INTERNATIONAL JOURNAL OF RESEARCH IN COMPUTER APPLICATIONS AND ROBOTICS
ISSN 2320-7345

AN ADAPTIVE INSTRUCTION DETECTION SYSTEM USING K-NEAREST NEIGHBORS CLASSIFIER WITH KMEANS AND KMEDIODS CLUSTERING ALGORITHMS

Muhammad M. Kwafha

mkwafha@bau.edu.jo

Information Technology Department, Al-Huson University College, Al-Balqa Applied University, Irbid, Jordan

Abstract

In recent years, the number of attacks has increased and intrusion detection has become the standard for information assurance. The goal of any intrusion detection system is to help computer systems prepare for and respond to attacks. The ultimate goal is to detect these attacks before access or during the access process. Firewalls do not offer complete protection and must nevertheless be supplemented by an intrusion detection system. Various tools and methods are used to monitor user and system events to detect illegal and abnormal activities in networks and systems. All these tools and methods are called intrusion detection systems (IDS). These systems have been implemented keeping in mind the use of artificial intelligence (AI) approaches. , such as Genetic Algorithm (GA), Decision Tree (DT), Expert System (ES), Neural Network (NN).In our research, we built an intrusion detection system that uses K-Nearest Neighbors (KNN) classifier with the help of KMeans and KMediods clustering algorithms. The system was trained and evaluated on the KDD 99 dataset. The results gave a good indication about the ability of the system to detect type of inserted data(attack or normal) especially for Normal, denial of service(DoS) and Prob types also they displayed the superiority of KMediods over KMeans due to outlier problem in KMeans.

Keywords: KNN, KMeans, KMediods, Intrusion Detection, Clustering

1. Introduction

Intrusion is one of the most important problems of e-security. Government departments, business organizations and individuals all of them working on increasing the security level of their systems through increasing their concerns about the issues of security.

Information infrastructure is very important and essential; we rely on it to support critical operations in huge systems such as banking, telecommunications, and other systems. Thus, intrusions into informational

systems become a significant threat on societies. It compromises the security of an information system through various means.

Attacks have a lot of classifications according to the researchers' views. There are four main categories of attacks according to the attackers' behaviors [1] that can be summarized as: DoS, Prob, User to Root (U2R) and Remote to Local(R2L).

To detect attacks of any type we mention above, we need to use one or more of IDS tools and approaches. The concept Intrusion Detection (ID) can be defined as: "any set of actions that attempt to compromise the Confidentiality, Integrity, and Availability (CIA) of a resource"[2]. ID aims to prevent Intrusions or at least to detect anomaly. It refers to a lot of techniques that have been developed to protect systems against malicious attacks over the past several years [3].

Typically, an ID system follows a two-step process. The first step includes: inspection of the system's configuration files to detect unacceptable settings. The second step includes: catch known methods of attack and recording system responses, these procedures are network-based, they considered the active component [4].

Detecting Intrusions needs a lot of procedures to handle their different types and resources. These procedures cost organizations a lot of effort, time and money. Intrusions compromise the Confidentiality, Integrity, and Availability (CIA) of resources, especially information [5].

Recently intrusion detection techniques have shifted from user-based and connection-based to process-based intrusion detection such as Statistical Analysis, Neural Network, Rule-Based Analysis, Bayesian Network, Finite State Machine and Data Mining [6].

Clustering is the method of grouping objects into meaningful subclasses so that members of the same cluster are quite similar and members of different clusters are quite different from each other. Therefore, clustering methods can be useful for classifying log data and detecting intrusions. In the proposed system we used two clustering methods called KMeans and KMediods clustering method.

2. Related Work

Many previous researches and studies shed light on network identification systems and associated methods and approaches to this topic. Many of these studies deal in depth with the main topic of the proposed research that facilitates our movement by clearing the route and direction of our movement to move towards, we will summarize some of these previous studies below:

Chen R. C., Cheng K. F., Hsieh C. F. [6] explained the differences between identification systems and firewall. The main difference between them is that the firewall is a manual passive defense system, in which the IDS can collect online packets from the network, and then it can monitor and analyze these packets. The IDS acts as a “second line of defense”.

Kaxienko P., Dorosz P. [7] gave a survey about IDS; it contains definition of IDS and what is not related to IDS, also defines different terminologies related to IDS and presents a brief review of the IDS architecture.

Scarfone K., Mell P.[8] Identification is defined as the process of monitoring events occurring in a computer system or network and analyzing them for signs of possible incidents, which are violations or imminent threats of violations of computer security policies, acceptable use policies or standard security practices.

Rhodes B.C., Mahaffey J.A., Cannady J.D. [9] described an innovative approach to identification that used self-organizing NN to recognize anomalies in a computer network data stream, where animalized detection attempted to identify the user model and detect the intrusion by knowing their identity, behaviors and for monitoring.

Sinclair Ch., Pierce L., Matzner S. [10] attempted to create an application to enhance the knowledge domain to detect a wide range of intrusions using machine learning (ML) techniques to create rules of the. Therefore, the amount of information must be processed less by humans than before. The planned work

will apply the Network Exploitation Detection Analyst Assistant (NEDAA) on a military subnet that combines artificial intelligence rule generation with conventional expert system (ES).

Pan Z. S., Lian H., Hu G. Y., Ni G. Q. [11] presented an integrated intrusion detection model based on neural network and expert system. It aims to leverage the classification capabilities of the neural network for unknown attacks and the expert system for unknown attacks. known attacks. It was designed to improve performance to detect all intrusions.

Kukielka P., Kotulski Z. [12] proposed an application of neural network (NN) as an application tool in IDS. The method proposed in this paper can use the NN learning property to discover new attacks.

Siraj M. M., Maaroof M. A., Hashim S. Z. M [13] proposed a new hybrid clustering model, based on unit range improvement (IUR), principal component analysis (PCA) and a unsupervised learning.

Faraoun K. M., Boukelif A. [14] proposed a new technique to improve the learning capabilities and reduce the computational intensity of a competitive learning multilayer neural network using K-means clustering algorithm. The K-means algorithm is first applied to the training dataset to reduce the amount of samples to present to the neural network, automatically selecting an optimal set of samples. When it is difficult to clearly define the rules as is the case in abuse detection or anomaly detection, the neural network is an appropriate method to do so (i.e. to define intrusions through training and learning).

Kato's V. [15] focused on evaluating the statistical method used by these methods to examine the data used in experimental identification. They evaluated 1200 observations with 42 features (variables), but the actual number of features used was only 30 features, these features were listed in Knowledge Discovery and Data Mining 99 (KDD 99).

3. Material and Methods

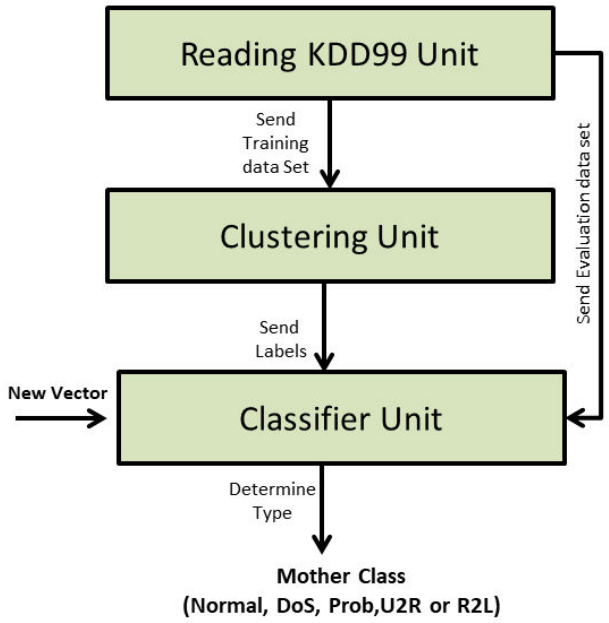
The objective of the proposed system is to perform the detection process more effectively, to do that a set of steps must be carried out as shown ulterior. Figure [1] shows the general framework of the prosed system

- Determine a training data from the set of KDD99 [16].
- Clustering training data using KMeans and KMediods algorithms.
- Extract the label for each cluster
- Use those labels in classification process.

3.1 Clustering Unit

It is the process of organizing objects into groups name clusters whose member are similar some way .where the cluster is a collection of objects which are similar to each other and share certain properties, they are also dissimilar to objects outside the cluster[17].

In clustering classes are unknown, we need to discover them from the data and grouping point into clusters based on how near they are one other. For this reason clustering sometimes referred to as unsupervised classification which means that class unknown and we look to discover them from the data. There exist many clustering algorithms in our work, we used KMeans and KMediods.



Figure[1]: Framework of proposed System

3.1.1 KMeans Method

The main idea of this technique is to define k centroids or means one for each cluster [68,70].The k-means method is simple and reasonably effective. The computational complexity of the algorithm is a function to the number of objects, number of clusters and number of iterations. KMeans algorithm is very sensitive to outlier's objects Figure [2] shows the main steps of KMeans procedure.

Input: The number of clusters K and a dataset for intrusion detection

Output: A set of K-clusters that minimize the squared-error criterion.

Algorithm:

1. Initialize K clusters (randomly select k elements from the data points as the Means).
2. Repeat Until Cluster Structure does not change.

2.1. Determine the cluster to which each data point belongs i.e. to the closest Mean.("closest" here is defined using any valid distance metric; " use Euclidean distance").

2.2. Calculate the New Means of the Clusters.

2.3. Change Clusters Centroids to means obtained, using step 2.1.

Figure[2]: Main Steps of KMeans

3.1.2 KMediods Method

KMediods it is more robust to noise and outliers as compared to KMeans. KMediods can be defined as the object of a cluster, whose average dissimilarity to all the objects in the cluster is minimal i.e. it is a most

centrally located point in the given data set[18.19].The objective of KMediods clustering is to find a non-overlapping set of clusters such that each cluster has a most representative point, i.e., a point that is most centrally located with respect to some measure, e.g., distance. These representative points are called KMediods. Figure[3] shows the main steps of KMediods.

Input: The number of clusters K and a dataset for intrusion detection

Output: A set of K-clusters that minimize the squared-error criterion.

Initialize: randomly select k elements from the n data points as the Medoids

Associate each data point to the closest Medoid. ("closest" here is defined using any valid distance metric; "use Euclidean distance").

3. For each Medoid m

3.1. For each non-Medoid data point o

3.1.1. Swap m and o and compute the total cost of the configuration

4. Select the configuration with the lowest cost.

5. Repeat steps 2 to 5 until there is no change in the Medoid in the old and new model.

Figure[3]: Main Steps of KMediods

3.1.3 Distance measure

An important step in most clustering is to select a distance measure, which will determine the similarity between nodes or objects of corpus. Here we used the Euclidian formula to compute the distance between elements. We considered the object as a vector in the space then computed the distance between object via Euclidian formal.

3.2 Classifier Unit

It fetches the label of each cluster that result from clustering unit in aim to determine the type inserted vector (Normal or Attack). In our work, we will useK-Nearest Neighbors (KNN) classifier for making the detection. It works by computing distance between labels and inserted vector; the minimum distance is taken to determine the mother class and type of vector. Figure[4] shows the main steps of KNN algorithm.

1. Input Labels and New Item

2. Set i=1

3. Loop

3.1 Compute Distance(New Item, $Label_i$)

3.2 If i==1 Then

Min=Distance(New Item, $Label_i$)

Mother Class= i

End if

3.3 If Min> Distance(New Item, $Label_i$) Then

Min=Distance(New Item, $Label_i$)

Mother Class= i

End if

3.4 Set i=i+1

3. if i> Number of labels Then Step 4 else Step 3

4. Return mother class

Figure[4]: Main steps of KNN algorithm

4. Result

The constructed system was trained and evaluated using KDD 99 dataset. We used 600 record for training and 400 for evaluation the result of correctness and error are given in table[1] and table[2] which shows the results of KMeans and KMediods respectively.

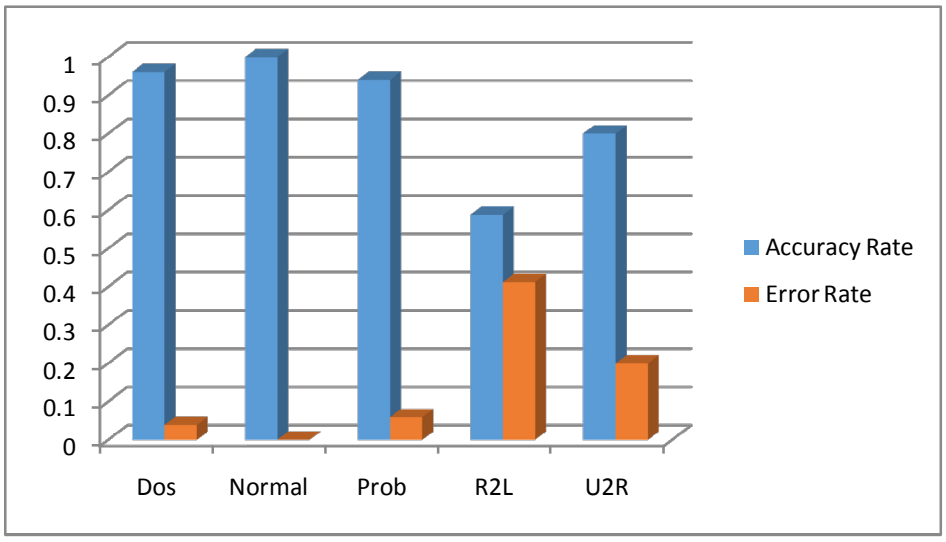
Table [1]: Accuracy and Error rate for KMeans algorithms labels

Class Name	Accuracy Rate	Error Rate
Dos	**0.961**	**0.039**
Normal	**1**	**0**
Prob	**0.941**	**0.059**
R2L	**0.588**	**0.412**
U2R	**0.8**	**0.2**

Table [2]: Accuracy and Error rate for KMediods algorithms labels

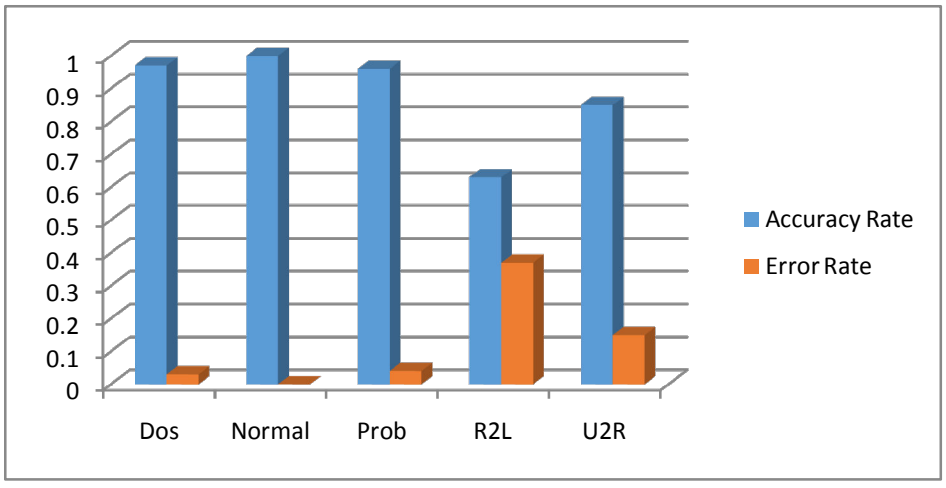
Class Name	Accuracy Rate	Error Rate
Dos	**0.97**	**0.03**
Normal	**1**	**0**
Prob	**0.96**	**0.04**
R2L	**0.63**	**0.37**
U2R	**0.85**	**0.15**

Figure[5]shows the accuracy and error rate graphically for the classifier that uses KMeans. The result gives a very good indication about the ability of classifier to detect Normal, DoS and Prob type, while a good result for R2L and U2R.



Figure[5]: KMeans Accuracy and Error Rate

The result of accuracy and error rate for KMediods labels classifier is shown Figure [6] which gives also a very good indication about the ability of classifier to detect Normal, DoS and Prob type, while a good result for R2L and U2R.
According to Figure[5] and Figure[6] the result of KMediods labels classifier is more better than the result of KMeans labels classifier due to no outlier in KMediods algorithm.



Figure[6]: KMediods Accuracy and Error Rate

5. Conclusion

The constructed system gives a very good result in detecting Normal, DoS and Probtypes. The results of KMediods are better than KMeans because there is no outlier in KMediods. We look in the future to use another types of classifiers and artificial intelligence algorithms with those clustering algorithms in order to improve them.

References

1. Bridges S. M., Vaughn R. B.(2000) "Intrusion Detection Via Fuzzy Data Mining", Accepted for 12th Annual Canadian Information Technology Security Symposium, June 19-23 2000.
2. Denning D. E. (1987) "An Intrusion-Detection Model", IEEE Transactions on Software Engineering, Vol. Se-13, No. 2, pp. 222-232 February 1987.
3. Izbasa C. "A Neural Network-Based IDS," Software, www.ieat.ro/researchreports/ids.pdf/download..
4. Kozushko H.(2003) "Intrusion Detection: Host-Based and Network-Base Intrusion Detection Systems", Independent Study, Software, http://infohost.nmt.edu/~sfs/Students/HarleyKozushko/Papers/IntrusionDetectionPaper.pdf
5. Xu X. (2006) "Adaptive Intrusion Detection Based on Machine Learning: Feature Extraction, Classifier Construction and Sequential Pattern Prediction",International Journal of Web Services Practices, Vol.2, No.1-2, pp. 49-58
6. Chen R. C., Cheng K. F., Hsieh C. F.(2009) "Using Rough Set And Support Vector Machine For Network Intrusion Detection", International Journal of Network Security & Its Applications (IJNSA), Vol. 1, No 1, pp.1–13., April 2009.
7. Kazienko P., Dorosz P.(2004), "Intrusion Detection System(IDS) part 1- (network intrusion; attack symptoms; IDS tasks; and architecture)", updated: June 14, 2004.
8. Scarfone K., Mell P.(2007) "Guide to Intrusion Detection and Prevention Systems (IDPS)", Recommendations of the National Institute of Standards and Technology Computer Security Division Information Technology Laboratory National Institute of Standards and Technology, February 2007
9. Rhodes B. C., Mahaffey J. A., Cannady J. D.(2000) "Multiple Self-Organizing Maps for Intrusion Detection," Paper for submission to the 23rd NISSC, Software.
10. Sinclair Ch., Pierce L., Matzner S.(1999), "An Application of Machine Learning to Network Intrusion Detection", 15th Annual Computer Security Applications Conference Phoenix, Arizona , December 6-10, 1999 .
11. Pan Z. S., Lian H., Hu G. Y., Ni G. Q. (2005) "An Integrated Model of Intrusion Detection Based on Neural Network and Expert Systems", The 17th IEEE international Conference on Tools with Artificial Intelligence.
12. Kukielka P., Kotulski Z. (2008) "Analysis of Different Architectures of Neural Networks For Applications in Intrusion Detection Systems", International Multi-conference on Computer Science and Information technology, pp.807 – 811.
13. Siraj M. M., Maaroof M. A., Hashim S. Z. M. (2009) "Intelligent Alert Clustering Model for Network Intrusion Analysis", Int. Advance. Soft Company. Appl. , Vol. 1, No. 1, July 2009.
14. Faraoun K. M., Boukelif A.(2006) "Neural networks learning improvement using the K-means clustering algorithm to detect network intrusions", Software
15. Katos V.(2007) "Network intrusion detection: Evaluating cluster, discriminant, and logit analysis," Science Direct, Information Sciences 177, pp. 3060–3073.
16. Langley P., Simon H. A.(1995), "Applications of Machine Learning and Rule Induction", Communications of ACM, Volume 38, Issue 11, New York, USA, November 1995.
17. Abu-Khalaf M. (2004) " EE 5322 Neural Networks Notes", October 24, 2004 , Personal Study , Software ,arri.uta.edu/acs/abumurad/EE5322/EE5322_NN_notes.pdf
18. Kaplantzis S., Mani N.(2006) "A Study On Classification Techniques For Network Intrusion Detection" . software
19. D. Lai and N. Mani, “Support vector machines and linear stationary models,” conversion report, Monash University, 2003.