International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

An Efficient Network Intrusion Based on Decision Tree Classifier & K-Mean Clustering using Dimensionality Reduction

Vandna Malviya[1], Anurag Jain[2]

[1]RITS Bhopal, MP, CSE Department, India

[2] RITS Bhopal, MP, CSE Department, India

Abstract: ***As the internet size grows rapidly so that the attacks on network. There is a need of intrusion detection system (IDS) but large and increasing size of network creates huge computational values which can be a problem in estimating data mining results this problem can be overcome using dimensionality reduction as a part of data preprocessing. In this paper we study two decision tree classifiers(J48, Id3) for the purpose of detecting any intrusion and comparing their performances .first we have applied data pre processing steps on each classifier which includes feature selection using attribute selection filter , Intrusion detection dataset is KDDCUP 99 dataset which has 42 features after preprocessing 9 selected attributes remains ,then discretization of selected attribute is performed ,simple k-Mean algorithm is used for analysis of data and Based on this study, we have concluded that J48 has higher classification accuracy with high true positive rate (TPR) and low false positive rate (FPR) as compared to ID3 decision tree classifiers.***

Keywords: Dimension reduction, weka, J48, ID3, KDDCUP 99

1. Introduction

The Intrusion detection system (IDS) is the process of monitoring the events occurring in network and analyzing the signs of possible attacks. Intrusion prevention is almost impossible to achieve hence our focus is on intrusion detection either before or after its success. As sooner as we are able to detect an attack, faster we can act. The losses and recovery from an attack is directly proportional to how quickly we are able to detect an intrusion.IDS can act as good deterrent to intruders. The term IDS was first coined in 1980 by James P Anderson [1] and then improved by D. Denning [2] in 1987.There are two basic approaches of intrusion detection anomaly detection and misuse detection. Misuse detection is the capability to identify a known sequence of activities in a system, often referred to as the signatures identified as - Threats. Anomaly detection is based on the prediction that normal user behavior is different from normal user. Using decision tree analysis, the logic of decision tree can be implemented in the intrusion detection system. A decision tree is a method of predicting unknown information which are converted into a tree like structures. Decision tree can prove itself more efficient in terms of retrieval of data for the purpose of making decisions. The decision tree starts with a root node which splits recursively per the possible conditions and its decision. In this paper two decision tree algorithms (J48, Id3) is used. While clustering is the process of selecting similar data from the dataset .goal of clustering algorithm is to generate minimum no of clusters to describe data. Here K-mean algorithm is used for clustering.

2. DataSet and Tool Description

2.1 Kdd Cup 99 Dataset

The KDDcup99[3] Intrusion Detection datasets is based on the 1998 DARPA initiative, which provides researchers of Intrusion Detection Systems (IDS) a benchmark on which to they evaluate different methodologies. In KDDCUP99 Attacks fall into one of four categories:

- *Denial of Service (DoS): In this attacker* tries to prevent legitimate users from using a service.
- *Remote to Local (R2L): In this attacker* does not have an account on the victim machine, hence tries to gain access.
- *User to Root (U2R): In this attacker* has local access to the victim machine and tries to gain super user privileges.
- *Probe: In this attacker* tries to gain information about the target host.

2.2 Simulation Tool

WEKA i.e., Waikato Environment for Knowledge Analysis (WEKA) is used as a simulation tool that would allow researchers easy access to state-of the art techniques in machine learning[6]. Weka provides three options:

- *Weka Explorer:* The Explorer has several panels which provides access to the main components of the workbench. It has panels i.e. Preprocess, Classify, Associate, Cluster.
- *Weka Experimenter*: It provides comparison of weka algorithms in systematic way.
- *Weka Knowledge Flow:* It provides better representation to the Explorer as a graphical front end to weka's core algorithm.

3. Model Classification

3.1J48 Decision Trees

It is a Java implementation of the C4.5 algorithm in the weka which is an open source data mining tool [4]. It builds decision trees from a set of training data, using information entropy. At each node of the tree, J48 chooses one attribute of the data that most effectively splits its set of samples into subsets. Its criterion is the highest and normalized information gain that results from choosing an attribute for splitting the data. The attribute which has highest normalized information gain is chosen to make the decision. For each attribute, the gain is calculated and the highest gain is used in the decision node.

3.2 ID3 Decision Trees

ID3or Iterative Dichotomiser 3 Algorithm [5] is a Decision Tree algorithm. It builds the tree from the top to down, with no backtracking. It makes use of Information Gain to select the most useful attribute for classification. ID3 is based on the Concept Learning System (CLS) algorithm. It is a Class for constructing an unpruned decision tree. ID3 has a drawback that it can only deal with nominal attributes. No missing values allowed in ID3. Empty leaves may result in unclassified instances.

3.3 Simple K-Mean Clustering

K-mean algorithm is a fast clustering algorithm to divide the data in to k groups [8]. Firstly it selects k points as the centroid of the k clusters. Then it calculates the Euclidean distance of each data point to centroid of each cluster. Data point is assigned to the cluster which has minimum Euclidean distance. New centroids are calculated by evaluating the mean of each cluster data points. The process is repeated until specified iterations achieved or same centroid evaluated in successive iterations.

4. Proposed Approach

***A.* Dimensionality Reduction:** Dimensionality reduction Feature extraction transforms a high dimensional data space into a data space of fewer dimensions.

Dimensionality Reduction Algorithm
Steps:
- ✓ Select the dataset.
- ✓ Perform attribute selection to filter out redundant & super flows attributes and
- ✓ Apply discretization for pre-processing the data

Using non redundant data attributes.
- ✓ Apply classification algorithms and compare their performances.
- ✓ Identify the Best One.

The original dataset consist of 41 attributes and one class label with their type of attack [7]. The following list out the attribute names:

*(i)41Attributes:*duration,protocoltype,service,Flag,src_bytes,dst_bytes,land,wrong_fragment,urgent,Hot,num_field_logins,logged_in,num_compromised,root_shell,su_attempted,num_root,num_file_creation,num_shells,num_access_files,num_outbounds_cmds,is_hist_login,is_guest_login,count,srv_count,serror_rate,srv_serror_rate,rerror_rate,srv_rerror_rate,same_srv_rate,diff_srv_rate,srv_diff_host_rate,dst_host_count,dst_host_srv_count,dst_hosdst_same_srv_rate,dst_host_diff_srv_rate,dst_host_same_src_port_rate,dst_host_srv_diff_host_rate,dst_host_serror_rate,dst_host_srv_serror_rate,dst_host_rerror_rate, dst_host_srv_rerror_rate.

Using proposed approach we obtained reduced dimensionality of 9 potential attributes which are listed as follows:
(ii) 9 *Attributes:* service, Flag, src_bytes, count, srv_count, dst_host_count,dst_host_srv_count,,dst_host_same_src_port_rate,label

B. System Architecture:
The suggested architecture for the system is presented in the figure 1.

In this first the KDDCup dataset is preprocessed and after applying different classification algorithm, the classification results are generated.

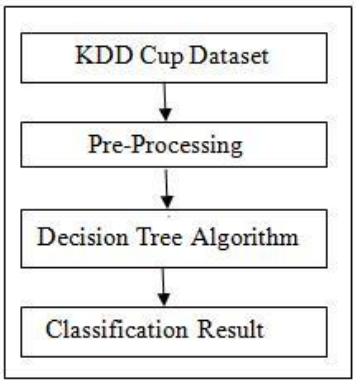

Figure 1: Flowchart of Proposed system

5. Model Evaluation

A Cross Validation
Cross-validation is a technique for showing how the results of a statistical analysis will generalize to an independent data set. It is mainly used where the goal is prediction, and estimating how accurately a predictive model will perform. 10-fold cross validation is frequently used method.

B Criteria for Evaluation
To estimate the performance of any model Accuracy, Sensitivity, Specificity, and Receiver Operating Characteristics Curve (ROC) along with Kappa statistics and correctly classified Instance are employed as major criteria. The accuracy, sensitivity and specificity were calculated by True Positive, False Positive, False Negative and True Negative.

a)Accuracy- Accuracy means probability that the algorithms can correctly predict positive and negative examples.
Accuracy=TP+TN/ (TP+ FN +TN+ FP)

b) *Sensitivity*- Sensitivity means probability that the algorithms can correctly predict positive examples. Precision (also called positive predictive value) is the fraction of retrieved instances that are relevant. High precision means that an algorithm returned substantially more relevant results
Precision = TP/ (TP+FP)

c) *Recall* - Recall (also known as sensitivity) is the fraction of relevant instances that are retrieved. High recall means that an algorithm returned most of the relevant results.
Recall=TP/ (TP+FN)

C Confusion Matrix

A confusion matrix is a visualization tool typically used in supervised learning (in unsupervised learning it is typically called a matching matrix). A confusion matrix that summarizes the number of instances predicted correctly or incorrectly by a classification model.

- The true positive rate (TPR) or sensitivity is defined as the fraction of positive examples predicted correctly by the model, i.e. TPR = TP / (TP + FN)
- The true negative rate (TNR) is defined as the fraction of negative examples predicted correctly by the model, i.e., TNR = TN / (TN + FP)
- False positive rate (FPR) is defined as the fraction of negative examples predicted as a positive class the model, i.e., FPR = FP / (TN + FP)
- The false negative rate (FNR) is the fraction of positive examples predicted as a negative class. i.e, FNR = FN / (TP + FN)

6. Simulation Result

Simulation shows that J48 has higher classification algorithm then existing algorithm and then we are comparing results with ID3 algorithm.

Table 1: Comparison of proposed J48 algorithm with existing algorithm and Id3

Algorithm	Classification accuracy	TP rate	FP rate	Precision	Recall	F-Measure
J48(C4.5)	99.1525	0.992	0.04	0.99	0.993	0.989
J48 proposed	99.5763	0.996	0.02	0.993	0.996	0.995
ID3	98.5876	0.994	0.01	0.97	0.98	0.985

Table 2: Comparison of proposed J48 algorithm with ID3 algorithm

Parameter	Classifier		
	J48 existing	J48 proposed	ID3
Correctly classified instances	702	705	698
Incorrectly Correctly classified instances	6	3	4
Kappa Statics	0.988	0.994	0.99
Mean Square Error	0.001	0.0007	0.0005
Root Mean Square Error	0.0271	0.0192	0.0223
Relative Absolute Error	1.5634	1.1701	0.8029
Root Relative Squared Error	15.4591	10.9477	12.7591

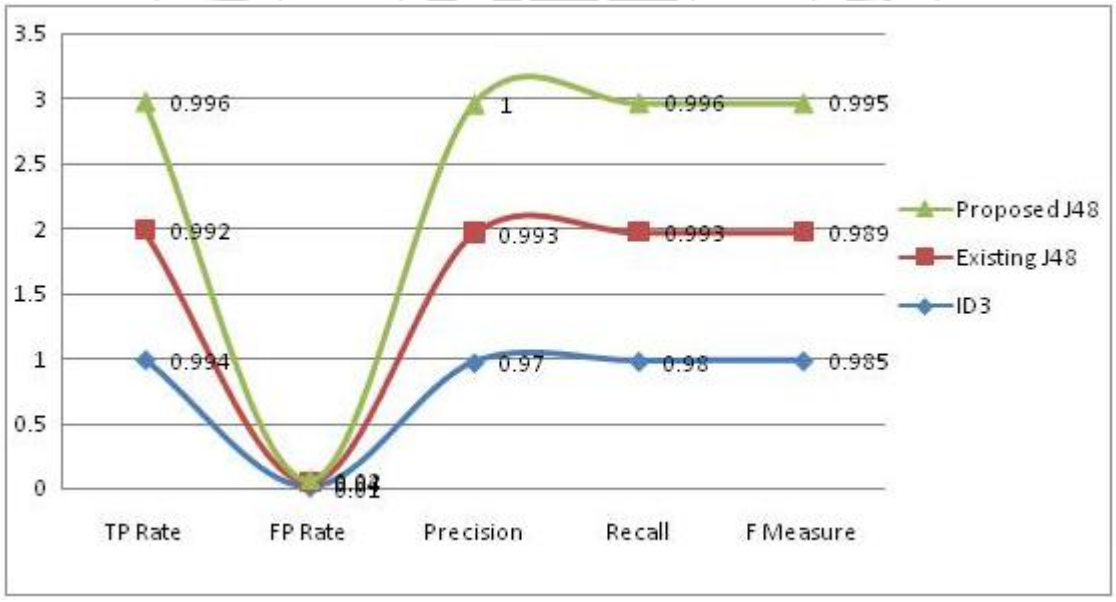

Figure 2: Comparison of proposed and existing algorithm with ID3

7. Conclusion

After the implementation of Decision Tree on an Intrusion Detection System, we have been able to classify the breach as attack or normal. Where the decision tree has been implemented to improve the precision of the system, the Dimensionality Reduction done on the data obtained from KDD cup 99 dataset. In previous approaches various has been used for the anomaly detection in proposed approach J48 has been used with modifications which shows improved results having higher accuracy rate with high true positive rate(TPR) and low false positive rate(FPR). The objective of this work is to study how decision tree algorithms are used in the classification the Intrusion Detection Attacks. In this

study dimension reduction plays an important part in this work to evaluate the Performance of two decision tree algorithms as J48 and id3 and their performance is compared using weka toolkit. Here the evaluation criteria is the Specificity, Accuracy, Sensitivity are evaluated to get the respective True Positive, false positive rate for both the algorithms result it is observed that J48 performs better classification and accuracy for reduced dimensionalities.

References

[1] James P. Anderson, "Computer security threat monitoring and surveillance," Technical Report 98-17, James P. Anderson Co., Fort Washington, Pennsylvania, USA, April 1980.
[2] D.Denning, "An Intrusion Detection Model", IEEE Transaction on Software Engineering, 13(2), 1987.pp.222-232
[3] The KDD Archive. KDD99 cup dataset, 1999. http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html
[4] Korting, Thales Sehn. "C4. 5 algorithm and Multivariate Decision Trees." Image Processing Division, National Institute for Space Research--INPE.
[5] Huang Ming, Niu Wenying, Liang Xu".An improved decision tree classification algorithm based on ID3" and the application in score analysis
[6] H. Witten, Eibe Frank, Len Trigg, Mark Hall," Weka: Practical Machine Learning Tools and Techniques with Java Implementations"
[7] R. Shanmugavadivu, Dr. N. Nagrajan "Network Intrusion Detection System Using fuzzy Logic" Indian Journal of computer Science and Engineering.
[8] Mac Queen, J. B.1967, Some Methods for Classification and Analysis of Multivariate Observations. Proc. 5th Berkeley Symp. Mathematical Statistics and Probability, University of California Press, pp 281- 297

Author Profile

Ms Vandna Malviya, M.Tech student, Department of Computer Science and Engg, Radharaman Institute of Technology and Science, Bhopal, M.P. India

Prof. Anurag Jain, Astt. Professor and Head, Department of Computer Science and Engg, Radharaman Institute of Technology and Science, Bhopal, M.P. India