
Update the best and worst of theagent
Calculate mass for all of agents
Calculate force and acceleration for all of agents
Update the velocity and the position
achieving the conditions
NO
YES
Terminate
End

Fig. 1.KM-GSA Hybrid Algorithm

In KM-GSA, the KM algorithm firsts selects K data randomly as the initial cluster center. Then, it assigns each point of data in the space to the cluster centroid that has the nearest distance to the point of data using Equation 7:

$$d(y_m, z_{ji}) = \sqrt{\sum_{i=1}^{d} d(y_{mi} - z_{ji})^2} \qquad (7)$$

where d (y_m, z_j) is a space between the point of data and the center of the cluster. To calculate the new cluster centroid in the KM-GSA, Equation 8 is used:

$$z_j = \frac{1}{n_j} \sum_{y_m \in c_j} y_m \qquad (8)$$

where C_j is the group of data samples related to the j_{th} cluster, and the number of data samples is defined as n_j. This process continues until a termination condition is met when the maximum number of iterations is exceeded or the value of the fitness function converges. Then, the GSA is implemented. The GSA generates an initial population of agents selected from the KM algorithm result randomly. Each agent is a possible solution for the clustering problem. In the state of clustering, each agent is used as the centroid of clusters. The fitness of the agents for the KM-GSA is determined according to the clustering criteria using Equation 9:

$$f(i) = \frac{\sum_{j=1}^{K} \sum_{y_m \in c_{ij}} d(y_m - z_{ij})^2}{N_m} \qquad (9)$$

where N is the Number of clusters, Nm is the number of data samples used as inputs to the clustering process, and z_{ij} is the j_{th} cluster centroid in solution offered by the i_{th} agent. The mass of each agent in the KM-GSA is determined according to its fitness using Equation 10:

$$M_i(t) = \frac{fit_i(t) - worst(t)}{\sum_{j=1}^{N} (fit_j(t) - worst(t))} \qquad (10)$$

where $fit_{i(t)}$ is the fitness value of the agent i at iteration t. The worst (t) is the maximum value of the fitness function, and the best (t) is the minimum value of the fitness function. Next, the KM-GSA first computes the output of all forces which act on certain agents by all other agents, then computes the acceleration of the agent using Equations 11 and 12:

$$F_i^d(t) = \sum_{j \in kbest, j \neq i} rand_j G(t) \frac{M_j(t) M_i(t)}{R_{i,j}(t) + \varepsilon} (x_j^d(t) - x_i^d(t) \qquad (11)$$

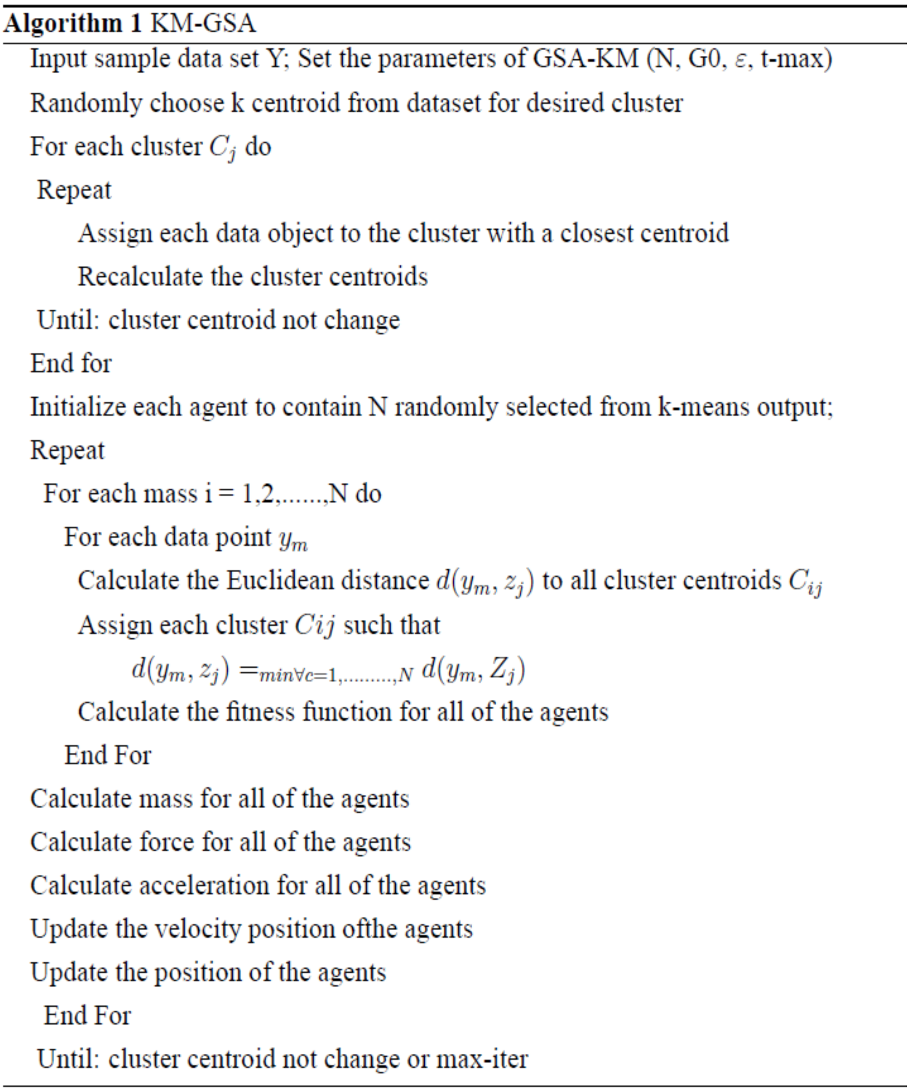
$$a_i^d(t) = \frac{F_i^d(t)}{M_i(t)} = \sum_{j \in kbest, j \neq i} rand_j G(t) \frac{M_j(t)}{R_{i,j}(t) + \varepsilon} (x_j^d(t) - x_i^d(t) \qquad (12)$$

where R_{ij} is Euclidian distance between two agents, and X_i and X_j are the positions of i_{th} and j_{th} agents, respectively. "is very small constant to avoid division by zero. In the KM-GSA, the new velocity is updated; then, the next position of a particle is updated, which indicates the cluster centers by adding the new velocity. The new velocity and position of every agent are updated using Equations 13 and 14:

$$v_i^d(t+1) = Rand_i \times v_i^d(t) + a_i^d(t) \qquad (13)$$

$$x_i^d(t+1) = x_i^d(t) + v_i^d(t+1) \qquad (14)$$

This process continues until the termination condition is met when the maximum number of iterations is exceeded or the value of the fitness function converges.



Algorithm 1 KM-GSA

```
Input sample data set Y; Set the parameters of GSA-KM (N, G0, $\varepsilon$, t-max)
Randomly choose k centroid from dataset for desired cluster
For each cluster $C_j$ do
 Repeat
    Assign each data object to the cluster with a closest centroid
    Recalculate the cluster centroids
 Until: cluster centroid not change
End for
Initialize each agent to contain N randomly selected from k-means output;
Repeat
  For each mass i = 1,2,......,N do
     For each data point $y_m$
       Calculate the Euclidean distance $d(y_m, z_j)$ to all cluster centroids $C_{ij}$
       Assign each cluster $Cij$ such that
            $d(y_m, z_j) =_{min\forall c=1,.........,N} d(y_m, Z_j)$
       Calculate the fitness function for all of the agents
     End For
Calculate mass for all of the agents
Calculate force for all of the agents
Calculate acceleration for all of the agents
Update the velocity position ofthe agents
Update the position of the agents
  End For
 Until: cluster centroid not change or max-iter
```

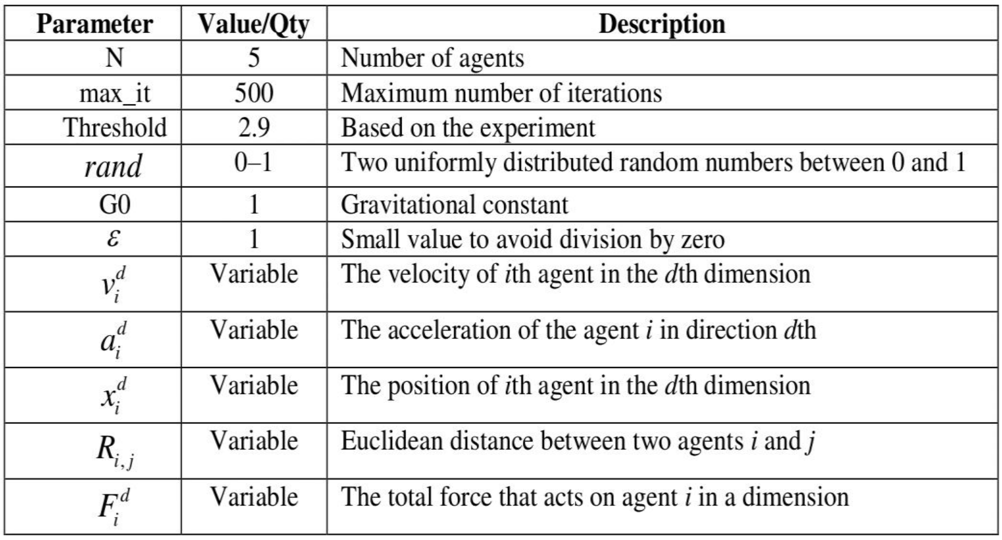
IV. EXPERMENT DATA

The KDD Cup1999 dataset was obtained from the 1998 DARPA Intrusion Detection Evaluation Program and prepared by MIT Lincoln Labs. It is the largest publicly available sophisticated benchmark for researchers to evaluate intrusion detection algorithms or machine learning algorithms. The KDD Cup 1999 dataset contains nine weeks of raw transmission control protocol (TCP) dump data from simulated US Air Force local area network which is injected with multiple attacks. Each TCP/IP connection has a total of 41 qualitative and quantitative features where some are derived features. Features were labelled from 1 to 41 and they are termed as f1, f2, f3,… and f41. The type of attacks belongs to four main categories, namely, Denial of Service (DoS), Remote to Local (R2L), User to Root (U2R) and Probing. This study, as in most of the research in the literature, used the 10 % version of the dataset consisting of 494,020 traffic connections with a similar ratio of attacks as in the full dataset [7, 8].

V. EXPERMENTAL SETUP

The training and testing data used in this study was comprised of 5,092 and 6,890 records respectively. The composition of these sample data maintains the actual distribution of KDD Cup 1999 data. In this paper, the experiments were performed separately for all four attack classes (probe, DoS, R2L and U2R) by randomly selecting data corresponding to that particular attack class and normal data only. Data scaling was done to ensure the training dataset was within the range of [0,1]. the number of iterations was 500 iterations and all the experiments were repeated 500 times (iterations) and the results were averaged. The parameter settings used in the experiment are shown in Table 1.

Table 1 Key parameter values used in KM-GSA



Parameter	Value/Qty	Description
N	5	Number of agents
max_it	500	Maximum number of iterations
Threshold	2.9	Based on the experiment
$rand$	0–1	Two uniformly distributed random numbers between 0 and 1
G0	1	Gravitational constant
ε	1	Small value to avoid division by zero
v_i^d	Variable	The velocity of *i*th agent in the *d*th dimension
a_i^d	Variable	The acceleration of the agent *i* in direction *d*th
x_i^d	Variable	The position of *i*th agent in the *d*th dimension
$R_{i,j}$	Variable	Euclidean distance between two agents *i* and *j*
F_i^d	Variable	The total force that acts on agent *i* in a dimension

In this paper, the KM-GSA was designed to improve the clustering process. To evaluate the performance of the proposed KM-GSA algorithm, the result of the KM-GSA was compared against the results of the KM algorithm and GSA. Each algorithm used the same training and testing dataset. While evaluating the KM-GSA algorithm, the number of clusters (N) was predefined as 5, 20 and 30. Then, the initial point for each cluster was arbitrarily selected, whereby the random instances were specifically used as the center points for the clusters. During the training, the best results were achieved when the number of clusters was 5. Based on this result, the number of clusters in this research was predefined as 5. The process of clustering algorithms was terminated when the maximum number of iterations was exceeded, or when the value of the objective function was close to zero. The results of the KM-GSA were evaluated in terms of effectiveness of detection which was represented by detection time as well as the detection rate, false positive rate and detection accuracy rate. The detection effectiveness of the KM-GSA was assessed by reference to the high detection accuracy, high detection rate, low false positive rate and low detection time.

VI. RESLUTS AND DISCUSSION

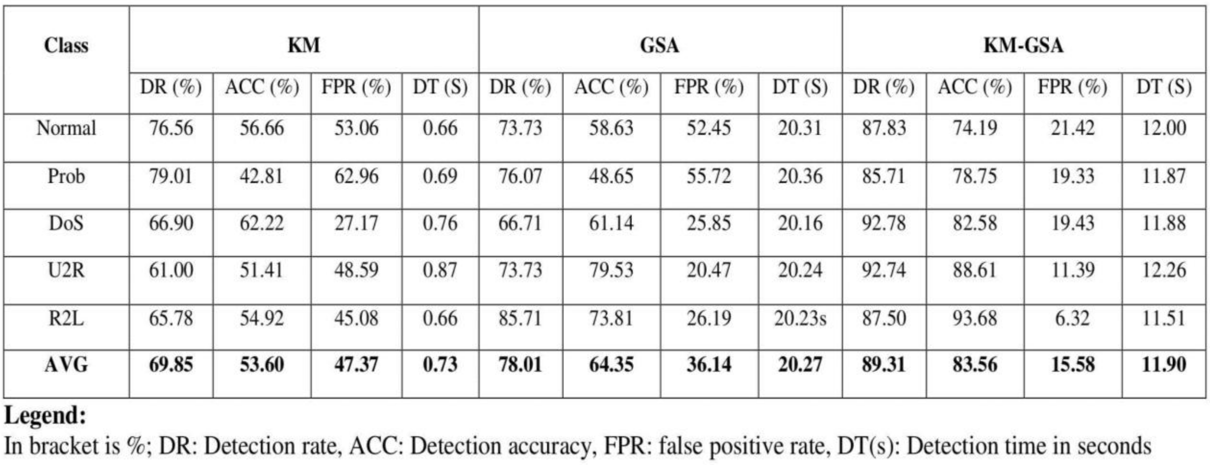
Standard measurements, such as the detection rate (DR), false positive rate (FPR), and detection accuracy rate (ACC), for evaluating the performance of KM-GSA algorithm are shown in Table 2.

Table 2 Description of performance measures

Performance Measures		Description
Percentage (%) Classification	Accuracy	Correctly classified as normal and attacks into their respective classes. It quantifies the discriminating capability of the classifier/model when presented with input data. $$\frac{TN + TP}{TN + TP + FN + FP}$$
	True Positive Rate (TPR) also known as Detection Rate (DR)	Measure the frequency of the targeted data correctly classified by the classifier/model as normal. $$\frac{TP}{TP + FN}$$
Error Percentage (%)	False Positive Rate (FPR) also known as False Alarm Rate (FAR)	Average number of normal traffic wrongly identified as malicious traffic (false alarm rate) $$\frac{FP}{TN + FP}$$

Table 3 summarizes the results of the KM algorithm, GSA and KM-GSA for detection accuracy, true positive rate, false positive rate and detection time for all traffic classes.

Table 3 Performance results for the three algorithms (KM, GSA and KM-GSA)



Class	KM				GSA				KM-GSA			
	DR (%)	ACC (%)	FPR (%)	DT (S)	DR (%)	ACC (%)	FPR (%)	DT (S)	DR (%)	ACC (%)	FPR (%)	DT (S)
Normal	76.56	56.66	53.06	0.66	73.73	58.63	52.45	20.31	87.83	74.19	21.42	12.00
Prob	79.01	42.81	62.96	0.69	76.07	48.65	55.72	20.36	85.71	78.75	19.33	11.87
DoS	66.90	62.22	27.17	0.76	66.71	61.14	25.85	20.16	92.78	82.58	19.43	11.88
U2R	61.00	51.41	48.59	0.87	73.73	79.53	20.47	20.24	92.74	88.61	11.39	12.26
R2L	65.78	54.92	45.08	0.66	85.71	73.81	26.19	20.23s	87.50	93.68	6.32	11.51
AVG	**69.85**	**53.60**	**47.37**	**0.73**	**78.01**	**64.35**	**36.14**	**20.27**	**89.31**	**83.56**	**15.58**	**11.90**

Legend:
In bracket is %; DR: Detection rate, ACC: Detection accuracy, FPR: false positive rate, DT(s): Detection time in seconds

The results showed that the KM-GSA outperformed the KM algorithm and the GSA in terms of the detection rate and detection accuracy in all five classes. The KM-GSA achieved the highest accuracy with average rates of 89.31% and 83.89 % for detection and detection accuracy, respectively. The KM algorithm achieved 69.85 % and 53.60 % for the detection rate and detection accuracy, respectively, and the GSA achieved 78.01%and 64.35% for the detection rate and detection accuracy, respectively. Based on the results, the KM-GSA achieved the lowest false positive rate compared to the KM algorithm and GSA in all five classes. The KM-GSA achieved the lowest false positive with an average rate of 15.57 %. Furthermore, the results (Table 3 above) showed that the detection time improved with the KM-GSA classifier with an average rate of 40 % compared to the detection time for the GSA.

Fig 2 and Fig 3 illustrate the comparison in terms of overall detection accuracy and false positive rate for the KM, GSA and KM-GSA algorithms. The results on the detection accuracy “Fig.1” showed that the KM-GSA achieved the highest accuracy, whereas the KM algorithm achieved the lowest accuracy. The KM-GSA improved the detection accuracy by 19.21 % and 29.96 % compared to the GSA and KM algorithm, respectively. The results on the false positive rate “Fig.2” showed that the KM-GSA achieved the lowest false positive rate, whereas the KM algorithm achieved the highest false positive rate. The KM-GSA reduced the false positive rate by 31.79 % and 20.56 % as compared to the KM and GSA, respectively. The KM-GSA outperforms the KM algorithm and GSA in terms of detection accuracy and false positive rate because it has features of both KM algorithm and GSA.

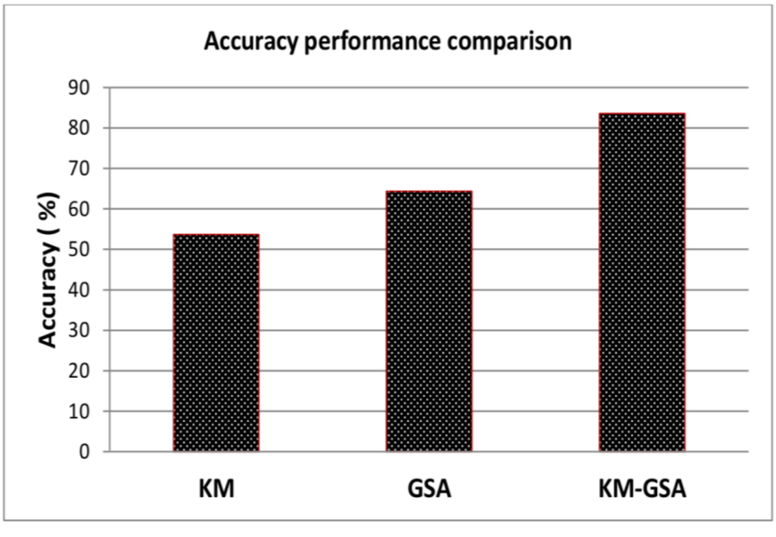


Fig. 2. Detection accuracy comparison of KM, GSA and KM-GSA

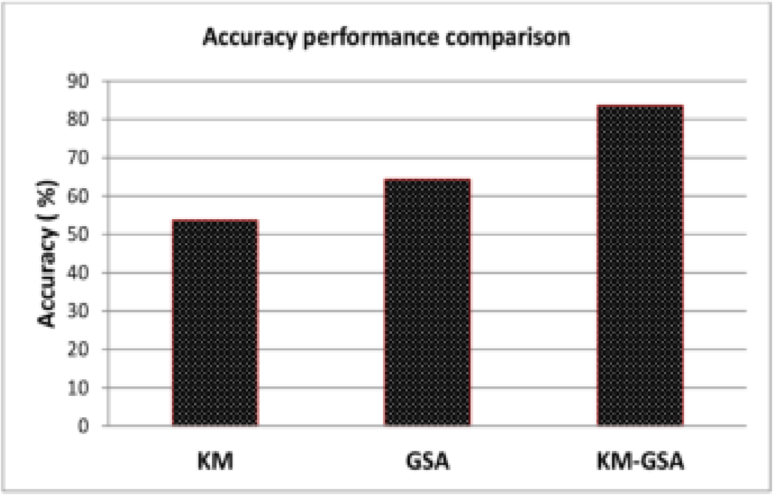


Fig. 3.False positive rate comparison of KM, GSA and KM-GSA

Fig 4 illustrates the results on the detection time behaviour of the KMGSA and GSA. As shown in the Fig 4, the detection time of the GSA was less than the detection time of the KM-GSA. The KM-GSA improved the detection time by 40 % as compared to the GSA. It is concluded that the KM-GSA increases the convergence speed of the GSA classifier because it has features of both the KM algorithm and the GSA.

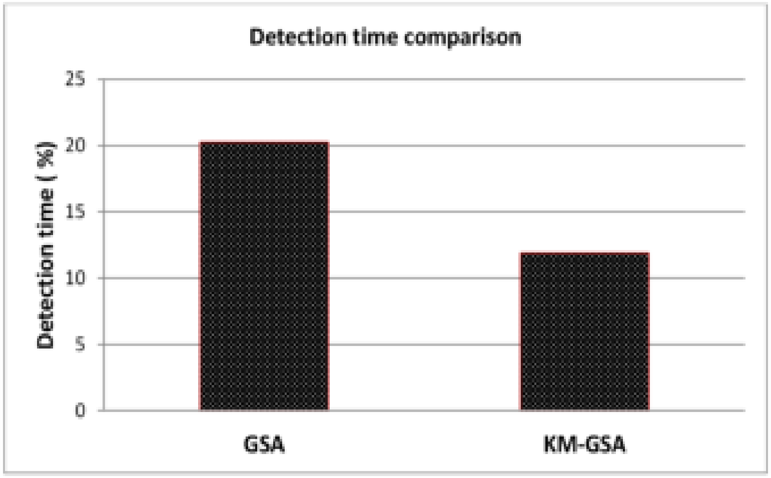


Fig. 4.Comparison of the detection time of the GSA and GSA-KM

VII. CONCLUSIONS

A new hybrid KM-GSA was designed to address the inefficient clustering technique which has an impact on the clustering process in ensemble clusters. The main purpose of the KM-GSA is to enhance the quality of the clustering result. The KM-GSA executes the KM algorithm on the selected dataset to find the near optimal centroids for the desired clusters. Then, the result of the KM algorithm is used to generate the initial population for the GSA. Finally, the GSA is used as the clustering algorithm to refine the clusters formed by KM. The KM-GSA was validated by comparing the obtained results against the KM algorithm and GSA. The results showed that the KM-GSA outperformed the KM algorithm and GSA in terms of detection accuracy and false positive rate.

REFERENCES

[1] K.. Wankhade, S. Patka, and R.. Thool, "An Overview of Intrusion Detection Based on Data Mining Techniques," *IEEE Conference on Communication Systems and Network Technologies (CSNT)*, pp. 626–629, 2013.

[2] D.Van der Merwe and A. P. Engelbrecht, " Data Clustering Using Particle Swarm Optimization*," In The IEEE Congress on Evolutionary Computation, CEC*, vol. 1, pp. 215–220, (2003).

[3] A. Hatamlou, S. Abdullah and H. “Nezamabadi-Pour, Application of gravitational search algorithm on data clustering,” *In 6th International Conference on Rough Sets and Knowledge Technology, RSKT, Springer* , pp. 337–346, (2011).

[4] G. Saini, and H. Kaur., *A Novel Approach Towards K-Mean Clustering Algorithm With PSO*, International Journal of Computer Science and Information Technologies , 2014 , vol. 5.

[5] A. Dastanpour, S. Ibrahim, , R. Mashinchi and A. Selamat, *Using Gravitational Search Algorithm to Support Artificial Neural Network in Intrusion Detection System*, SmartCR, 2014, vol.4.

[6] P. Gogoi, D. Bhattacharyya, B. Borah, and J. K. Kalita, *MLH-IDS: a multilevel hybrid intrusion detection method*, The Computer Journal, 2014, vol .57.

[7] S. Mukkamala, A. H. Sung, and A.Abraham, “Intrusion Detection Using Ensemble of Soft Computing Paradigms,” *In Intelligent Systems Design and Applications*, pp. 239–248, (2003).

[8] C.F.Tsai, Y.F. Hsu, C.Y. Lin and W.Y. Lin, Intrusion Detection by Machine Learning: A review, *Expert Systems with Applications*, 2009, vol. 36.