JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

A Novel Network Intrusion Detection System Using Decision Tree Technique

Adithya Sharma.B.S
M-Tech Scholar,Department of Computer Science and Engineering,BMS Institute of Technology And Management , Bengaluru,Karnataka,India.

Anjan K Koundinya,
Department of Computer Science and Engineering,,
BMS Institute of Technology And Management ,Bengaluru,Karnataka,India..

Ashwini.N
Department of Computer Science and Engineering,
BMS Institute of Technology And Management ,Bengaluru,Karnataka,India.

ABSTRACT:

A Network Intrusion Detection System is a protection system which monitors computer network activity and alerts the network administrator if any malicious activity. Intruders make many efforts to obtain entry to the network and damage the data of the organization. As a result, the most critical feature of any sort of organization is security.

Intrusion detection has become a major research topic as a result of these factors. The two forms of IDS are signature-based and anomaly- based. The decision tree method in our proposed work is based on the signature based IDS. The importance of feature selection and split value in the construction of a decision tree cannot be overstated.

The algorithm in this research is intended to make the model faster in detection and accurate. The information gained is used to choose the most relevant features and the split value is selected to ensure that the classifier is unaffected by the most common values. On the basis of a variety of attributes, experiments are conducted on the NSL-KDD dataset. The time taken by the categorizer to build the model and the precision it achieves are examined. The suggested Decision Tree Split method can be employed for signature-based intrusion detection,according to the findings.

1. INTRODUCTION

An Intrusion Detection System (IDS) is a monitoring system which monitor's a network for malicious activity and sends a notification to the administrator of a network to events which don't meet security criteria. There are two types of network intrusion detection systems: anomaly-based and signature-based.

A signature-based intrusion detection system (IDS) uses a multitude of methods to detect similarity in the systems behavior and already identified threats have been saved in the signature database. The anomaly-based intrusion detection system recognizes network actions that are not consistent with the regular behaviors documented in the system profile database. There are several classifiers that can be used to detect tampering.Some are based on trees, like decision trees [1] and random forests [2], while others are governed by rules., like oneR [3], others, on the other hand, are function-based. like SVM(Support Vector Machine) [4].

A Decision Tree is a tree-like structure made up of internal nodes that reflect a test of an attribute, branches that indicate the test's result, as well as the leaf nodes that supply the class label.The categorization the route from the root node to the leaf determines the regulations. Because it is the most evident attribute for dividing the data, the root node is chosen first to separate each piece of input data.

The tree is made up of defining qualities and their matching values that will be applied to examine the data supplied at each intermediary node. Following the creation of the tree, it may figure newly entering data by travelling from a a leaf node to a root node, pausing along the way, at all internal nodes, and verifying the properties at each node [5].

Determining the value to use for node splitting in the tree is the most challenging component of designing adecision tree.

Decision trees can analyse data and detect crucial network traits that indicate malicious activity. Many real- time security systems can benefit from analysing a huge quantity of data on intrusion detection. It can detect trends and patterns that might aid in further study, the creation of threat signatures, as well as other monitoring responsibilities.

Decision trees have an advantage over other categorization systems in that they give a comprehensive set of rules that are easy to comprehendand combine with real-time technologies[6]. This dataset has 41 characteristics. NSL-KDD is the most recent open source dataset for network intrusion detection.The classifier will take longer to identify intrusion and its performance will suffer if a complete feature set is employed for categorization input data.

As a result, before conducting any categorization, we must decrease this collection using a feature selection strategy. Feature selection is used to eliminate features that are irrelevant or redundant. There are several feature selection methods in the literature, including information gain [7], PCA (Principle Component Analysis), and GA (Genetic Algorithm). There are several classifiers available for network data classification, including KNN (k- nearest neighbor), SVM, ANN (Artificial Neural Network), and decision tree. C4.5 constructs using the idea of information entropy to build a decision tree from a batch of training data.

The algorithm selects an attribute at each node of the tree, that most efficiently divides the every class in the provided training set into smaller subsets. The gain ratio is the dividing factor in this case. To make the decision, the gain ratio of the characteristic with the highest gain ratio is selected. [8].

2. RELATED WORK

On the NSL-KDD data set, Gaikwad and Thool [9] used the Genetic Algorithm (GA) to pick significant characteristics. The GA chooses 15 characteristics from a total of 41 in the data set. With decision tree as a classifier, these 15 characteristics provide 79 percent accuracy on test data, and the model takes 176 seconds to develop.

Information gain, gain ratio, and correlation-based feature selection are among the strategies discussed by Bajaj and Arora [10]. In their work, they choose 33 characteristics for categorization out of 41 and compare theoutcomes of several classifiers.

The Simple Cart method has the maximum accuracy of 66.77 percent, whereas the C4.5 decision tree's classification result is only 65.65 percent. Alazab et al. [11] use information acquisition and decision trees to choose characteristics to identify both previous and new threats.

Thaseen and Kumar [12] employed two effective feature selection methods: correlation-based featureselection (CFS) and consistency-basedfeature selection (CFS) (CONS).

In this study, CFS is used to choose 8 characteristics, and Naive Bayes is used to classify them. , C4.5 decision tree, and AD (Alternating Directions) are types of decision trees. The outcomes of each decision tree have been compared. CONS chooses ten helpful traits from a total of 41.Various strategies, such as Random, are used to classify the data Random forest and Decision Tree have both been investigated.

Information gain, gain ratioQuantitative Particle SwarmOptimization using Principle Component Analysis (PCA) and Optimized Least Significant Particles (OLSP-QPSO) are among the feature selection methodologies used to assess and analyse the effectiveness of IDSs (PCA).The OLSP-QPSO approach has a larger amount of attributes reductions and a low false alarms and good rate of detection than other feature selection strategies.

3. DECISION TREE SPLIT (DTS)ALGORITHM

The C4.5 decision tree method is used in the Decision Tree Split (DTS) technique.The issue mainly for the construction of decision tree is the node split value.The suggested algorithm introduces a novel method for determining split value.The algorithm'sstages are as follows:

1. The decision tree's leaf node is built by picking the class that all of the provided training samples belong to.
2. Calculation of gain ratio by splitting the information of split value of attribute.
3. Computation of Information gain.
4. Calculation of entropy
5. By taking all value average in the domain to chose the split value attribute.
6. Determination of highest gain ratio attribute.
7. Building decision node which splitsthe dataset.
8. Splitting the property allows you to repeat steps 1 through 4 on all subsets.

3.1 THE PROPOSED ALGORITHM IMPORTANCE.

For the selection of split value, algorithm sorts attribute value first.Then then sorted values, say, $Q_i + Q_{i+1}. . . . .Q_n$, the gain ratio is derived by selecting a lower value of and Q_{i+1} as a split value and as a threshold value.

The split value for that node is decided based on the value that delivers the maximum gain ratio.We employ a simple and effective strategy instead of all these computations, which makes the process more complex and difficult to grasp.In order to determine the split value, the attribute values do not need tobe sorted in our function.

The split value is determined by averaging the values in a given attribute's domain at each node.It provides all of the values in the domain the same weight,constructing the classifier completely impartial towardsthe most common values in theattribute's domain.

This constraint can be solved by only evaluating qualities with a higher information gain value than the average information gain.

4. DEVELOPMENT AND TESTING

The method runs on a Windows-10 64- bits computer with 8 GB of RAM and an i5 core.The algorithm proposed is compared with the existing algorithmslike CART,C4.5 and AD Tree.

The analysis of the algorithm is done based different parameters like time taken for classifier to build the model,true positive rate, false positiverate and accuracy.

True Positive(TP) are the predictions which are projected to be normal appropriately.True Negative(TN) are the predictions which shows instance correctly.False Positive(FP) are the predictions which are correctly as attack when they don't exist. False Negative(FN) are the predictions which shows normal when the attack is reality.Accuracy is calculated by the total number of accurate forecasts.It is computed as

$$\boxed{Accuracy = \frac{TP + TN}{TP + TN + FP + FN}}$$

True Positive Rate(TPR) is computed as

$$\boxed{TPR = \frac{TP}{TP - FP}}$$

False Positive Rate(FPR) is computed as

$$\boxed{FPR = \frac{FP}{TP - FP}}$$

4.1 DATASET

The proposed techniques efficiency is calculate by experimenting with NSL- KDD dataset, that is a version revised of KDD’99 data set.This constraint can be solved by only evaluating qualities with a higher information gain value than the average information gain.

The NSL-KDD divides traffic two categories, normal and abnormality, for binary categorization.The trials were carried out on a 125973-record data set for training and a 22544-record data set for testing.The method then constructs, trains, and tests the decision tree using the data set with these specified qualities

.

4.2 RESULT AND ANALYSIS

The suggested algorithm's performance is compared to the performance of several approaches.The accuracy in identifying attacks on the NSL KDD testdataset is used to compare the findings.

The findings are based on research that employs approaches like Self-Organizing Maps (SOM), Hoeffding Tree, and Ripple Down Rule Learner Intrusion Detection (RDRID) to train and test their detection model.As demonstrated in Fig. 1, our suggested technique for creating decision trees is effective in attack detection.

The suggested approach, as well CART, the Naive Bayes (NB) Tree, and the AD Tree are examples of other classifiers, that are assessed using the NSL- KDD test dataset.Several classifiers' time is also assessed, and a bar graph is displayed in Fig. 2.The true positive rate of DTS is higher than that of other techniques.

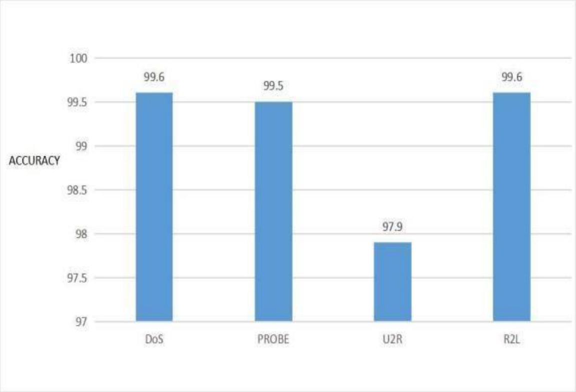

Fig 1- Accuracy of detection of several attacks

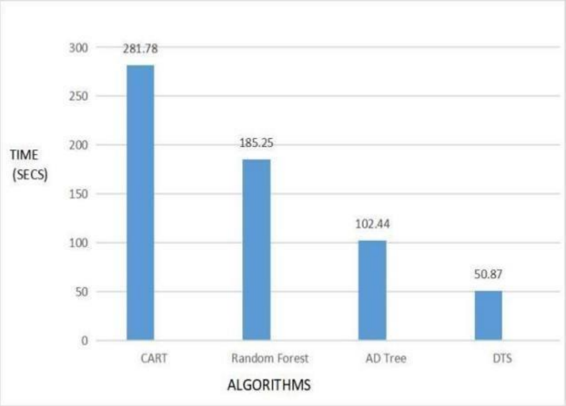

Fig 2 - Comparison of time to constructionmodel of various algorithms

Table 1- Detection Accuracy

SLNO	ATTACKS	DETECTION ACCURACY
1	DoS	99.6%
2	Probe	99.5%
3	R2L	97.9%
4	U2R	99.6%

5. CONCLUSION AND WORK INTHE FUTURE

The administrator of the network will be able to detect incoming traffic using decision tree and know whether the incoming traffic is malicious or non- malicious traffic.The model also separates the non malicious and malicious traffics,By changing the value of split value calculation By taking the split value computation and modifying it, the aggregate of all values in an attribute's domain.The algorithm gives equal weightage fro all values of the domain.It gives good accuracy by taking less number of attributes into count in less amount of time.From experiment's result,its concluded that, the suggested intrusion detection system is more precise and efficient in detecting the attacks in the network Our Future-goal is to increase the split value by employing ideas like geometric mean,which provides domain values uniform weighting.

REFERENCES

[1] P. Aggarwal, and S.K. Sharma, An Empirical Comparison of Classifiers to Analyze Intrusion Detection, Proc. of Fifth International Conference anAdvanced Computing and Communication Technologies, 2015.

[2] Sicato, Jose Costa Sapalo, et al. "A comprehensive analyses of intrusion detection system for IoT environment." *Journal of Information Processing Systems* 16.4 (2020): 975-990..

[3] Avaliable on- http://www.saedsayad.com/oner.htm ,08-07-2022,1;00pm.

[4]M. F. Elrawy, A. I. Awad, and H. F. A. Hamed, "Intrusion detection systems for IoT-based smart environments: a survey," Journal of Cloud Computing, vol. 7, article no. 21, 2018..

[5] C. Kolias, A. Stavrou, J. Voas, I. Bojanova, and R. Kuhn, "Learning Internet-of-Things security 'Hands-On'," IEEE Security & Privacy, vol. 14, no. 1, pp. 37-46, 2016.

[6] J. Markey, Using Decision Tree Analysis for Intrusion Detection: A How-To Guide, SANS Institute InfoSec Reading Room, June, 2011.

[7] Onyebuchi, Okwume B. "Signature based network intrusion detection system using feature selection on android." *Signature* 11.6 (2020): 551-558.

[8] Alsaadi, Husam Ibrahiem, et al. "Computational intelligence algorithms to handle dimensionality reduction for enhancing intrusion detection system." (2020).

[9] D.P. Gaikwad, and R.C. Thool, Intrusion Detection System Using Bagging with Partial Decision Tree Base Classifier,Proc. of the 4th International Conference on Advances in Computing, Communication andControl, 2015.

[10] K. Bajaj, and A. Arora, Improving the Intrusion Detection using Discriminative Machine Learning Approach and Improve the Time Complexity by Data Mining Feature Selection Methods, International Journal of Computer Science, vol. 76, Aug, 2013.

[11] A. Alazab, M. Hobbs, J.Abawajy, and M. Alazab, Using Feature Selection for Intrusion Detection System, International Symposium on Communications and Information Technologies, 2012.

[12] S. Thaseen, and Ch. A. Kumar, An Analysis of Supervised Tree Based Classifiers for Intrusion Detection System, In Proc. of the 2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering, Feb, 2013.

JETIR