Anomaly Intrusion Detection System based on Unlabeled Data

Salima Benqdara
University of Benghazi
Benghazi, Libya

ABSTRACT
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. Therefore, this paper has proposed a new AIDS based on Supervised and unsupervised methods that effectively detects attacks with a low false positive rate. The proposed approach consists of ensemble clusters with an efficient clustering technique, and enhancing the capability of the detection classifier by utilizing an efficient method. Experimental results showed an improvement in the detection accuracy which scored 97.0% on the overall accuracy and 0.03 % on the false positive rate for all classes of network traffic. Hence, this validates the proposed GSA-based AIDS.

General Terms
Security, Algorithms.

Keywords
Network Intrusion Detection, ensemble clusters, unlabeled data.

1. INTRODUCTION
The importance to safeguard computer network against confidentiality, integrity and availability breaches is an important issue. Security policies or firewalls have difficulty in preventing such attacks because of the hidden vulnerabilities contained in software applications. It is important to have a detecting and monitoring system to protect important data. An Intrusion Detection System (IDS) is a protection system that plays an important role to protect or secure networks. IDS is an automated system which can detect a computer system invasion by using an audit trail provided by the operating system or by using a network monitoring tools. The main target of the IDS is to monitor network events automatically to detect malicious. Therefore, IDS is required as an additional wall for protecting systems despite the prevention technique. 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.

Supervised and unsupervised methods are promising approaches to detecting new attacks. Supervised methods, also known as classification methods, require a labeled training dataset. The dataset contains both normal and anomalous samples that build a training model. Supervised methods are more efficient in network classification compared to unsupervised methods since they have access to more information [1]. However, supervised anomaly detection methods depend on a labeled training dataset, making the intrusion detection process error-prone, costly and time consuming. Since labeled is often done manually, any mistake in labeling the training data may lead to detection ineffectiveness. On the other hand, unsupervised anomaly detection schemes allow training based on unlabeled data sets, but this suffers from clustering problems. Better accuracy rate can be achieved by merging two or more machine learning algorithms to construct the ensemble clustering [2].The ensemble clustering was introduced as a prominent technique for improving the robustness, stability and accuracy of unsupervised classification solutions [3,4]. It combines multiple partitions generated by different clustering algorithms, which exploit the strong points of each individual algorithm to generate a better final outcome [5, 6]. Ensemble clustering provides better performance than single clustering algorithms in several respects [3, 4, 5].

The choice of classifiers to classify the data traffic is an issue because they can affect the accuracy and classification of an attack. Various machine learning methods are used to classify intrusion detection datasets such as the decision tree, nave Bayesian, neural network and (SVM). The SVM is a margin-based classifier based on small sample learning with good generalization capabilities, and is commonly used in the application of classification [7, 8]. The SVM outperforms in the important aspect of robustness and efficiency in the network classification. It can manage the problem of imbalanced attacks which can otherwise lead to poor detection performance. This problem occurs due to the small learning sample size of low-frequent attacks compared to high-frequent attacks. Moreover, SVMs outperform the neural network in the important aspects of scalability, training time and prediction accuracy [9]. SVM is commonly used in IDSs because of its robustness and efficiency in the network classification [10]. However, one of the primary problems of SVM is how to select the kernel function and its parameter values. This problem is a crucial step in handling a learning task with SVM since it has an impact on the classification accuracy [8, 11].

In order to address the problems described in the previous paragraph, this paper proposes the AIDS based on the Gravitational Search Algorithm (GSA-based AIDS) based on unsupervised and supervised techniques. The proposed system consists of data scaling, ensemble clusters and a hybrid classifier.

The rest of the paper is organized as follows: Section 2 discusses the related works on the ensemble clusters and a

hybrid classifier 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.

2. RELATED WORK

Based on a review of the literature [(4, 12, 13], detection accuracy is improved by ensemble classifiers approach and hybrid classification approach.

Gao et al. (2010) presented an intrusion detection framework based on a parallel clustering ensemble algorithm (PEA-IDS). The parallel clustering algorithm is used to achieve high accuracy and a low false alarm rate. The PEA-IDS system is composed of two stages. In the first stage, the PEA-IDS is used to preprocess the dataset. The parallel clustering ensemble is proposed from different KM algorithms and the KM algorithms use different initial k cluster centers to find different partitions. In the second stage, the EA algorithm is used to combine the results of multiple clustering into a single data partition. Then, the scheme uses the PEA to detect the abnormal network behavioral patterns. The performance of the PEA-IDS framework was tested on the KDD 99 dataset and the results showed that it produced a great improvement in time and efficiency.

Wang et al. (2010b) proposed a hybrid approach to the design of an IDS. The proposed approach combines the support vector classifier and ABC algorithm (ABC-SVM). The ABC parameter parameters and beneficial features for the SVM. The results showed that the ABC-SVM approach achieved high accuracy rates.

Govindarajan and Chandrasekaran (2011) introduced a hybrid architecture to design an intrusion detection model based on an ensemble classifier. The proposed model consists of an MLP algorithm and radial basis function. In comparison to a single base classification method, the proposed approach was better in terms of performance. The results indicated that the ensemble classifier of the MLP achieved better detection accuracy against the ensemble classifier of the radial basis function classifier. The proposed approach represented a significant improvement in terms of prediction accuracy in intrusion detection.

Manekar and Waghmare (2014) proposed an IDS based on the machine learning technique. The proposed system consists of two machine learning algorithms: SVM and PSO. In the first step in the proposed system, the PSO algorithm is used to optimize the value of the C and parameters and important features for the SVM. In the second step, the parameters and features are used to train the SVM. The results showed that the proposed system k improved the detection accuracy compared to a single SVM classifier.

Kuang et al. (2014) proposed a new intrusion detection system composed of kernel principal component analysis (KPCA) and GA with SVM. The N-KPCAGA- SVM system consists of two stages. In the first stage, KPCA is used to reduce the dataset and extract the features of the normalized data. The second stage deals with the detection classifier. The GA is used to optimize the accuracy of the SVM classifier by detecting the subset of the best values of kernel parameters for the SVM classifier. The results showed that the classification accuracy of the proposed system achieved faster convergence speed and better detection accuracy compared with a single SVM classifier.

Dastanpour et al. (2014) presented an approach for an IDS composed of the ANN algorithm and GSA optimization. The proposed system consists of two stages. In the first stage, the ANN algorithm is executed on the training dataset and the recognition results of the ANN are sent to next stage. In the second stage, the recognition results of the ANN are classified by the hybrid GSA-ANN algorithm. The KDD 99 dataset was used to evaluate the proposed system, with the results showing that the GSA-ANN hybrid approach achieved high accuracy compared with a single ANN algorithm.

Patka (2014) presented an IDS based on a clustering ensemble. The proposed system is composed of the IG method and KM clustering algorithm. The IG method is used to select the important features from the dataset. The clustering ensemble based on the KM algorithm is used to improve the performance of the IDS and achieve high accuracy and a low false alarm rate. The initial clusters are formed using the KM algorithm. The number of k cluster centroids is calculated by dividing and merging the clusters. Then, the points with the higher density are selected as the initial centroids and cluster formation is performed again to detect normal and anomalous records. The divide and merge step helps in calculating the k number of cluster centroids. The KDD 99 dataset was used to test the performance of the system, with the results showing that the system achieved a high detection rate and low false alarm rate.

3. GRAVITATIONAL SEARCH ALGORITHM (GSA)

Gravitational search algorithm is one of the latest heuristic optimization algorithms, which was first introduced by Rashedi et al. (2009) as a new stochastic population-based optimization tool based on the metaphor of gravitational interaction between masses. The GSA is constructed on the law of Newtonian Gravity; every particle in the universe attracts every other particle with a force that is directly proportional to the product of their masses and inversely proportional to the square of the distance between them. In the algorithm, all the individuals can be viewed as objects with masses. The objects attract each other by the gravity force, and the force makes all of them move towards the ones with heavier masses. The objects transform information by the gravitational force, and the objects with heavier masses become heavier [18].

To describe the GSA, consider a system with N masses (agents) in which the position of the ith mass is defined as follows:

$$X_i = \left(x_i^1, \ldots\ldots, x_i^d, \ldots\ldots\ldots, x_i^n\right) \quad (1)$$

The mass of each agent is calculated after computing a current population's fitness as follows:

$$M_i(t) = \frac{m_i(t)}{\sum_{j=1}^{N} m_j(t)} \quad (2)$$

Where

$$m_i(t) = \frac{fit_i(t) - worst(t)}{\sum_{j=1}^{N}(fit_i(t) - worst(t))} \quad (3)$$

Where fit_i (t) represent the fitness value of the agent i at time t. best (t) and worst (t) are the best and worst fitness of all agents, respectively and defined as follows:

$$best(t) = \min_{j\in\{i,\ldots,N\}} fit_j(t)$$
$$worst(t) = \max_{j\in\{i,\ldots,N\}} fit_j(t) \quad (4)$$

To compute the acceleration of an agent, the total forces from a set of heavier masses that act on it should be considered based on the law of gravity (Equation 5), followed by the calculation of an agent acceleration using a law of motion (Equation. 6. After that, the next velocity of an agent is calculated as a fraction of its current velocity added to its acceleration (Equation 7). Then, its next position can be calculated using Equation8.

$$F_d^i(x) = \sum_{j\in kbest, j\neq i} rand_j G(t) \frac{M_j(t)M_i(t)}{R_{i,j}(t)+\varepsilon}\left(x_j^d(t) - x_j^d(t)\right) \quad (5)$$

$$a_i^d(t) = \sum_{j\in kbest, j\neq i} rand_j G(t) \frac{M_j(t)}{R_{i,j}(t)+\varepsilon}\left(x_j^d(t) - x_i^d(t)\right) \quad (6)$$

$$V_i^d(t+1) = rand_i \times V_i^d(t) + a_i^d(t) \quad (7)$$

$$X_i^d(t+1) = X_i^d(t) + V_i^d(t+1) \quad (8)$$

4. PROPOSED APPROACH

The problem identified in this study is that supervised anomaly detection requires a labeled training dataset which makes the intrusion detection process error-prone due to the manual labeling of the data. The imbalanced data problem arises due to the difficulty in identifying the decision boundaries between normal and abnormal behaviors. Furthermore, the detection classifier (SVM) has a drawback regarding the selection of the kernel function and its parameter values. In order to address the study problems, in this paper GSA-based AIDS proposed

consists of two main phases whereby each phase provides the output which is essential for the next phase.

4.1 Phase 1: Conversion of Unlabeled Data to Higher-Quality Labeled Data

The first phase deals with data scaling and converting the unlabeled data to labeled data. Data scaling is done to ensure the training datasets are within the range of [0, 1]. Then, the unlabeled dataset is converted to a labeled dataset by designing ensemble clusters to produce a higher quality labeled dataset. Phase 1, addresses further improvements to the ensemble cluster approach by improving the clustering method through the design of a hybrid clustering algorithm to efficiently label the data. The output of this phase is a higher-quality labeled dataset which is used to classify the traffic data in the next phase. Three clustering algorithms (KM-GSA, KM-PSO and FCM) were chosen to design the ensemble clusters, since they gave better detection accuracy and minimum false positive rate. Each of the clustering algorithms is used to cluster the unlabeled data into a new set of labeled data. The steps of the three clustering algorithms (KM-GSA, KM-PSO and FCM) are explained in the Algorithm 1, Algorithm 2, and Algorithm 3. The outputs of the individual clustering algorithms are sent to an un-weighted voting scheme to select the output of the ensemble clusters, which produces a higher-quality labeled dataset. The architecture of ensemble clusters is illustrated in Figure 1.

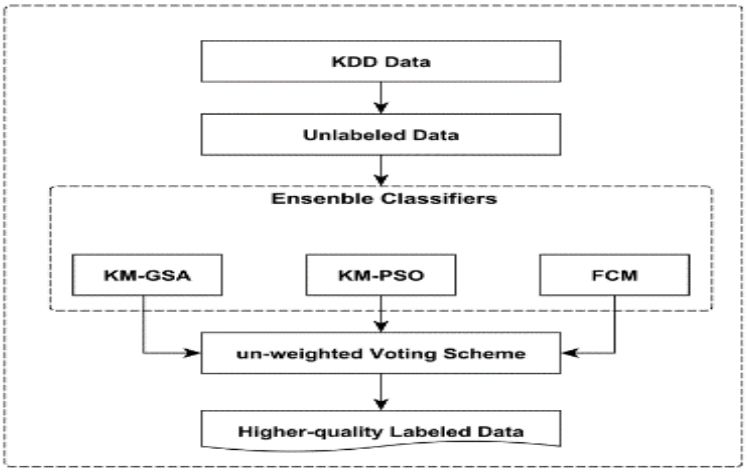

Fig 1: Architecture of ensemble clusters

4.2 Phase 2: Detection Classifier

Phase 2 focuses on the improvement of the detection classifier by applying a hybrid classification approach. The hybrid classifier is designed based on a combination of the GSA and SVM algorithms as shown in Figure 2. The GSA is introduced as an optimization technique to optimize the SVM parameters. The GSA starts with n-randomly selected agents and searches for the optimal agent iteratively. Each agent is an m dimensional vector and represents a candidate solution. The SVM classifier is built for each candidate solution to evaluate its performance through evaluation of the fitness function. The fitness function value is based on the

classification accuracy of the SVM classifier. The GSA guides the selection of potential subsets that lead to the best prediction accuracy. The outcome of the final detection stage is the identification of attacks through the improved detection effectiveness of the proposed GSA-based AIDS. The detailed steps of the algorithm are explained in the Algorithm 4

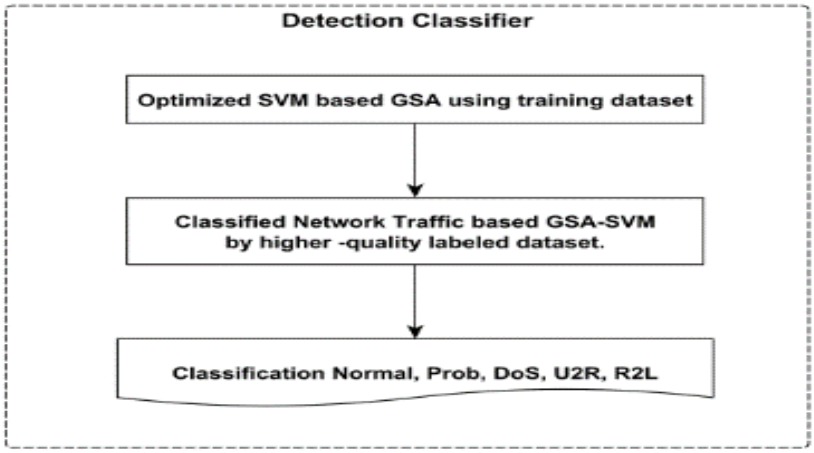

Fig 2: Architecture of detection classifier

5. 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 labeled 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 [19, 20].

6. EXPERMENTAL SETUP

This section describes the experimental setup to evaluate the proposed GSA-based AIDS and its components. In order to evaluate and compare the effectiveness of the proposed system, this study used the KDD cup 1999 dataset. The dataset contains 4,940,000 traffic connections consisting of normal network traffic and 24 types of attacks from four categories of attacks, namely, probe, DoS, U2R and R2L attacks. . In this study, 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]. In this study, the number of iterations was 500 iterations and all the experiments were repeated 500 times (iterations) and the results were averaged. To produce a high-quality labeled dataset, in the training phase, the ensemble clustering is first applied to each class of the unlabeled KDD cup 1999 training dataset. Each of the clustering algorithms is used to cluster the unlabeled data into a new set of labeled data. The output of the ensemble cluster training phase consists of five class-specific classifiers: normal, probe, DoS, U2R and R2L. The output of the individual clustering algorithms is sent to an un-weighted voting scheme to select the output of the ensemble clusters for each class, which produces higher-quality labeled data. In order to improve the effectiveness of the detection by the GSA-based AIDS, a new high-quality labeled dataset is produced. The higher-quality labeled dataset is used to test and validate the GSA-based AIDS. Standard measurements, such as the detection rate (DR), false positive rate (FPR), and detection accuracy rate (ACC), for evaluating the performance of GSA-based AIDS are shown in Table 1.

7. RESLUTS AND DISCUSSION

The GSA-based AIDS was evaluated in terms of the overall accuracy, detection rate and false positive rate. In order to evaluate the effectiveness of the higher-quality labeled dataset to address this limitation and improve the effectiveness of detection, the GSA-based AIDS was validated using a higher-quality labeled dataset and the KDD Cup 1999 test dataset. The performance results of the GSA-based AIDS using the higher quality labeled dataset were then benchmarked against the performance results of the GSA-based AIDS using the KDD Cup 1999 test dataset. Table 2 presents a summary of the results achieved by the GSA-based AIDS using a higher-quality labeled dataset and using the KDD Cup 1999 test dataset.

The results in Table 2 showed that the GSA-based AIDS using the higher-quality labeled dataset outperformed the GSA-based AIDS using the KDD Cup 1999 test dataset in terms of detection accuracy, detection rate and false positive rate. The GSA-based AIDS using the higher-quality labeled dataset achieved a high detection rate and accuracy, with an average rate of 96.85 % and 97.05 % respectively. However, the GSA-based AIDS using the KDD Cup 1999 test dataset achieved 91.64 % and 86.70 % for the detection rate and detection accuracy, respectively. The GSA-based AIDS using the higher-quality labeled dataset achieved a lower false positive rate than the GSA-based AIDS using the KDD Cup 1999 test dataset in all five traffic classes with an average of 0.03%. The results as set out in the table showed that the GSA-based AIDS using the higher-quality labeled dataset improved the detection accuracy for U2R and R2L classes by 12.0 % and 23.06% , respectively, compared to the GSA-based AIDS using the KDD Cup 1999 test dataset.

Figure 3 illustrates the detection accuracy of the GSA-based AIDS using the KDD Cup 1999 test dataset and the GSA-based AIDS using the high-quality labeled dataset with respect to all five classes. The results indicated that the GSA-based AIDS using the high-quality labeled dataset obtained high detection accuracy compared to the GSA-based AIDS using the KDD Cup 1999 test dataset. The results indicated that the GSA-based AIDS using the high- quality labeled dataset and the GSA-based AIDS using the KDD Cup 1999 test dataset provided high detection accuracy on the DoS class. The GSA-based AIDS using theKDD Cup 1999 test dataset obtained the highest detection accuracy on the DoS class. However, the GSA-based AIDS using the KDD Cup 1999 test dataset obtained the lowest detection accuracy on the U2R class. The GSA-based AIDS using the high-quality labeled dataset obtained the lowest detection accuracy on the U2R class. The GSA-based AIDS using the high-quality labeled dataset had the highest detection accuracy for all classes. It can be concluded that the GSA-based

AIDS using the high-quality labeled dataset outperforms the GSA-based AIDS using the KDD Cup 1999 test in terms of detection accuracy for all classes because of the advantages gained from using the higher-quality labeled dataset to classify the data traffic.

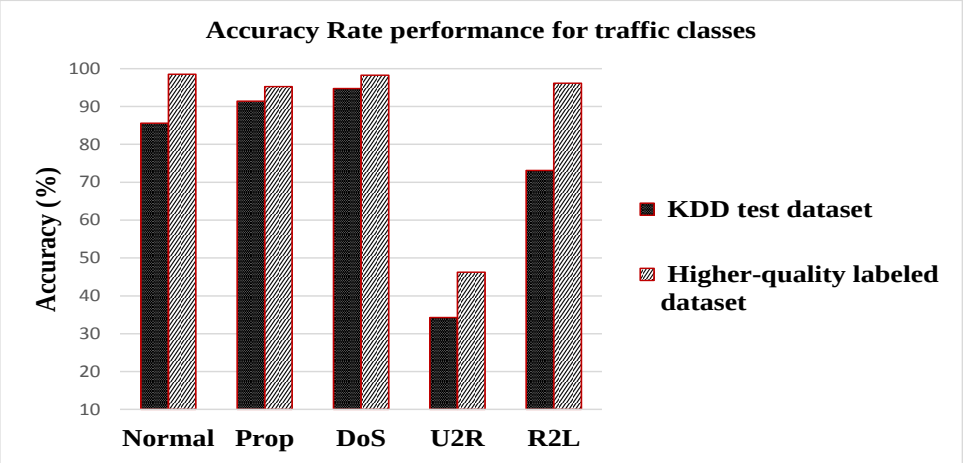

Fig. 3. Comparison detection accuracy of GSA-based A-IDS on different test dataset

Figure 4 illustrate the comparison in terms of the false positive rate for the GSA-based AIDS using the higher-quality labeled dataset and the GSA-based AIDS using the KDD Cup 1999 test dataset. The results on the false positive rate indicated that the GSA-based AIDS using the higher-quality labeled dataset achieved the lowest false positive rate. The GSA-based AIDS using the higher-quality labeled dataset reduced the false positive rate by 0.17 % compared with GSA-based AIDS using the KDD Cup 1999 test dataset. Overall, the results showed that the GSA-based AIDS using the higher quality labeled dataset outperformed the GSA-based AIDS using the KDD Cup 1999 test dataset in terms of false positive rate because of the advantages of using the higher-quality labeled dataset to classify the data traffic.

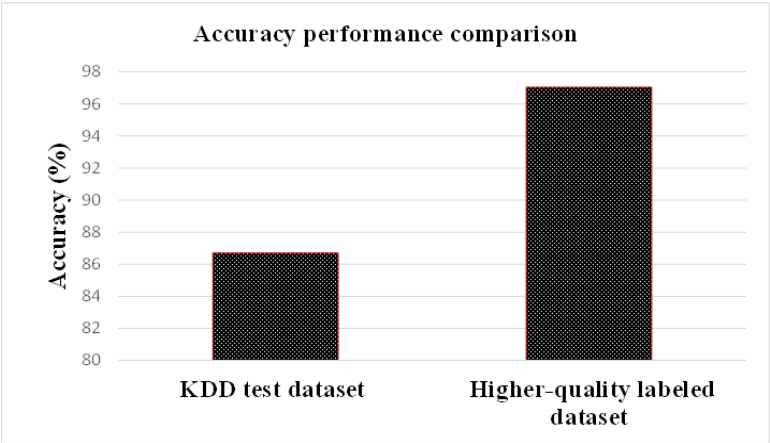

Fig 4: False positive performance of GSA-based AIDS on different test datasets

Figure 5 illustrates the detection accuracy of the GSA-based AIDS using the GSA-SVM classifier and SVM classifier with respect to the five traffic classes. The results showed that the GSA-based AIDS obtained high detection accuracy when using the GSA-SVM classifier compared to the GSA-based AIDS using the SVM classifier. The results also showed that the GSA-based AIDS using the GSA-SVM classifier and the GSA-based AIDS using the SVM classifier

achieved the highest detection accuracy on the DoS and normal classes. The GSA-SVM classifier and SVM classifier obtained the lowest detection accuracy on the U2R class; however, the GSA-based AIDS using the GSA-SVM classifier obtained similar detection accuracy on the normal and DoS classes. The GSA-based AIDS using the GSA-SVM classifier had the highest detection accuracy for all classes. The GSA-based AIDS using the GSA-SVM classifier outperformed the GSA-based AIDS using the SVM classifier in terms of the detection accuracy for all classes because it included the GSA as an optimization technique to optimize the SVM parameters.

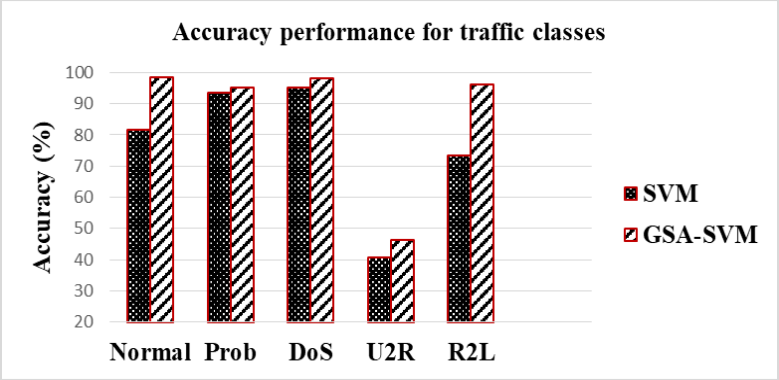

Fig 5: Detection accuracy performance of GSA-based AIDS with different classifiers

8. CONCLUSION

The aim of this paper is to design and develop the GSA-based AIDS as an effective IDS. The proposed system produces higher-quality labeled data by implementing ensemble clusters with an efficient clustering technique, and enhancing the capability of the detection classifier by utilizing an efficient method. The ensemble clusters technique was designed to convert unlabeled data to higher-quality labeled data and the GSA-SVM classifier was designed to enhance the classification process in the detection classifier. The results showed an improvement in detection effectiveness when using the high-quality labeled dataset which scored 97.0% on the overall accuracy and 0.03 % on the false positive rate. The GSA-based AIDS using the higher-quality labeled dataset outperformed the GSA-based AIDS using the KDD 99 test dataset. The detection accuracy improved by 10.84 % while the false positive rate reduced by 0.17 % when using a higher-quality labeled dataset. The detection accuracy of the GSA-based AIDS using the GSA-SVM classifier improved by 6.95 % ,while the false positive rate reduced by 0.07 % as compared to the GSA-based AIDS using the SVM classifier.

9. REFERENCES

[1] Yadav, M. R, and Kumbharkar, P. B, "Intrusion Detection System with Supervised Learning Algorithms ", International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE), Vol. 4(4), (2014), 305–310.

[2] . Tsai, C. F, and Lin, C.Y, "A triangle Area based Nearest Neighbors Approach to Intrusion Detection. ", Pattern Recognition, 43(1), (2010), 222–229.

[3] Gao, H., Zhu, D, and Wang, X. 2010. A parallel clustering ensemble algorithm for intrusion detection system. In The Ninth International Symposium on Distributed Computing and Applications to Business Engineering and Science, DCABES.IEEE, 450–453.

[4] Patka, S. (2014). Intrusion detection model based on data mining technique. In International Conference on Advances in Engineering and Technology, ICAET. IOSR-JCE, 34–39.

[5] Ghaemi, R., bin Sulaiman, N., Ibrahim, H, and Mustapha, N, " A review: accuracy optimization in clustering ensembles using genetic algorithms", Artificial Intelligence Review, 35(4), (2011), 287–318.

[6] Bahri, E., Harbi, N, and Huu, H. N. 2012. A Multiple Classifier System Using an Adaptive Strategy for Intrusion Detection. In International Conference on Intelligent Computational Systems, ICICS. 7–8.

[7] Majid, A., Khan, A. and Mirza, A. M, " Combination of support vector machines using genetic programming", International Journal of Hybrid Intelligent Systems, 3(2), (2006), 109–125.

[8] Kuang, F., Xu, W, and Zhang, S, "A novel hybrid KPCA and SVM with GA model for intrusion detection ", Applied Soft Computing, 18, (2014), 178–184.

[9] Srinivas, M. and, Andrew, H. 2003. Feature selection for intrusion detection using neural networks and support vector machines. Transportation Research Board, winter, 1–11.

[10] Tsai, C. F., Hsu, Y. F., Lin, C.Y, and Lin, W.Y, " Intrusion Detection by Machine Learning: A review. Expert Systems with Applications", 36(10), (2009), 11994–12000.

[11] Ranaee, V., Ebrahimzadeh, A. and Ghaderi, R, "Application of the PSO– SVM Model for Recognition of Control Chart Patterns", ISA transactions, 49(4), (2010), 577–586.

[12] Peddabachigari, S., Abraham, A., Grosan, C, and Thomas, J, "Modeling Intrusion Detection System Using Hybrid Intelligent Systems", Journal of network and computer applications, 30(1), (2007), 114–132.

[13] Kausar, N., Samir, B. B, and Hussin,M, " Efficient Intrusion Detection system based on support vector machines using optimized kernel function. ", Journal of

Theoretical and Applied Information Technology, 60(1), (2014), 55–63.

[14] Wang, J., Li, T, and Ren, R. 2010b. A real Time IDS Based on Artificial Bee Colony-Support Vector Machine Algorithm. In The Third International Workshop on Advanced Computational Intelligence (IWACI). IEEE, 91–96.

[15] Govindarajan, M, and Chandrasekaran, R, "Intrusion Detection Using Neural Based Hybrid Classification Methods", Computer networks, 55(8), (2011), 1662–1671.

[16] Manekar, V, and Waghmare, K, " Intrusion Detection System using Support Vector Machine (SVM) and Particle Swarm Optimization (PSO) ", International Journal of Advanced Computer Research, 4(3), (2014), 25–30.

[17] Dastanpour, A., Ibrahim, S., Mashinchi, R, and Selamat, A, "Using Gravitational Search Algorithm to Support Artificial Neural Network in Intrusion Detection System", Smart CR, 4(6), 426–434.

[18] Rashedi,E., Nezamabadi ,H., and Saryazdi, S., "Filter modeling using gravitational search algorithm " , Engineering Applications of Artificial Intelligence, to be published, 2010.

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

[20] Tsai, C.-F., Hsu, Y.-F., Lin, C.-Y, and Lin, W.-Y, "Intrusion Detection by Machine Learning: A review ", Expert Systems with Applications, 36(10), (2009), 11994– 12000.

10. APPENDIX

Algorithm 2 KM-PSO

Input sample data set Y; Set the parameters of KM-PSO $(\omega, N, C1, C2, t-max)$
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 particle to contain N randomly selected from k-means output
Repeat
 For each particle 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
Update the cluster centroids
 End For
Until: cluster centroid not change or max-iter

Algorithm 3 FCM

Initialize the membership function values u_{ij} , i = 1,2,..., n; j = 1,2,..., c; Set the parameters of FCM $(n, m, \varepsilon, t - max)$

Repeat

Calculate the cluster center

Compute Euclidian distance

Update the membership values

Until $\left| U^{(t)} - U^{(t-1)} \right| < \varepsilon \, or \, max - itr$

Algorithm 4 GSA-SVM

Initialize the position and velocity of agents randomly,Set the parameters of GSA-SVM (N, G0, ε, tmax)

Repeat

For each mass $i = 1, 2, ..., Ndo$

Train SVM

Evaluate fitness function of each agent

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 of agents

Update the position of the agents

End For

Until: cluster centroid not change or max-iter

Retrain SVM and classification results

Table 1. 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 2. Performance results for GSA-based AIDS

Class	GSA-based AIDS using KDD Cup 1999 test dataset			GSA-based AIDS using higher-quality labeled dataset		
	ACC (%)	FPR (%)	DR (%)	ACC (%)	FPR (%)	DR (%)
Normal	85.57	0.13	74.34	98.52	0.03	99.99
Prob	91.39	0.11	94.74	95.25	0.09	100
DoS	94.74	0.19	91.86	98.28	0	94.74
U2R	34.26	0.66	26.57	46.26	0.41	46.3
R2L	73.11	0.39	34.33	96.17	0	92.67
AVG	**86.20**	**0.29**	**91.64**	**97.04**	**0.03**	**96.85**

Legend:
In bracket is %; ACC=Detection accuracy, FP=False positive rate; DR=Detection rate
AVG: this average excluded the up normal value