The International Journal Of Engineering And Science (IJES)
|| Volume || 6 || Issue || 1 || Pages || PP 93-97 || 2017 ||
ISSN (e): 2319 – 1813 ISSN (p): 2319 – 1805

Hybrid Approach for Intrusion Detection Model Using Combination of K-Means Clustering Algorithm and Random Forest Classification

Muhammed Kabir Gambo[1], Azman Yasin[2]
[1,2]*School of Computing, Universiti Utara Malaysia,*

----------**ABSTRACT**----------
Any violation of information security policy with malicious intent is regarded as an intrusion. The fast evolving new kind of intrusions poses a very serious threat to system security, although there has been the rapid development of several security tools to counter the growing threats, intrusive activities are still growing. Many Intrusion Detection models have been implemented since the concept of Intrusion Detection emerged, but the majority of the existing Intrusion detection models have many drawbacks which include but not limited to low accuracy in detection, high false alarm rates, adaptability weakness, inability to detect new intrusions etc. The main aim of this study is proposing a model that combined simple K-Means clustering Algorithms and Random Forest classification technique that will have minimum false alarms rate and high accuracy detection rate. The experiment was carried out in WEKA 3.8 using the NSL-KDD dataset to process the dataset and obtained the results. At the end of training and testing of the proposed study, the results indicated that the proposed approach achieved improved accuracy and reduced false alarm rates by 99.98% and 0.14% respectively.

Keywords: *Random Forest, K-Means, Clustering, Classification, NSL-KDD data set, WEKA.*

Date of Submission: 11 January 2017 Date of Accepted: 16 January 2017

I. INTRODUCTION

The studies in intrusion detection systems' field have grown astronomically in the recent decades due to different stories and bad experiences on network attacks. And because the Internet is becoming the main medium for communication and the unprecedented surge of network technologies for getting important and demanding information always, attackers take advantage of this to inflict harm on the victims' systems and networks for their malicious intent.

Intrusion Detection system is an efficient defense technique against network attacks as well as host attacks. It monitors key nodes of computer systems or networks, collects, analyzes, audit records security logs and network packets. (Hu, Li, Xie, & Hu, 2015)**.**

(Elbasiony, Sallam, Eltobely, & Fahmy, 2013) Used weighted K-means and Random Forest classification, the experiment worked very well except that KDD CUP99 dataset was used and the results were 98.3% Detection Rate and 1.6% false alarm rate.

(Yassin, Udzir, & Muda, 2013) Proposed integrated machine algorithms and Naïve Bayes to minimize false alarm rate and improve accuracy rate. The results show significant improvement in the accuracy rate with 99.0% when compared with previous studies with the same approach. However, false alarm rate was high at 2.2%.

In (Tahir et al., 2015) K-means clustering algorithms was combined with support vector machine to formed hybrid intelligent system, the Authors were able to obtain 96.24% accuracy and 3.715% alarm rate.

II. INTRUSION DETECTION SYSTEM

Intrusion detection systems are often classified by the way they detect the attacks. But in general terms, there are two categories of IDS: Anomaly based and Signature based. The Signature based system perform similar fashion with most antivirus systems. They maintained a database of the signatures that might detect a particular type of attack and compare incoming traffic to those signatures if there is any similarity it triggers an alarm. The drawback of this type of detection methods is that it relied solely on the signature database to detect an attack (Onuwa, 2014).

Anomaly based detection typically works by taking the baseline of the normal traffic and activities taking place on the network. They now compare the current state of the traffic on the network against this baseline to detect patterns that are not normally present in the traffic. But it also has its drawback which is false alarms.(Bhuyan, Bhattacharyya, & Kalita, 2014).

III. DESIGN OF PROPOSED HYBRID TECHNIQUE

The proposed study applies clustering and classification techniques.The key concept of clustering is to group similar data in one cluster and the unrelated data in another cluster.(Hu et al., 2015). Distance is used to evaluate the similarity of two different samples, if the distance is between two sample is shorter, then the similarity is higher. The distance between samples points and the cluster center is regarded as the objective function.

$$J_c = \sum_{i=1}^{c} \sum_{p \in C_i} \|p - M_i\|^2 \quad (1)$$

From the above formula, M_i is the average number of cluster C_i, p is the data point inside the clusters in all the iterating process, afterward repeat the calculation of the cluster centers and that become the next iteration reference. In any of the two iterations, a comparison of the objective function is made, the smallest among them is the one closer to the best (Hu et al., 2015).

The ultimate goal of classification is building a system from classified objects in order to classify objects that were not previously seen as accurately as possible. And based on the available information of the classes and the type of classification, the classifier output can be presented in many forms. Example; Rules, Trees, etc. (Chauhan, Kumar, Pundir, & Pilli, 2013).

The study used K-Means clustering algorithm to separate and then label the data for the corresponding groups before applying Random Forest classifier for classification.

IIIA. K-MEANS CLUSTERING

It is an unsupervised machine learning algorithm popularly applied to solve most of the known clustering issues in machine learning and data mining and it is very easy to implement. K-mean clustering is the most common technique for analyzing raw data. It aids intrusion detection even if the training data is not labeled, it can also detect new and unknown intrusions.(Kumar, Chauhan, & Panwar, 2013).

IIIB. RANDOM FORESTS

Random Forest is a data analysis approach and predictive modeling, it is also an approach to data exploration, it generates many trees by using recursive partitioning then aggregate the results. Each of the trees is constructed separately by using a bootstrap sample of the data when the bagging technique is applied (Chauhan et al., 2013). It is also an amalgamation of tree predictors in a manner that each tree relies on the amount of random vector sampled independently with equal circulation for the whole trees in the forest.

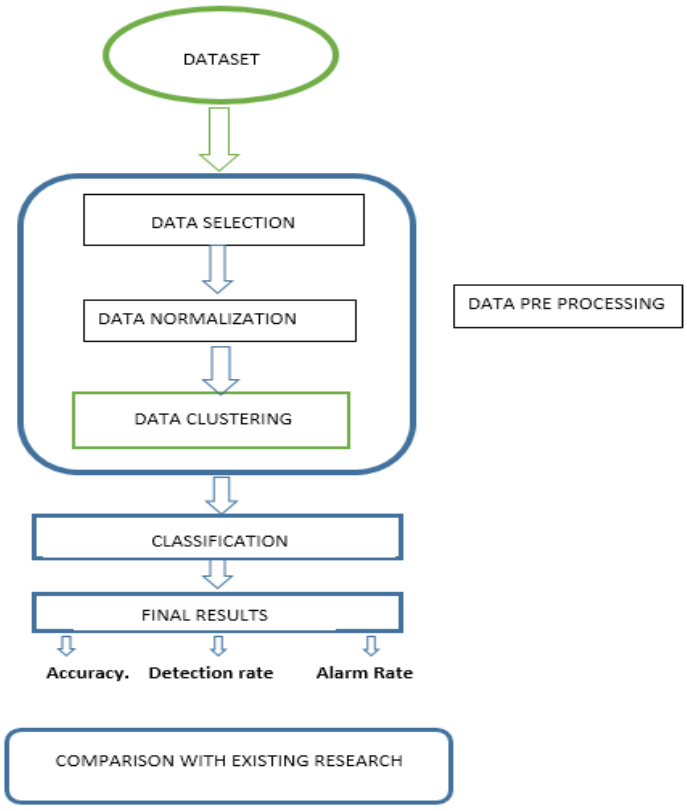

Figure 1. proposed hybrid intelligent approach

IV. EXPERIMENT SETUP

The experiment was carried out in WEKA 3.8 using NSL-KDD dataset. WEKA supports all kinds of tasks related to machine learning and data mining such as preprocessing, regression, clustering, classification, feature selection and visualization(Panwar, 2014).

IVA. Dataset descriptions

NSL-KDD data set was used for the experiment. It is the upgraded version of KDDCUP '99 Intrusion detection dataset (Tesfahun & Bhaskari, 2013). KDDCUP '99 have some inherent issues of large unnecessary records that make the learning algorithms favors the most recurrent records and restrain it from detecting the minority records (Tahir et al., 2015).

In NSL-KDD dataset each of the class is either normal or intrusion. The five main attack classes in the dataset are; i) Remote to Local (R2L) ii) User to Root (U2R) iii. Probe iv) Denial of Service (DOS) and v) normal. Each instance in a dataset is the network connection.

IVB. Data Pre-Processing

The aim of pre-processing is to make original NSL-KDD intrusion dataset applicable input for the classification. Preprocessing also reduces vagueness and produce accurate information to detection engine. In addition, preprocessing arranges the network data by grouping and handles the incomplete dataset.

IVC. Data normalization

Dataset normalization plays very important part in the preparation of data prior to classification. Normalizing the input data will assist in accelerating the learning phase and boost the performance of intrusion detection even if the datasets are too enormous. Without normalization, features with greater values dominate the features with smaller values (Moussaid & Toumanari, 2014)

V. RESULTS AND DISCUSSION

VA. Clustering results

The K-Means algorithms result obtained after the clustering was performed using Euclidean distance measure grouped the dataset into normal and anomaly. The number of Iteration was 10 on full dataset. All attributes were normalized in the range of 0 – 1, and the number of clusters was set up to two. The outcome of the results indicated 81% as an anomaly while 19% as normal behavior.

VB. Classification results

After applying all the preprocessing steps, the classification phase was performed using Random forest technique with test mode of **10-fold cross validation and full training set.** Random forest classification divides the network behavior to normal and abnormal and assigns the attack behavior to its specific category. The confusion matrix was realized from the classification of the proposed hybrid intelligent approach using full NSL-KDD intrusion dataset. From **125,973** connection instances. Table 1 shows the obtained confusion matrix by connection records in testing the proposed approach.

Table 3: results of Confusion Matrix for Classification (number of connection records)

Actual	**Predicated**	
	Attack	Normal
Attack	True Positive (TP)= 67,317	False Negative (FN) = 26
Normal	False Positive (FP) = 80	True negative (TN) = 58550

The obtained confusion matrix for classification of the proposed approach was calculated as shown in Table 2. Clearly, the result indicated a high rate of detection. **99.98** percent attack was detected from 67,343 real attacks, at the same time **0.02** percent regarded as normal. The total number of the normal connections of records in the NSL-KDD testing dataset which is **22,544** was classified as **99.86** percent as normal and **0.14** percent as an attack using 41 training features..

Table 4: the results in percentage confusion Matrix

Actual	**Predicated**	
	Attack	Normal
Attack	TP = 99.98%	FN = 0.02%
Normal	FP = 0.14%	TN= 99.86%

VC. Performance Evaluation

The performance evaluation of the proposed approach consists of two phases. First, a mathematical equation was applied and the second phase was carried out by comparing the result of the proposed approach and existing hybrid intelligent approaches. The results rely on the measurement metrics which was obtained from the classification of the proposed approach. It combined the K-means clustering and Random forest classification algorithms,

The accuracy (A), implies the total number of connections correctly classified including normal and intrusive connections. The detection rate (DR), is the number of attacks detected when it happened. Lastly, False alarm rate (FAR), is the number of attacks detected when there was none in the actual sense.

Table 5 presents the results of accuracy, detection rate, and false alarm rate as follows.

Table 5: Result of Performance Evaluation

Metric	**Formula**	**Value**
Accuracy	***(TP+TN)/(TP+ TN+FP+FN)***	**99.98%**
Detection rate	***(TP)*** / (TP+FP)	**99.86%**
False Alarm Rate	(***FP) / (FP+TN)***	**0.14%**

The second phase was the evaluation process to enable correlation with existing intelligent approaches for network intrusion detection to verify that, the proposed approach has improved the detection rate and decrease the false alarm rate. Base on the foregoing, the proposed approach was compared with five of some of the current hybrid intelligent approaches for network intrusion detection. The table below shows the comparison and differences between these approaches in detection rate.

Table 6: Existing Approaches and the Prop osed Hybrid Intelligent Approach comparison table

AUTHOR/YEAR	**TECHNIQUES**	**DATASET**	**ACCURACY RATE**	**ALARM RATE**
(Patra & Map, 2013)	SOM + PCA	NSL-KDD	93.01%	5.4%
(Abraham, 2010)	NB + PCA	NSL-KDD	94.84%	4.4%
(Tahir et al., 2015)	K-Means + SVM	NSL-KDD	96.24%	3.715%
(Govindarajan, 2014)	RBF + SVM	NSL-KDD	98.46%	-
(Elbasiony et al., 2013)	Weighted K-means + Random forest	KDD CUP 99	98.3%	1.6%
(Yassin et al., 2013)	K-means + NB	NSL-KDD	99.0%	2.2%
The Proposed hybrid Approach (2016)	**Simple K-Means + Random forest**	**NSL-KDD**	**99.98%**	**0.14%**

VI. CONCLUSION

The proposed study analyzed NSL-KDD CUP 99 dataset by applying K-Means clustering and Random Forest Classification techniques. K-Means enabled the clustering of attacks present in the training dataset into four major categories giving a better representation of the clusters. Confusion matrix was used to produce the results. Also, the process of performance evaluation was done using three measurement metrics. In the end, correlation of the results of the proposed approach was made with the existing network intrusion detection approaches, the results obtained indicated a significant improvement in the detection, accuracy rate of **99.86%, 99.98% respectively** and False alarm rate reduced to **0.14%** when compared with the existing hybrid models.

REFERENCES

[1]. Abraham, A. (2010). Discriminative Multinomial Naïve Bayes for Network Intrusion Detection, 5–10.

[2]. Bhuyan, M. H., Bhattacharyya, D. K., & Kalita, J. K. (2014). Network Anomaly Detection : Methods , Systems and Tools, 16(1), 303–336.

[3]. Chauhan, H., Kumar, V., Pundir, S., & Pilli, E. S. (2013). A Comparative Study of Classification Techniques for Intrusion Detection. http://doi.org/10.1109/ISCBI.2013.16

[4]. Elbasiony, R. M., Sallam, E. A., Eltobely, T. E., & Fahmy, M. M. (2013). ELECTRICAL ENGINEERING A hybrid network intrusion detection framework based on random forests and weighted k-means, 753–762.

[5]. Govindarajan, M. (2014). A Hybrid RBF-SVM Ensemble Approach for Data Mining Applications, (February), 84–95. http://doi.org/10.5815/ijisa.2014.03.09

[6]. Hu, L., Li, T., Xie, N., & Hu, J. (2015). False Positive Elimination in Intrusion Detection Based on Clustering, 519–523.

[7]. Kumar, V., Chauhan, H., & Panwar, D. (2013). K-Means Clustering Approach to Analyze NSL-KDD Intrusion Detection Dataset, (4), 1–4.

[8]. Moussaid, N. E. L., & Toumanari, A. (2014). [Overview of Intrusion Detection Using Data-Mining and the features selection.

[9]. Onuwa, O. B. (2014). ORIENTAL JOURNAL OF Improving Network Attack Alarm System : A Proposed Hybrid Intrusion Detection System Model.

[10]. Panwar, S. S. (2014). OF COMPUTER © I A E M E DATA REDUCTION TECHNIQUES TO ANALYZE NSL-KDD DATASET, 21–31.

[11]. Patra, M. R., & Map, A. S. O. (2013). Enhancing Performance of Intrusion Detection through Soft Computing Techniques. http://doi.org/10.1109/ISCBI.2013.17

[12]. Tahir, H. M., Hasan, W., Said, A., Zakar-, N. H., Katuk, N., Kabir, N. F., … Yahya, N. I. (2015). HYBRID MACHINE LEARNING TECHNIQUE FOR INTRUSION DETECTION SYSTEM, (209), 464–472.
[13]. Tesfahun, A., & Bhaskari, D. L. (2013). Intrusion Detection using Random Forests Classifier with SMOTE and Feature Reduction, 128–133. http://doi.org/10.1109/CUBE.2013.31
[14]. Yassin, W., Udzir, N. I., & Muda, Z. (2013). ANOMALY-BASED INTRUSION DETECTION THROUGH K- MEANS CLUSTERING AND NAIVES BAYES CLASSIFICATION, (49), 298–303.