Imperial Journal of Interdisciplinary Research (IJIR)
Vol-3, Issue-11, 2017
ISSN: 2454-1362, http://www.onlinejournal.in

Intrusion Detection Using Hierarchical Clustering Followed By Signature Approach & Random Forest Classifier

Ms. Meghana Solanki[1] & Mr. Pranav Pathak[2]
[1] Computer department DYPCOE, Ambi, Talegaon Pune, India
[2] Department of Computer Science & Engg Arvind Gavali College of Engg., Satara, India

Abstract: *As the need of technology is expanded, intrusion revelation has become an emerging campus for analysis. Intrusion Detection System (IDS) tries to recognize as well as announce the action of applicants as either normal or anomaly. An IDS is not only a nonlinear but also complicated dilemma. It pledges with network traffic data. Many IDS approaches have been introduced. It produces different layers of accuracy. Due to which evolution of adequate as well as vigorous Intrusion detection system is required. The consequence display that planned miniature is potent with immense detection rate. In this paper, we have constructed a model for intrusion detection system using hierarchical clustering followed by signature based approach and random forest algorithm.*

1. Introduction

Intrusion activities are gaining day by day due to the gain in usage of network. There is raising in attack on computer system in recent years. There is need of perspective and robust intrusion detection system. The main aim of IDS is to search out invasion in normal information analysis. James Anderson first time invented IDS. There are two types of Invasion detection system: Host based as well as network based. The data held on individual computer systems are inspected by Host based intrusion detection systems. The data circulated among computer are evaluated by network based systems. An intrusion is also known as a malicious activity in the internet. Any activity that defies security protocol of the network is defined as invasion [1]. Intrusion detection system (IDS) is software as well as hardware. It is used to carry out the act of searching unauthorized use of a computer. It also searches telecommunications network which fill the gaps not only in the firewall but also anti-viruses. An IDS affords monitoring of applicants. It also and analyze user as well as system activity.IDS can audit not only configuration of system but also vulnerabilities. It assess the of critical system integrity as well as data files. It afford statistical analysis of activity patterns based on the matching with known attacks as well as analyze abnormal activity along with operate system audit [2]. The types of computer attacks detected by IDS are classified into 3 categories namely: (a) scanning attacks, (b) denial of service (DOS) attacks, and (c) penetration attacks [3]. In most IDS however, there is a high occurrences of false positives as well as false negatives which can be embarrassing to cope with the network administrators. In case of false positive, an IDS incorrectly recognizes a benign activity as a malicious In case of false negative, an IDS is unable to find out a malicious activity.

2. Literature Survey

In this paper [4], an author described, Data mining as a searching tool for knowledge in data. It also searches different types of patterns in data. In this paper [5], an author proposed a new version of the KDD data set which is known as the NSL-KDD. In this paper [6], an author proposed That NSL-KDD is a diminished version of the KDD'99 dataset. The NSL-KDD posses the same characteristics as the KDD'99 except one are it does not consist of the redundant data items of the KDD'99. It does not contain twin records which make it not only unbiased to frequent but also redundant entries.In this paper [7], an author proposed network IDS using Random forest as well as PSO. Appropriate features for classifying intrusions are selected by binary PSO [7]. In this paper [8], an author proposed Multistage filtering for network IDS .Authors used enhanced adaboost with decision tree algorithm along with Naive bayes. It detects frequent attacks in networks. In this paper [9], an author proposed A Hybrid Intelligent Approach for IDS. The performance of resultant model is improved by the combination of classifiers.In this paper [10], an author proposed IDS using Random forest and SVM. In this paper [11], an author proposed feature selection based Hybrid IDS using K-means as well as Radio basis function. He introduced hybrid technique which combines K-means and SVM.

3. Proposed Work

The architecture of Proposed IDS is shown in fig 1.It consists of 5 steps: loading of dataset, Pre-processing, Hierarchical classification, feature selection & Random forest algorithm.

3.1 NSL-KDD Data Set

The NSL-KDD data set has 30,000 entries and 43 attributes among them the 41 attributes are similar to the KDD'99. The data label is the 42nd attribute. The level of difficulty is the 43rd attribute. There are 22 different verities of data.

3.2 Pre-processing

Pre-processing cleans data inconsistent data as well as noise. It not only combines but also removes redundant entries. Pre-processing al convert the dataset attributes into numeric data and after that save into a readable format.

3.3 Hierarchical Method

Hierarchical decomposition of the given set of data objects is constructed by the hierarchical method. The classification of the hierarchical method depends on the way in which hierarchical decomposition is done. There are two categories of hierarchical methods: Agglomerative Approach and Divisive Approach.

Agglomerative approach is nothing but the bottom-up approach. In this method, we consider each object forming a separate group. It merges the objects or groups which are close to one another. This process continues until all of the groups are united into one or until there is condition terminate.

Divisive approach is nothing but the top-down approach. In this method, we consider all of the objects which are located in the same cluster. In case of the continuous iteration, a cluster is divided up into smaller chunks. This process continues until l each object in one cluster or there is condition terminate. This method is intransigent means once a union or distribution is done, it can never be unaccomplished.

3.4 Feature selection

Feature selection (FSS) is a pre developing step generally applied in data mining. It is not only impressive in dimensionality contraction but also evacuates irrelevant features which result in increase in accuracy. It is related to the problem of recognizing those features which are fruitful in anticipating class. There are 3 categories of feature selection method: 1) filter method, 2) wrapper method and 3) embedded method.

3.5 Random Forest (RF)

Random forest (RF) is an all at once classifier. It is used to boost the accuracy. Random forest contains variety of decisions trees. Random forest possesses low classification error in comparison with other conventional classification algorithms. Splitting each node is depending on number of trees, minimum node size along with number of features used for classification.

There are some advantages of RF are listed below.

1) Forests generated by RF can be saved for forthcoming quotation.

2) The problem of over fitting is beaten by Random forest.

3) Accuracy as well as variable importance is unquestionably set up by RF.

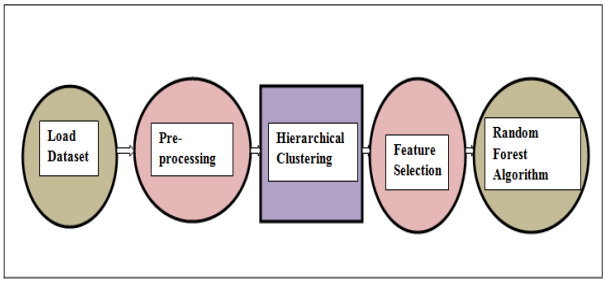

Figure 1. Architecture of Proposed IDS

4. Algorithm

We use Random forest Algorithm for proposed Intrusion Detection System

Input: NSL-KDD dataset

Output: Classification of different type of attacks

Step 1: Load the dataset
Step 2: Application pre-processing approach.
Step 3: Chunk the dataset using hierarchical clustering.
Step 4: Segregate the dataset into training and test.
Step 5: Select the best set traits.
Step 6: Dataset is handover to random forest for classification
Step 7: Calculate accuracy, Detection rate.

5. Results

We used NSL-KDD dataset for our IDS study. NSL-KDD dataset consists of 43 attributes. The class label is last attribute. We did assessment on various number of Random forest trees.

DR_N= no. of true normal data detected/ no. of normal data detected*100

DR_A= no. of true attack data detected/ no. of attack data detected*100

ER= DR_N + DR_A

Accuracy= Samples correctly classified in test data / Number of samples in test data

Where

DR_N= Detection rate for normal data
DR_A= Detection rate for attack data.
ER=Efficiency Rate

The consequences displayed an efficiency of 86.51%, 73.20%, 65.50% and 57.23% depending on the number of chunks used (25, 50, 75, or 100). Further, it can be observed that as the number of chunks increases over the data type numbers, the detection rate as well as efficiency rate, decreases.

Table 1. Hierarchical clustering results

	Cluster or chunk number			
	25	50	75	100
Genuine Normal Data Detected	13902	12354	11261	9196
Total Normal Data Detected	27512	23520	22507	21373
Genuine Attacks Detected	312	431	586	701
Total Attacks Detected	761	2472	3319	4797
Normal Data Detection Rate	57.63%	50.09%	49.69%	45.30%
Attack Data Detection Rate	31.53%	17.21%	13.02%	11.53%
Efficiency Rate	86.51%	73.20%	65.50%	57.23%

It is obvious from Tables 2 and 3 that our proposed miniature earned high Detection Rate to segregate the intervention. For DOS invasion our proposed miniature gained an accuracy of 99.83%, which is 6% greater than J48 algorithm.

Table 2. Performance Measure for Random Forest (No. of trees = 100).

Attack Type	**Accuracy**	**Detection Rate**
DoS	99.83	99.94
Probe	99.83	99.92
R2L	99.83	99.92
U2R	99.83	99.94

Table 3. Performance Measure for J48 Tree.

Attack Type	**Accuracy**	**Detection Rate**
DoS	99.35	99.6
Probe	99.39	99.5
R2L	99.34	99.6
U2R	99.38	99.6

6. Conclusion

The Results of hierarchical clustering displayed that a higher efficiency rate is gained. The efficiency rate depends on the correct number of clusters. In this paper we consider the Random Forest (RF) algorithm for searching four types of invasion such as DOS, probe, U2R and R2L. Feature selection is implemented on the NSL KDD dataset not only to shrink dimensionality but also to diminish superfluous and trivial factors. The proposed miniature is assessed with the help of NSL KDD data set. We did comparison of our newly proposed approach with j48 classier in case of accuracy as well as Detection Rate. Our empirical result proved that accuracy as well as Detection Rate for four types of attacks is increased.

7. References

[1] Bischof, H., Leonardis, A., and Selb, A. "MDL principle for robust vector quantisation. Pattern Analysis and applications". 2:59-72,1999.

[2] SANS Institute. "Understanding Intrusion Detection System". 2001.

[3] Bace, R., and Mell, P. "Intrusion Detection System", NIST Special Publications SP800. November. 2001.

[4] Han, K., Kamber, M., Pei, J. "Data Mining Concepts and Techniques". Third Edition. Morgan Kaufmann, Elsevier Inc. 2012. ISBN 978-0-12-381479-1.

[5] The DARPA Intrusion Detection Data Sets. Lincoln Laboratory Massachusetts Institute of Technology. Available at: www.II.mit.edu

[6] Patel A.,Sammarvar,S., and Naik, A. "DataMining Vs.Statistical Techniques for Classification of NSL-KDD Intrusion Data". International Journal of Computer Science and Information Technologies, Vol 5(4), 2014.ISSN:075-9646.

[7] Arif Jamal Malik, Waseem Shahzad and Farrukh Aslam Khan, "Network Intrusion Detection Using Hybrid Binary PSO and Random Forests Algorithm", *Security and Communication Networks*, (2012).

[8] P. Natesan and P. Balasubramanie," Multi Stage Filter Using Enhanced Adaboost for Network IDS", *International Journal of Network Security and its Applications*, vol. 4, no. 3, (2012).

[9] Mrutyunjaya Panda, Ajith Abraham and Manas Ranjan Patra," A Hybrid Intelligent Approach for Network Intrusion Detection", *UCCTSD*,pp. 1–9, (2012).

[10] Md. Al Mehedi Hasan, Mohammed Nasser, Biprodip and Shamim Ahmad," Support Vector Machine and Random Forest Modeling for IDS",*JILSA*, pp. 45–52, (2014).

[11] Ujwala Ravale, Nilesh Marathe and Puja Padiya, "Feature Selection Based Hybrid Anomaly Intrusion Detection System Using K Means and RBF Kernel Function", *ICACTA*, pp. 428–435, (2015).