International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-3S, February 2019

Network Intrusion Detection System using K-Means Clustering and Gradient Boosted Tree Classifier

Nandini Rebello, Manamohan K

Abstract— Network intrusion detection is an important and dynamic research area because the internet is always subjected to an ever increasing number of security threats. As the type of attacks appearing is continuously changing, there is a need for developing adaptive and flexible security features. This is where anomaly-based network intrusion detection techniques are important to protect the network against malicious activities. In literature, many such intrusion detection systems have been proposed till date. In this paper, a hybrid model for intrusion detection by performing K-means clustering to form cluster models of the dataset and input it to the Gradient Boosted Tree classifier has been proposed. In order to evaluate the performance metrics the NSL-KDD dataset was used. The proposed model showed improved results having high detection rate of 99.3% and low false alarm rate of 0.19%.

Index Terms— Anomaly detection, k-means clustering, gradient boosted tree classification, Intrusion Detection

I. INTRODUCTION

With growing changes in the internet today the possibilities and opportunities available are endless. But with this unlimited opportunities the risk of attacks and intrusions have also increased. Though firewall, antivirus and other access control mechanisms provide some level of security, additional security tool is required to strengthen the security of information system and communication. Intrusion Detection Systems aims to provide such a security mechanism against malicious intrusions.

An intrusion attempt is an unauthorized and deliberate attempt to access valuable information and temper it. Intruders who exploit such valuable information are very proficient in using the latest and sophisticated programming techniques available. Attacks can be classified into following types (a) Denial of Service (DoS): this form of attack prevents the host from using the resources that are required to operate correctly; (b) Worms and viruses: this form of attack makes use of the network in order to exploit other hosts; (c) Compromises: using known vulnerabilities this type of attack obtains privileged access to a host [1].

Intrusion Detection Systems (IDS) is normally categorized into two types based on the information source i.e. host-based or network based. Host based systems analyses events mainly related to the operating system such as process identifiers and system calls. Network based systems monitors' network traffic and identifies suspicious patterns in the packets. Intrusion detection systems assumes that malicious activities are different from the normal activities and hence can be detected. Previously proposed network based system were mainly rule based. These type of systems had drawbacks, such as it cannot be extended to detect new features, all intrusions cannot be specified using rules, and also the procedure to encode rules is slow and expensive [2].

Therefore, IDS can be categorized into signature based intrusion detection which makes use of a set of rules or known signature attacks and hence, detects only known attacks by matching the incoming data with the data stored in the signature database. Signature based systems thus cannot detect new or unknown attacks. And another way of detection is, anomaly based systems that searches for patterns whose behavior is not as expected in a dataset. Anomalies can be any behavior which is previously not known and hence cannot always be called as an attack. Anomaly based system can therefore detect previously unknown attacks but have a disadvantage of a high false positive rate.

In order to analyze large datasets and predicting known and unknown behavior, data mining techniques can be used. This includes machine based learning techniques, classification techniques and clustering techniques. One of the application area of data mining is intrusion detection. Clustering technique is mainly used to group data having similar properties and to detect outliers. Similarly, classification is a technique that is used to predict the class label of a data object based on previously encountered data objects. In this paper, a hybrid network intrusion detection model has been proposed using both of these data mining techniques. To form clusters of the data set k-means clustering algorithm was used, and then different Gradient Boosted tree classifiers for each of the clusters were trained.

The remaining sections of the paper has been organized as follows; literature review that discusses the various hybrid models that have been proposed in the research field, next is methodology in which the proposed model has been described, followed by results and discussions and then conclusion.

II. LITERATURE SURVEY

Intrusion compromises on the security of communication networks in terms of availability, integrity and confidentiality. An intruder can be an insider or an outsider trying to gain unauthorized control and entry of a system. To

Manuscript published on 28 February 2019.

* Correspondence Author (s)
Nandini Rebello, Dept. of Computer Science & Engineering, Manipal Institute of Technology, MAHE, India. (E-mail: nandinirebello@gmail.com)
Manamohan K, Dept. of Computer Science & Engineering, Manipal Institute of Technology, MAHE, India. (E-mail: sri.manmohan@manipal.edu)

protect network systems, intrusion detection systems provides mechanisms that gathers and analyses information from a host and identifies possible security breaches. Anomaly based intrusion detection systems assume that all intrusive activities are anomalous, which may result in activities that are anomalous but not intrusive, hence being falsely signaled as intrusive. This is known as false positive.

In the research field, numerous hybrid intrusion detection models have been developed. In this literature survey a few of the recent works carried out in this area have been compared and studied. S. T. Ikram and A. K. Cherukuri proposed a model where in the chi-square feature selection method was used to reduce the number of attributes of the dataset along with multi class Support Vector Machine (SVM) classifier. By optimizing the parameter it ensures that the predicted label for the test set is accurate. This model showed an improved true positive rate and reduced false alarm rate [3].

S.Varuna and P.Natesan, proposed a hybrid intrusion detection method using k- means clustering algorithm to find the distance between each object in the dataset and the number of centroids, using this new features were formed. This new features were then input to the classifier for training and detection [4]. Naïve Bayes classifier is a probabilistic model and it does not consider a feature based on the presence of any other feature, because of which this model provides a better detection rate for attacks like U2R and R2L attack type.

N. A. Biswas et al., proposed a model for intrusion detection using Principal Component Analysis technique for selecting relevant features, K-means clustering to group attacks of a specific type and Artificial Neural Network technique is used on each of the clusters trained, with the testing dataset [5]. The final result using different neural network is formed by aggregating the neutral network. Compared to high frequent data, result showed better values for less frequent data like U2R and R2L. K. K. Vasan and B. Surendiran worked on finding out the minimum number of Principal Components required to reduce the dimensionality of network data [6]. Based on the results obtained, the first 10 Principal Components are effective for classification algorithms, which reduced the number of attributes from 41 to 28.

W. L. Al-Yaseen et al., proposed a hybrid multi-level intrusion detection model using Extreme Learning Machine (ELM) and Support Vector Machine. This model proved to be efficient in detecting unknown and known attacks. It modifies the k-means algorithm to generate a reduced and high quality dataset. The main objective was to separate the original training data into five categories and the new training data comprising of fewer instances which in turn reduced the classifiers training time. To train the model support vector machine and extreme learning machine is used on this new training data set. This hybrid ELM and SVM model proved to be more reliable and gave better detection performance that did not cause large fluctuations [7].

S. Aljawarneha, M. Aldwairia and M. B. Yassein proposed an ensemble hybrid model consists of following classifiers: J48, Meta Pagging, Random Tree, REPTree, AdaBoostM1, DecisionStump and Naïve Bayes. The model achieved a higher percentage in successfully correctly classifying the instances when compared to the individual classifiers and also obtained lower false positive rate and highest percentage for true positive rate [8]. Although an improvement in the accuracy rate for probe and R2L is required.

III. METHODOLOGY

The proposed method uses K-means clustering and Gradient Boosted Tree classifier to model a Network Intrusion Detection System. The clustering algorithm used to form clusters of data that reduces the training data set to a new data set is K-means. This training dataset is used to train the Gradient Boosted Tree classifier and the model is tested using the test dataset.

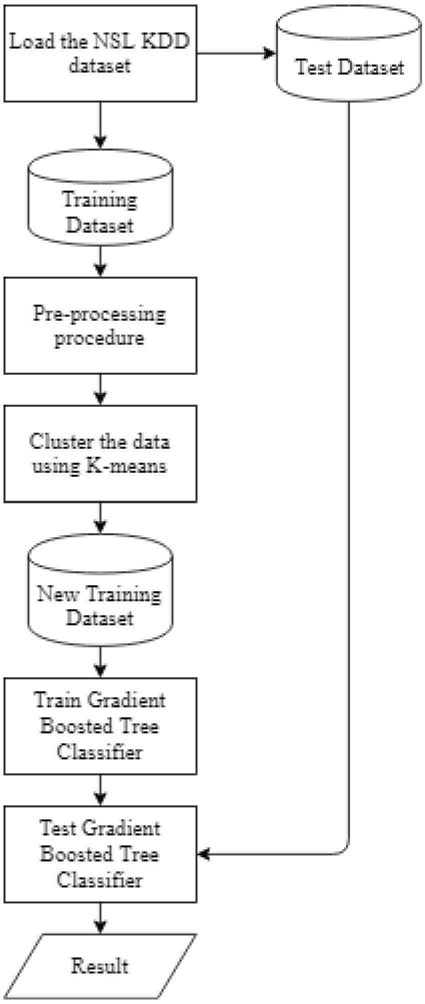
Load the NSL KDD dataset
Test Dataset
Training Dataset
Pre-processing procedure
Cluster the data using K-means
New Training Dataset
Train Gradient Boosted Tree Classifier
Test Gradient Boosted Tree Classifier
Result

Fig. 1 Proposed Approach

A) NSL – KDD Data Set

In order to perform analysis, the NSL-KDD dataset is used, which is a subset of the original KDD dataset. Each entry is described by 41 features and is a network connection, out of which 34 are continuous, 7 are nominal and label attribute to indicate whether the connection is of the type normal or any of the different attack types [9].

B) Pre – processing

In order to perform feature selection, the technique used is Information Gain. Suppose the training set consists of S samples with its corresponding m class labels and the set contains si samples of class I [10]. In order to classify a given sample, the needed Information is calculated using the formula:

$$I(s1, s2, \ldots, sm) = -\sum_{i=1}^{m} \frac{si}{s} \log_2\left(\frac{si}{s}\right) \qquad (1)$$

Feature set *F* with values { f_1, f_2, …, f_v } can divide the training set into v subsets { S_1, S_2, …, S_v } where *Sj* is the subset which has the value *fj* for feature *F*. Furthermore let *Sj* contain *sij* samples of class *i*. Entropy of the feature *F* is given by:

$$E(F) = \sum_{j=1}^{v} \frac{s1j + .. + smj}{s} * I(s1j, \ldots, smj) \qquad (2)$$

Information Gain for F can be calculated as:

Gain (F) = I (*s1*,...,*sm*) − E(F) (3)

C) K- Means Clustering

K-means clustering is used to partition the dataset into clusters, such the data objects within a cluster are similar and the data objects belonging to two different clusters are not similar. In order to calculate the similarity measure, the Euclidean distance formula is used. Initially it selects k objects randomly as the cluster centres, where k is the number of clusters and reassigns the each object to a cluster it is most similar, based on the mean value of the objects. It repeats this process until there is no change.

In this paper, K-means clustering is used to form cluster models which are then used by the Gradient Boosted tree classifier. As K-means cannot handle categorical data, only the numeric features are used in order to build the cluster models.

D) Gradient Boosted Tree Classifier (GBT)

Classification is a procedure used for predicting the class label of new instances on the basis of a training set containing observations whose class label is known. In order to train the cluster models Gradient Boosted Tree Classifier is used, this classifier is built for improving the performance of classification and regression trees (CART). This classifier grows trees in a sequential manner and based on the results of the trees obtained in the previous step, new decision trees are obtained. In order to predict the final output, the tree ensemble makes use of a K additive function for a given data set with x features and s data samples. That is, the sum of predictions from each tree determines the final prediction of a given sample S [11].

E) Performance Metrics

The following parameters have been considered for evaluating the performance of the proposed idea.

Accuracy:

A = (TP+TN) / (TP+TN+FP+FN) (4)

Detection Rate:

DR = (TP) / (TP+FP) (5)

False Alarm Rate:

FAR = (FP) / (FP+TN) (6)

The performance metrics are determined by using a confusion matrix, as depicted in table I.

Table I Confusion Matrix

	Normal	**Attack**
Normal	True Positive	False Positive
Attack	False Negative	True Negative

TP=True Positive, indicates actual positives that have been identified correctly,

TN= True Negative, indicates an attack is not detected when there is no attack,

FP= False Positive, indicates incorrect detection of normal as an attack,

FN= False Negative, indicates that the system has failed to detect an attack.

An excellent style manual for science writers is [7].

F) Proposed Algorithm

The algorithm described below gives the step-by-step procedure used in order to carry out the experiment and record the performance metrics of the proposed model.

Input: NSL – KDD Data set

Output: Classification of data set based on normal and attack type

Step 1: Load NSL KDD data set.

Step 2: Divide the data set into two parts; 80% training and 20% test dataset.

Step 3: Apply preprocessing technique – eliminate duplicate and missing values. And apply information gain for feature selection.

Step 4: Use k-means to cluster the dataset into two types normal and attack.

Step 5: New training dataset formed in step 4 is used to train the gradient boosted tree classifier.

Step 6: Use the test data set on the trained classifier in step 5 to determine the accuracy of the model.

Step 7: Record the performance metrics, that is, the false alarm rate (FAR), Detection rate (DR) and accuracy.

IV. RESULTS AND DISCUSSIONS

In order to conduct the experiment, the NSL – KDD data set was used. Out of the 42 attributes, the last attribute contained the class label indicating whether the data object is of the type normal or attack. 80% of the data set was used to train the model, while 20% was used to test the model.

Before applying the clustering and classification techniques on the dataset, the dataset was pre – processed where by, data objects containing null values as well as those containing duplicate values were eliminated. Once the data was cleaned, feature selection was done using the Information Gain technique. Table II shows the comparison of the results obtained in terms of accuracy, detection rate and false alarm rate for the various techniques used.

In this paper, only two class labels have been used in order to compute the performance metrics, that is, normal and attack. In order to perform a comparison analysis of the results obtained, the proposed approach is compared to the results obtained in [12], which uses only k-means clustering for intrusion detection.

Table II Comparison of the results obtained

Method	Accuracy (A)	DR	FAR
K-means clustering in [12]	81.61%	40.8%	4.03%
K-means clustering with feature selection	81.06%	99.7 %	1%
K-means clustering and GBT classifier	99.6%	99.3 %	0.19%

Compared to the results obtained in [12], it can be observed that the accuracy is only slightly improved with K-means clustering with feature selection, but the detection rate increased and false alarm rate reduced by a considerable amount.

Also, accuracy of the system increases and false alarm rate decreases when performing clustering followed by classification as proposed in this paper. The results obtained shows that the proposed methodology can achieve low false alarm rate, high detection rate and good accuracy.

However, in order to accurately determine the performance it would be necessary to consider the different type of attacks, that is, DoS, Probe, R2L, and U2R.

V. CONCLUSION

In this paper a novel approach for network intrusion detection is proposed using k-means clustering and gradient boosted tree classifier. The results obtained proved that using a classifier algorithm after forming clusters of data gives an accurate and high detection rate model with low false alarm rate. However, the experiment conducted was based only on normal and attack type class labels. For future work, analysis can be carried out considering the different types of attack, that is, DoS, probe, U2R, and R2L

REFERENCES

1. M. H. Bhuyan, D. K. Bhattacharyya, and J. K. Kalita, “Network Anomaly Detection: Methods, Systems and Tools”, IEEE Communications Surveys & Tutorials, VOL. 16, NO. 1, First Quarter, 2014.
2. M.A. Jabbar, S. Samreen, “Intelligebt Network Intrusion Detection Using Alternating Decision Trees”, International Conference on Circuits, Controls, Communications and Computing (I4C), IEEE, 2016.
3. S. T. Ikram , A. K. Cherukuri, “Intrusion detection model using fusion of chi-square feature selection and multi class SVM”, Journal of King Saud University – Computer and Information Sciences (2017) 29, 462–472, 2017
4. S.Varuna, Dr.P.Natesan, “An Integration of K-Means Clustering and Naïve Bayes Classifier for Intrusion Detection”, 3rd International Conference on Signal Processing, Communication and Networking (ICSCN), 2015.
5. N. A. Biswas, W. M. Tammi, F. M. Shah, S. Chakraborty, “FP-ANK: An Improvised Intrusion Detection System with Hybridization of Neural Network and K-Means Clustering over Feature Selection by PCA”, 18th International Conference on Computer and Information Technology (ICCIT), 2015.
6. K. K. Vasan, B. Surendiran, “Dimensionality reduction using PrincipalComponent Analysis for network intrusiondetection”, Perspectives in Science (2016) 8, 510—512. Published by Elsevier GmbH, 2016.
7. W. L. Al-Yaseen, Z. A. Othman, M. Z. A. Nazri, “Multi-level hybrid support vector machine and extreme learning machine based on modified K-means for intrusion detection system”, Expert Systems With Applications 67 (2017) 296–303, , Published by Elsevier, 2017.
8. S. Aljawarneha, M. Aldwairia, M. B. Yassein, “Anomaly-based intrusion detection system through feature selectionanalysis and building hybrid efficient model”, Journal of Computational Science, Published by Elsevier, 2017.
9. Olusola, Adetunmbi A., Adeola S. Oladele, and Daramola O. Abosede. "Analysis of KDD’99 Intrusion detection dataset for selection of relevance features." In Proceedings of the World Congress on Engineering and Computer Science, vol. 1, pp. 20-22. 2010.
10. H. G. Kayacık, A. N. Zincir-Heywood, M. I. Heywood, “Selecting Features for Intrusion Detection: A Feature Relevance Analysis on KDD 99 Intrusion Detection Datasets”. in Proc. 3rd Annual Conference on Privacy, Security and Trust, 2005.
11. B. A. Tama1, K. H. Rhee1, “An in-depth experimental study of anomaly detection using gradient boosted machine”. Neural Computing and Applications, pp 1–11.
12. S. Duque, Dr.M. N. bin Omar, “Using Data Mining Algorithms for Developing a Model for Intrusion Detection System (IDS)”, Procedia Computer Science 61 (2015) 46 – 51, Published by Elsevier, 2015.
13. “Nsl-kdd data set for network-based intrusion detection systems”. Available on: http://nsl.cs.unb.ca/NSL-KDD.