An Adaptive Intrusion Detection Model based on Machine Learning Techniques

Salima Omar
Universiti Teknologi Malaysia
Faculty of Computing

Asri Ngadi
Department of Information Technology
Universiti Teknologi Malaysia
Faculty of Computing

Hamid H. Jebur
Universiti Teknologi Malaysia
Faculty of Computing

ABSTRACT

Intrusion detection continues to be an active research field. Even after 20 years of research, the intrusion detection community still faces several difficult problems. Detecting unknown patterns of attack without generating too many false alerts remains an unresolved problem. Although recently, several results have shown that there is a potential resolution to this problem. Anomaly detection is a key element of intrusion detection in which perturbations of normal behavior suggest the presence of intentionally or unintentionally induced attacks, faults, and defects. This paper proposes a hybrid machine learning model based on combining the unsupervised and supervised classification techniques. Clustering approach based on combining the K-means , fuzzy C-means and GSA algorithms to obtain the normal patterns of a user's activity, the technique is used as the first component for pre-classification to improve attack detection. Then, a hybrid classification approach of Support Vector Machine (SVM) and Gravitational Search Algorithm (GSA) algorithm will be used to enhance the detection accuracy.this research used the KDD CUP 1999 to get initial results, which were encouraging.

Keywords

Supervised Machine Learning, Unsupervised Machine Learning, Network Intrusion Detection.

1. INTRODUCTION

Information is now becoming ubiquitous with the infrastructure like the Internet. Sensitive information exposure is inevitable with the increasing use of the Internet. Studies covering the prevention, detection and the forensic aspect of computer network attacks have long been researched on. The prevention techniques such as encryption, Virtual Private Network (VPN) and firewall alone seem to be inadequate. It reduces exposure rather than monitors or eliminates vulnerabilities in computer systems [11]. It is important to have a detecting and monitoring system to protect important data.

An Intrusion Detection System (IDS) is an automated system that can detect a computer system intrusion either by using the audit trail provided by an operating system or by using the network monitoring tools. The main goal of intrusion detection is to detect unauthorized use, misuse and abuse of computers by both system insiders and external intruders [14]. IDS does not eliminate any preventive mechanism but provides the defense in safeguarding the computer system. In IDS, misuse and anomaly are the two types of detection approaches. Misuse detection can detect known attacks by constructing a set of signatures of attacks, while anomaly detection recognizes novel attacks by modeling of normal behaviors [25]. The outcome of this modeling is called reference model. A significant deviation from the model of reference indicates a potential threat. The anomaly detection approach is popular because it is seen as a possible approach to detecting unknown or new attacks [5] [8] [24]. Unfortunately, the anomaly detection approach suffers high false alarm especially when IDSs use pattern recognition algorithms in operational environments [10].

This paper proposes an adaptive intrusion detection model by introducing two different modules namely adaptive hybrid clustering approach based on combining the K-means , fuzzy C-means and GSA algorithms for modeling the normal patterns of a user's activity to improve attack detection and a hybrid classification approach based on combining the Support Vector Machine (SVM) and Gravitational Search Algorithm (GSA) algorithm to enhance the detection accuracy .

The remainder of this paper is organized as follows. In section 2, we presents the related work, In section 3, K-means , fuzzy C-means , gravitational search algorithm and SVM classifiers are discussed first and then, we discussed the proposed adaptive intrusion detection model in details with the system architecture. The initial results are shown in section 4, while section 5, summarizes the work and points out what we will do in future..

2. RELATED WORK

Hua TANG and Zhuolin CAO [4] proposed a machine learning based algorithm for intrusion detection, which uses a combination of neural networks and support vector machines. However, they have used all the features of the KDD cup dataset. Lee et.al [16] proposed a data-mining framework for constructing intrusion detection models by mining normal patterns of audit data. Mukkamala et.al [18] integrated data mining techniques with IDS to develop efficient security system. Shanghai and Yingxu [15] discussed that the key idea is about applying data mining Techniques namely, classification, meta-learning, association rules, and frequent episodes to audit data for computing misuse and anomaly detection models that accurately capture the actual behavior (i.e., patterns) of intrusions and normal activities. Although all these proposed detection models can detect a high percentage of old and new PROBING and U2R attacks, they miss a large number of new DOS and R2L attacks..

Theodoros Lappas et.al [23] in their work mostly focused on data mining techniques that are being used for such purposes, and then presented a new idea on how data mining can aid IDSs by utilizing bio clustering as a tool to analyze network traffic and enhance IDSs. S. Peddabachigan et.al [19] investigated some new techniques for intrusion detection and evaluated their performance based on the KDD Cup 99 Intrusion data. They have explored Decision Tree (EC4.5) and SVM as intrusion models and then designed a hybrid EC4.5-SVM model and arrived at a new ensemble approach with

EC4.5, SVM, DT-SVM. S. Sun et.al [20] proposed a hybrid intelligent system that uses a new algorithm called weighted support vector clustering algorithm, which is applied to the anomaly detection problem. Their experimental results showed that their method achieves a high detection rate with low false alarm rate.

Giacinto et al. [8] took a slightly different approach. Their anomaly IDS was based on a modular multiple classifier system where each module was designed for each group of protocols and services. The reported results showed that this approach provides a better trade-off between generalization abilities and false alarm generation than that provided by an individual classifier trained on the overall feature set. Mrudula Gudadhe et al. [17] have demonstrated a new ensemble boosted decision tree for intrusion detection system. The underlying idea of this approach is to combine simple rules to form an ensemble such that the performance of the single ensemble is improved.

3. A-IDS TECHNIQUES

In our model, we use a hybrid machine learning model based on combining the unsupervised and supervised classification techniques. Clustering approach based on combining the K-means , fuzzy C-means and the Gravitational Search Algorithm (GSA) for modeling the normal patterns of a user's activity, which combines an unsupervised clustering algorithm with the GSA technique. The technique is used as a first component for "pre-classification" to reduce the data dimensionality and improve attack detection. Then, a hybrid Support Vector Machine (SVM) and Gravitational Search Algorithm (GSA) will be used to enhance the detection accuracy. The subsequent subsections will briefly describe these techniques

3.1 K -Means Clustering

The K-means algorithm, starting with k arbitrary cluster centers in space, partitions the set of giving objects into k subsets based on a distance metric. The centers of clusters are iteratively updated based on the optimization of an objective function. This method is one of the most popular clustering techniques, which are used widely, since it is easy to be implemented very efficiently with linear time complexity [3]. However, the K-means algorithm suffers from several drawbacks. The objective function of the K-means is not convex and hence, it may contain many local minima [21].

3.1.1 *Fuzzy C-Means (FCM)*

Fuzzy c-means (FCM) is a method of clustering, which allows one piece of data to belong to two or more clusters. This method was introduced by Bezdek [2] in 1974, and it is frequently used in pattern recognition. Fuzzy c-means (FCM) algorithm is one of the most popular fuzzy clustering techniques because it is efficient, straightforward, and easy to implement. However, FCM is sensitive to initialization and is easily trapped in local optima .

3.2 Gravitational Search Algorithm (GSA)

Gravitational search algorithm (GSA) is a newly developed stochastic search algorithm based on the law of gravity and mass interactions (Rashedi et al., 2009). GSA is based on the physical law of gravity and the law of motion. The gravitational force between two particles is directly proportional to the product of their masses and inversely proportional to the square of the distance between them [6]. GSA a set of agents called masses has been proposed to find the optimum solution by simulation of Newtonian laws of gravity and motion [20]. GSA has a greater ability to explore the whole search space and avoids being trapped in the local optima very well .The global search ability of GSA is superior to that of other well-known heuristic optimization methods in most cases [1,6,7] .

3.3 Support Vector Machine (SVM)

Support vector machines (SVM) are learning machines that plot the training vectors in high dimensional feature space, labeling each vector by its class. SVMs classify data by determining a set of support vectors, which are members of the set of training inputs that outline a hyper plane in feature space. Computing the hyper plane to separate the data points leads to a quadratic optimization problem. There are two main reasons that we used SVMs for intrusion detection. The first reason is that its performance is in terms of execution speed, and the second reason is scalability. SVMs are relatively insensitive to the number of data points, and the classification complexity does not depend on the dimensionality of the feature space [22].

3.4 System Architecture

Figure.1 shows the overall architecture of the Adaptive Intrusion Detection Model, which is now under construction. The functionality of each component of the system can be described as follows.

3.4.1 Adaptive Hybrid Clustering Approach

In this module, we will apply clustering approach based on combining the K-means , fuzzy c-means and GSA algorithms for obtaining the normal patterns of a user's activity, which combines an unsupervised clustering algorithm with the GSA, the technique is used as the first component for pre-classification to improve attack detection.

3.4.1.1 Hybrid KM-GSA Algorithm

K-means is a simple and efficient algorithm that is widely used for data clustering. However, its performance depends on the initial state of centroids and may trap in local optima. The gravitational search algorithm (GSA) is an effective method for searching the problem space to find a near optimal solution [7]. The KM-GSA algorithm is a hybrid data clustering algorithm based on GSA and k-means (GSA-KM), which uses the advantages of both algorithms. The hybrid KM-GSA algorithm will be built on two main steps. In the first step, KM-GSA applies k-means algorithm on selected dataset and tries to produce near optimal centroids for desired clusters , which will be applied by the GSA algorithm. The output of the k-means algorithm, which has been achieved in the previous step, will be applied as candidate solutions by the GSA algorithm in the second step. This process generates a high-quality initial population, which will be used in the next step by the GSA algorithm. Finally, in the third step, GSA will be employed for determining an optimal solution for the clustering problem.

3.4.1.2 Fuzzy C-Means (FCM) Algorithm

Fuzzy c-means (FCM) algorithm is one of the most popular fuzzy clustering techniques because it is efficient, straightforward, and easy to implement [2]. The clustering approach based on fuzzy c-means algorithm will be applied to the selected dataset for modeling the normal patterns of a user's activity to improve attack detection

The key in this module is to improve the performance of the SVM classifier. The clustering algorithm is used for analysis data set to build the normal behavior model and reduce dataset. Clustering analysis help to find the boundary points, which are the most qualified data points to train SVM between two classes. Thus, the system will be greatly shortened the training time, and has also achieved better detection performance in the resultant SVM classifier.

3.4.2 *Hybrid Classification Approach*

This module attempts to increase the classification accuracy rate in SVM classification by employing a hybrid approach based on combining the gravitational algorithm search (GSA) and support vector. The gravitational algorithm search (GSA) will be used to optimize the input selection for the SVM kernel parameter setting.

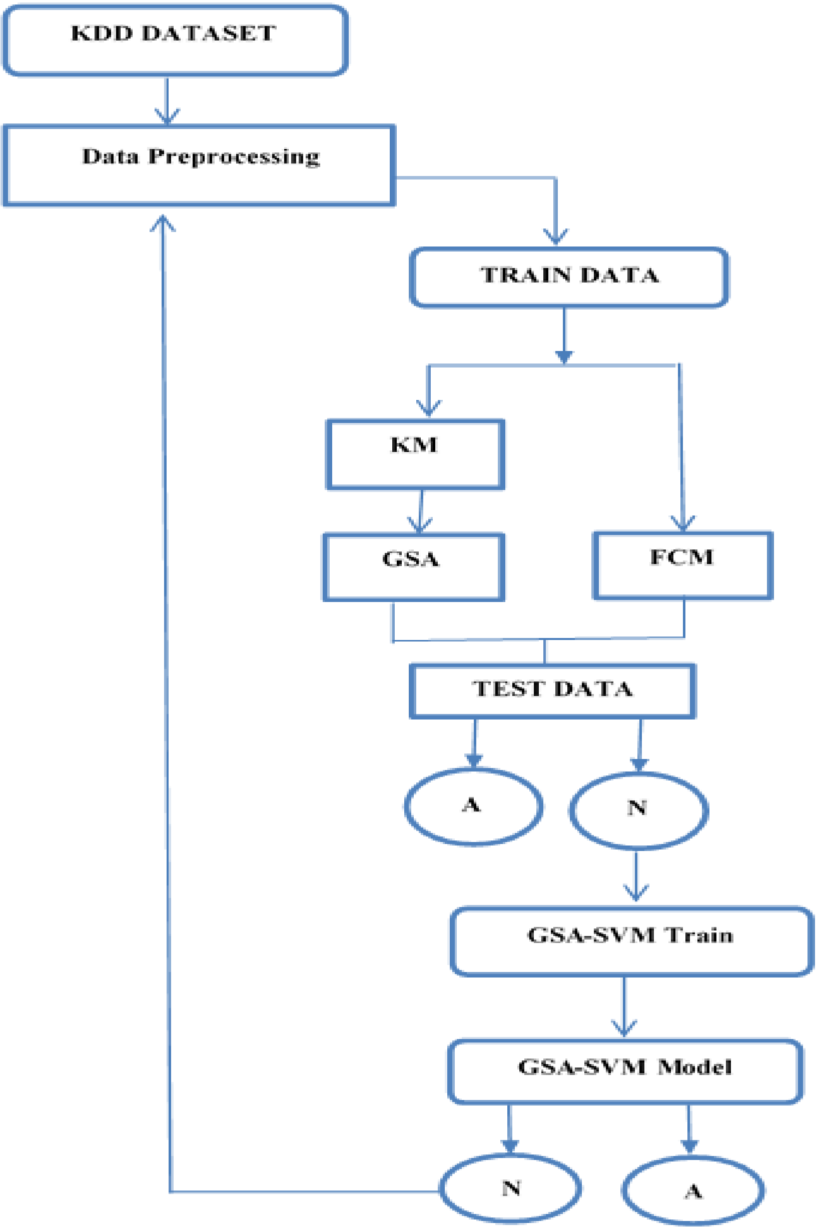


Fig .1 Proposed Adaptive Intrusion Detection Model

The hybrid GSA-SVM algorithm searching process, starts with initializing population of agents by using the result from phase 1(normal profile), and searches for the optimal agent (mass) iteratively. Each agent represents a candidate solution.

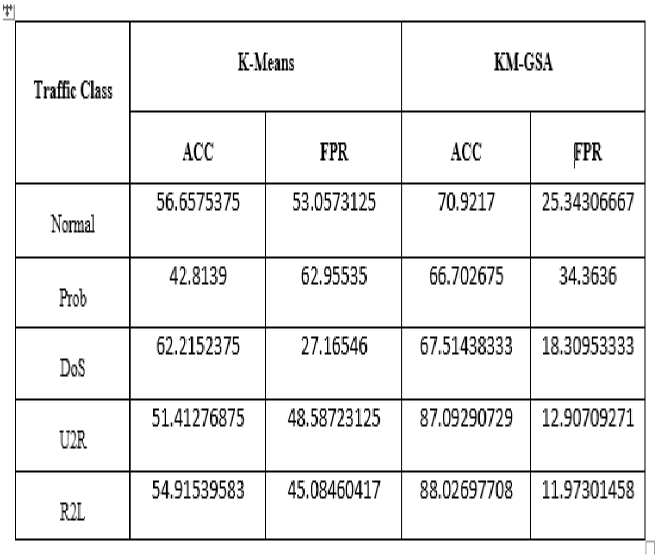
SVM classifier is built for each candidate solution to evaluate its performance. GSA algorithm guides the selection of potential subsets that lead to best prediction accuracy. The algorithm uses the most fit agents to contribute to the next generation of n-candidate agents. Thus, on the average, each successive population of candidate agents fits better than its predecessor. This process continues until the value of the fitness function converges.

4. INITIAL RESULTS

This research utilized the dataset KDD Cup 1999, which is the largest publicly available and sophisticated benchmarks for researchers to evaluate intrusion detection. This study, as most of the literature research, used 10% version of the dataset consisting of 494,020 traffic connections with similar ratio of attacks as in the full dataset [26,27]

We got the initial results of the first stage, which includes using both the hybrid KM-GSA and KM for clustering to obtain the normal patterns of a user's activity. The results showed that the hybrid KM-GSA achieve better results than KM , especially for the accuracy(ACC) and false positive rates (FPR) as shown in tables 1and 2.

Table 1: Comparison of accuracy and false positive rate between K-Means and GSA algorithms



Traffic Class	K-Means		KM-GSA	
	ACC	FPR	ACC	FPR
Normal	56.6575375	53.0573125	70.9217	25.34306667
Prob	42.8139	62.95535	66.702675	34.3636
DoS	62.2152375	27.16546	67.51438333	18.30953333
U2R	51.41276875	48.58723125	87.09290729	12.90709271
R2L	54.91539583	45.08460417	88.02697708	11.97301458

5. FUTURE WORK

Three different machine learning algorithms GSA, KM, FCM and SVM are used to construct the adaptive hybrid clustering. The hybrid classifier approaches provide efficient techniques for anomaly based intrusion detection. This model is now at the infant stage of development. More results could be obtained once we finish deploying the system.

6. REFERENCES

[1] Abarghouei ,A. , Ghanizadeh, A. , and hamsuddin, M.," Advances of soft computing methods in edge detection", Int. J. Advance Soft Comput, Appl., vol. 1, n. 2, 2010, 162-203.

[2] Bezdek,J.1974. Fuzzy mathematics in pattern classification. Ph.D. thesis, Ithaca, NY: Cornell University.

[3] Chen ,Y., and Ye, F.2004. Particle swarm optimization algorithm and its application to clustering analysis. In Proc. The IEEE International Conference on Networking in Sensing and Control.

[4] Hua TANG, D., and Zhuolin CAO, " Machine Learning-based Intrusion Detection Algorithms", In Journal of Computer Information Systems, Vol.5, No.6 ,1825-1831, 2009.

[5] Denning D,E. (1987). An Intrusion Detection Model. IEEE Transactions on Software Engineering. SE-13.

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

[7] Rashedi,E., Nezamabadi,H., Saryazdi, S.," GSA: A gravitational search algorithm", Information Sciences, vol. 179, 2232-2248, 2009.

[8] Forrest, S., Hofmeyr, S., Somayaji, A., and Longstaff, T. (1996). A sense of self for Unix Processes. In Proceedings of the 1996 IEEE Symposium on Security and Privacy. IEEE Computer Society.

[9] Giacinto, G., Roli, F,. and Didaci, L. (2003a). " Fusion of multiple classifiers for intrusion detection in computer networks," Journal of Pattern Recognition, Vol. 24 , 1795-1803.

[10] Giorgio, G., Fabio, R., Luca, D,. 2007. Fusion of multiple classifiers for intrusion detection in computer networks, Proc. IEEE Conference in Network Security.

[11] Gosh, A. K., Schwartzbard, A., and Schatz, M. (1999). Learning Program Behavior Profiles for Intrusion Detection. In Proceeding of the Workshop on Intrusion Detection and Network Monitoring.

[12] Hossain, M., and Bridges, S. (2001). A Framework for an Adaptive Intrusion Detection System With Data Mining. In Proceedings of the 13th. Annual Canadia Information Technology Security Symposium.

[13] Jiong Zhang, Mohammad Zulkernine.2006. Anomaly based network intrusion detection with unsupervised outlier detection. Proc. IEEE Communication Society.

[14] Kim, J., Bentley, P., Aickelin, U., Greensmith, J., Tedesco G., and Twycross, J., " Immune System Approaches to Intrusion Detection – A Review", Natural Computing. 2007, 413-466.

[15] Shanghai ,L., and Yingxu, L.2009. A Data Mining Framework for Building Intrusion Detection Models Based on IPv6. In Proceedings of the 3rd International Conference and Workshops on Advances in Information Security and Assurance. Seoul, Korea, Springer-Verlag.

[16] Lee, W., and Stolfo, S.1998. Data mining approaches for intrusion detection. In Proceedings of the 7th USENIX security symposium .

[17] Mrudula, G., Prakash, P., and Kapil , W.2010. A new data mining based network intrusion detection model. International Conference on Computer and Communication Technology.

[18] Mukkamala, R., Gagnon ,J., and Jaiodia, S.2000. Integrating data mining techniques with intrusion detection methods. In Research Advances in Database and Information systems security.

[19] Peddabachigari,S., Abraham,A., C., Grosan, and Thomas,J, " Modeling intrusion Detection system using hybrid In intelligent systems", Journal of Network and Computer Applications, 2007.

[20] Sun ,S., and Wang ,Y.2009. A Weighted Support Vector Clustering Algorithm and its Application. In Network Intrusion Detection In First International Workshop on Education Technology and Computer Science Vol. 1.

[21] Selim, S., and Ismail, M. "K-means type algorithms: a generalized convergence theorem and characterization of local optimality", IEEE Transaction of Pattern Analysis Machine Intelligent, 1984.

[22] Srinivas ,M., Guadalupe,J., and Amdrew,.S.2002. Intrusion Detection using Neural Networks and Support Vector Machines. In Proceedings of the International Joint Conference on Neural Networks.

[23] Theodoros, L., and Konstantinos, P., " Data Mining Techniques for Network Intrusion Detection" System. In techrepublic January, 2007.

[24] Warrander, C., Forrest, S., and Pearlmutter, B. (1999). Detecting intrusions using system calls: alternative data models. In proceedings of the 1999 IEEE Symposium on Security and Privacy. IEEE Computer Society.

[25] Xu, X., and Wang, X. (2005). An Adaptive Network Intrusion Detection Method Based on PCA and Support Vector Machines. Proceedings of First International Conference on Advanced Data Mining and Applications, ADMA 2005, Wuhan, China.

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

[27] Mukkamala, S., Sung A. H. and Abraham, A. (2003). Intrusion Detection Using Ensemble of Soft Computing Paradigms. Proceedings of 3rd. International Conference on Intelligent Systems Design and Applications. Advances in Soft Computing, Springer Verlag, Germany, 239-248.