Computing, Information Systems, Development Informatics & Allied Research Journal
Vol. 6 No. 4. – www.cisdijournal.net

A Network Intrusion Detection System: Enhanced Classification via Clustering Model.

Balogun, A. O., Balogun, A. M., Adeyemo, V. E. & Sadiku, P. O.
Department of Computer Science,
University of Ilorin
Ilorin, Kwara State, Nigeria.
bharlow058@gmail.com

ABSTRACT

The aim of developing an IDS is to build a system that oversee the general protection of a network from attacks both from within and without, and doing so accurately. Optimization of IDS has also being receiving attention from the research community due to its large volumes of security audit data. In developing an IDS, most dataset used have high dimension in which only few attributes are needed for building an IDS – feature selection is used to solve this little problem. In this paper, we present and analyze the performance of some machine learning algorithm which performs classification via clustering using the KDDcup'99 dataset. Using the WEKA tool, simulations were ran and results was deduced after applying the proposed models to the dataset containing all the type of attacks.

Keywords: Classifier, Clustering, Data mining, Feature Selection, Intrusion Detection, KDD Dataset, Machine Learning..

1. INTRODUCTION

Consequent to the ever increasing rate of networks of computers and it related usage, increment to global information infrastructure is on high demand resulting to large repository of information which if left unguarded can lead to dangerous situations all around the globe. The protection of these networks of computers and the information saved or passing across it from cyber-attacks, computer viruses and worm, potential information theft and leakage is of utmost importance. Defending against these malicious hazards had saw to the development of a lots of computer security techniques including firewalls, cryptography, anomaly and misuse intrusion detection system among all others, wherein intrusion detection had strong ground for protection amidst them all due to its defending complex and dynamic intrusion behaviors[1].

Intrusion detection system, a type of security management system employed to collect and analyze information within a computer network or in a computer in order to help build a protection mechanism from various attack, is recording massive research works on its system using data mining which have been flooded by many researchers recently as it have the potency of classifying and detecting anomalies within a network [2]. IDS detection efficiencies are becoming better day by day as improvements are being discovered.

However, most IDS that are being developed are created using classifier and little effort are made on clustering technique in comparison to the rate of classification technique. Clustering technique which is an unsupervised form of classification [3], a data mining category of machine learning algorithms that separate the training dataset based on similar characteristics [4] can be also be used for classification purposes – which is the aim of this research. Clustering algorithms breaks dataset into groups in respect of the information found in the data describing it. It ensures that the attributes of a certain group are alike and distinct to other attributes in other groups [5].

Conclusively, this paper will present an IDS model that will be developed using classification via clustering technique and the considered algorithms will be two famous clustering algorithms which are Expectation Maximization (EM) and K-Means algorithms. More so, the performance evaluation of these algorithms was carried out on KDDCup'99 datasets and these algorithms performance will be enhanced by preprocessing the dataset using feature reduction technique.

2. RELATED WORKS

A research work carried out by [6] presented the usage of EM and K-means on random projection and feature reduction was done on the dataset using the principal component analysis (PCA). It was observed in the research that despite PCA computational intensity, it was only marginally better than random projection. According to [7], the authors also applied EM and K-Means and also performed feature selection using J48 (decision tree) algorithm. They evaluated its performance on R2L, and Vote datasets using both the original and the reduced dataset. A hybridized model of EM, K-Nearest Neighbor and Genetic Algorithms was used to remove data that are difficult to learn in order to achieve better results [8].

However, this research haven study the way clustering technique has been used focused on the way it can be used to build an IDS model. This is done by evaluating the classifiers performance on the KDDCup'99 dataset and also on each category of attack, thereby exposing its potency for categorizing attack generally and specifically. This research is pivotal in the case of having to detect and classify an unknown or unlabeled attack.

3. METHODOLOGY

This paper presents classifiers algorithms belonging to classification via clustering categories. The algorithms classifies by making sure the number of clusters they generate tallies with the number of labels found in the dataset in order to obtain a useful model. The selected classifiers are Expectation Maximization (EM), and K-Means. More so, the reduction of the dataset dimensionality popularly called feature selection was carried out using CFS Subset attribute evaluator via Best First search method, serving as the data preprocessing process before the algorithms got trained and tested for analysis. This algorithm on its own selects the vital attributes needed for classification.

The choice of dataset used in this paper is aimed at knowing the overall performance of the selected algorithms for classifying attacks and also to know how best in can classify an attack. We used the 10% of KDDCup'99 dataset (which contain one type of normal data and 24 various attacks categorized under four groups namely DoS, Probing, U2R, and R2L) and the datasets containing each category of attack was also used.

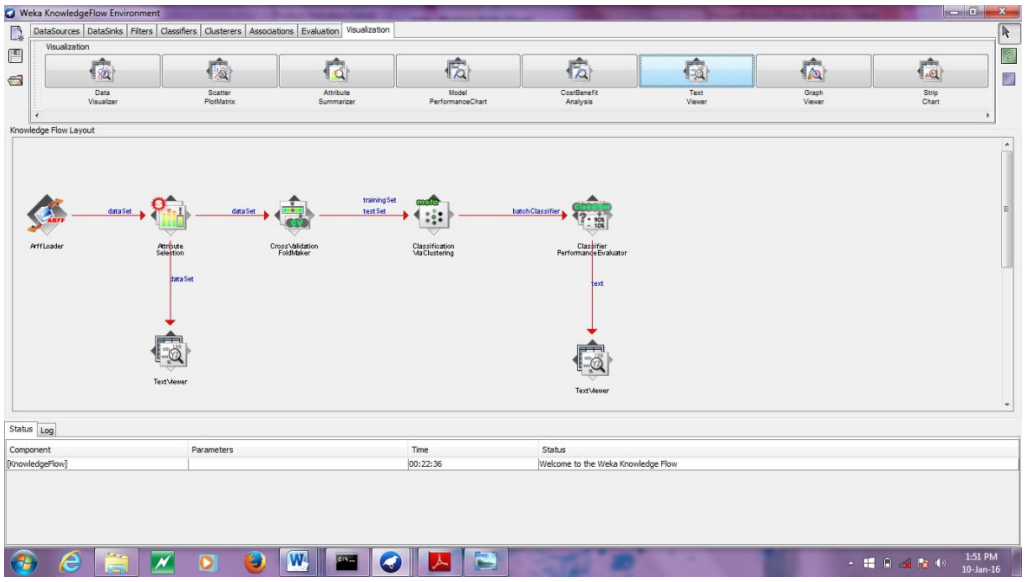

Figure 1: Proposed System Architecture (WEKA).

Arff Loader: This loads the datasets which will be used to train and test the algorithm.

Attribute Selection: This module see to data preprocessing. The CFS Evaluator is used here with "Best First" search method.

Cross Validation Maker: This is the process of breaking the dataset in 10 fold (which is the default fold) and passing the fold one after the other to the classifier.

ClassificationviaClustering: This module is where the clustering algorithm is selected, where training and testing of the selected algorithm is done.

Classifier Performance Evaluator: This module evaluates and measures the performance of the selected algorithm.

Text Viewer: This module receives and display outputs in text format from any module.

3.1 Performance Evaluation

The results of the classifiers used will be evaluated and measure using the following parameters: correctly classified instances (%), incorrectly classified instances (%),TP (True Positive) rate, FP (False Positive) rate, and TT (Training Time of the algorithm on each dataset).

3.2 Evaluation Setup

The experiments were carried out on a HP probook 6470b laptop with the following configurations Intel(R) Core(TM)i5-3230M, CPU 2.60GHz, 6GB RAM (5.55 GB usable), 64-bit operating system whose platform is Microsoft Windows7 Professional (Service Pack 1) . The latest Weka – an oen source machine learning package was used for setting up the experimental and evaluation environment (Weka 3.6.11).

3.3 Feature Selection Algorithm

Cfs Subset Evaluator: is a supervised attribute filter that can be used to select attributes. It works by evaluating the worth of a subset of attribute by considering the individual predictive ability of each feature along with the degree of redundancy between them. Its parameter was set as: locally predictive= True, missingSeparate= False.

Best First Search Method: searches the space of attribute subsets by greedy hillclimbing augmented with a backtracking facility. Its parameter was set to: direction= Forward, lookupCacheSize= 1, searchTermination= 5, startset= Nil.

3.5 Classifier Algorithm

Expectation Maximization: is an iterative model based clustering method for solving problems, applied where data contain latent variables or considered incomplete. For this experiment, EM was performed using it these parameters: maxIterations= 100, seed= 100.

K-Means: K-means algorithm classifies objects by their membership to one of the k groups, k chosen a priori. Determining a cluster membership is based on the centroid of the group after being calculated, each object to the group is assigned with the closest centroid. We ran K-Means with the following parameters: distance\Function = EuclideanDistance – R first-last, maxIteration= 500, numClusters = 2, seed = 10.

4. RESULTS AND DISCUSSION

The results of the data preprocessing module are discussed below. The feature selection which is the reduction of the dimensionality of the dataset before training and testing of the classifier was carried out. Table 1 below shows the result of applying CFS attribute evaluator using Best First search method on the datasets.

Table 1: Details of the datasets used for classifiers evaluation

Dataset	No of Initial attribute	No of Actual training attribute (with label)	No of Instances
10% KDD Cup	42	12	487271
Dos	42	9	387504
Probing	42	7	4107
Remote – to – Local	42	7	1126
User – to – Root	42	6	52

Table 2: Performance evaluation of the classifier algorithms – correctly classified instances, incorrectly instances, TP rate, FP rate, and Training Time (TT).

Classifier		KDD	Dos	Probe	R2L	U2R
Expectation Maximization (EM)	Correctly Classified Instances (%)	90.0519	92.0602	78.4027	93.2504	61.5385
	Incorrectly Classified Instances (%)	9.8865	5.2278	8.4246	6.3943	38.4615
	True Positive Rate	0.901	0.946	0.903	0.936	0.615
	False Positive Rate	0.062	0	0.04	0.283	0.3
	Training Time (secs)	5437.08	5913.09	32.62	1.01	0.14
K- Means	Correctly Classified Instances (%)	78.8879	99.1977	61.7969	87.389	65.3846
	Incorrectly Classified Instances (%)	21.1121	0.8023	38.2031	8.6146	34.6154
	True Positive Rate	0.789	0.992	0.618	0.91	0.654
	False Positive Rate	0.073	0.016	0.222	0.732	0.359
	Training Time (secs)	9.03	6.91	0.03	0.02	0.02

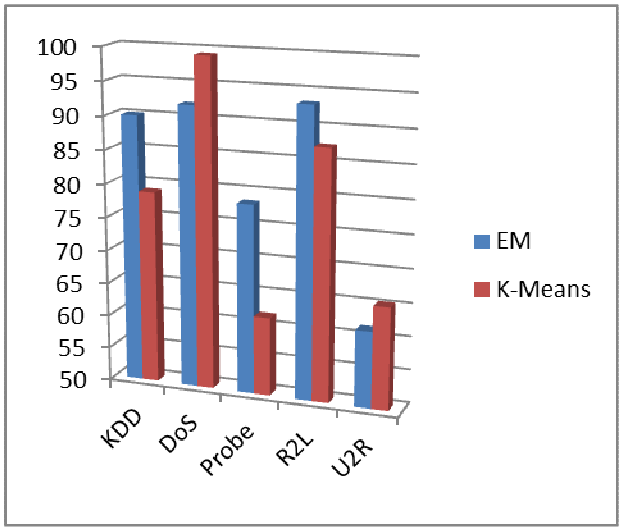

Figure 1: Accuracy of the Algorithms.

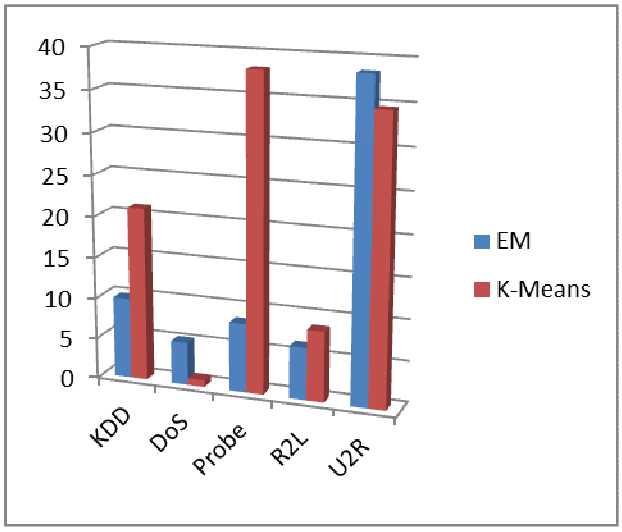

Figure 3: Inaccuracy of the Algorithms.

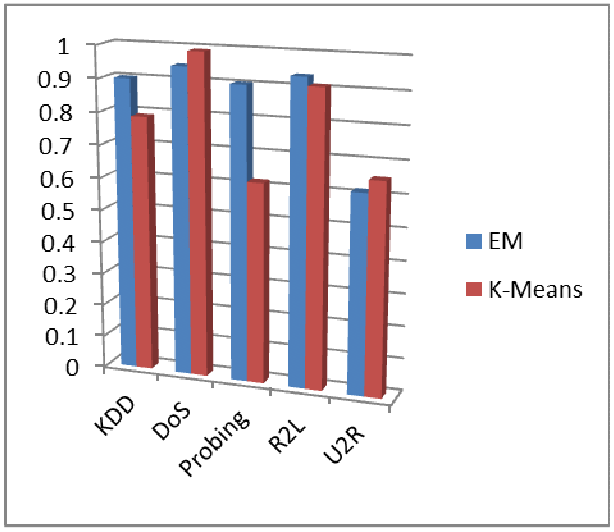

Figure 2: True Positive of the Algorithms.

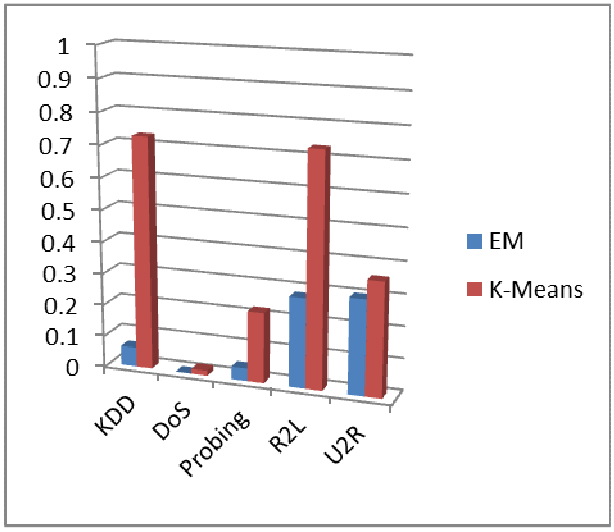

Figure 4: False Positive the Algorithms.

5. CONCLUSION AND RECOMMENDATION

From the analysis, the experimental result revealed that feature selection improved the performance of both K-means and EM algorithm on the datasets, though EM didn't perform well on the U2R and DOS datasets as compared to K-means. This experiment recorded that there was general improvement of both algorithms on the dataset. The performances of the algorithms were well above average in terms of accuracy (correctly classified instances). Evaluating the results of the models on KDD dataset revealed EM algorithm is better than its K-means counterpart though not in all respect. Nevertheless, the results of both models on various categories of attack are credible and differ slightly in some cases. However, we strongly recommended further research be carried out on how clustering technique can be used for classification, using this research as a benchmark. Further researches should see to how improvements can be made on classification via clustering techniques.

REFERENCES

[1] Ma, Y., Choi, D., and Ata, S. (2008.): APNOMS 2008, "Application of Data Mining to Network Intrusion Detection: Classifier Selection Model", LNCS 5297, pp. 399–408, 2008.

[2] Jaiganesh, V., Mangayarkarasi, S., and Sumathi, P., (2013) "Intrusion Detection Systems: A Survey and Analysis of Classification Techniques". IJARCCE, Vol. 2, Issue 4, April 2013.

[3] Osama, A., (2008). "Comparisons between Data Clustering Algorithms". The International Arab Journal of Information Technology, Vol. 5, No. 3.

[4] Yong, G.J., Min S.K. and Jun H. (2014). "Clustering Performance Comparison using K-means and Expectation Maximization Algorithms." Biotechnology & Biotechnological Equipment, 28:sup1, S44-S48.

[5] Ian H.W., Eibe F. and Mark A.H. (2011). "Data Mining: Practical Machine Learning Tools and Techniques (3rd edition)". Morgan Kaufmann Publishers Inc.,San Francisco, CA, USA.

[6] Neil, A., Andrew, S. and Doug, T (n.d). "Clustering with EM and K-Means".

[7] Balogun A. O., Mabayoje M. A., Salihu S and Arinze S. A. "Enhanced Classification via Clustering Techniques using Decision Tree for Feature Selection". International Journal of Applied Information Systems 9(6):11-16, September 2015. Published by Foundation of Computer Science (FCS), NY, USA.

[8] Mehmet, A., I. Cigdem and A. Mutlu. (2010). "A hybrid classification method of K Nearest Neighbor, Bayesian Methods and Genetic Algorithm," Expert Systems with Applications vol. 37, p. 5061–7.