IOSR Journal of Engineering (IOSRJEN)
ISSN: 2250-3021 Volume 2, Issue 7(July 2012), PP 54-57
www.iosrjen.org

Knowledge Based Analysis of Statistical Tools in Attack Detection

Renu Yadav[1,] Dr. Kanwal Garg[2]
[1]*M. Tech Scholar Department of Computer Science & Applications Kurukshetra University Kurukshetra, Haryana India*
[2]*Assistant Professor Department of Computer Science & Applications Kurukshetra University Kurukshetra, Haryana India*

Abstract: - *With the network playing more and more important effect in the modern society, crimes using computer network are presenting the obvious trend of escalation. In order to ensure network security, one technique adopted for detecting abnormal or unauthorized behavior is the Intrusion Detection System (IDS). Various data mining techniques are applied in an offline environment to add more depth to the network defense in order to determine the various attacks or threats to the network. This work focuses on finding the best classifier with respect to accuracy in detecting various attacks on the tcpdump data so that mechanisms can be incorporated in Intrusion Detection Systems to detect the misclassified types of attacks efficiently. This helps to reduce the rate of misreporting and thus enhancing the efficiency of the Intrusion Detection System.*
Keywords: - *Intrusion Detection System (IDS), data mining, decision trees, CART, J48.*

I. INTRODUCTION

Data mining refers to knowledge discovery in database, these data are usually numerous, incomplete, indistinct and random. Data, if as original information, can be termed as knowledge but generally, knowledge refers to concepts, rules and restraints.

The methods used for finding knowledge can be mathematical or non-mathematical; it can be deductive or inductive. The available knowledge can be used for optimizing enquiry, manage information, control progress and make intellectual decision. Therefore, data mining can be regarded as a crisscross subject that helps people in application of data from low and simple inquiry to discover knowledge in data and support decision.

In the analysis of Intrusion Detection System, the data circulating in network has the following characteristics: mass data, even if a small commercial website, the number of data message sent and received are quite impressive and incomplete whose transportation is busy, data message which overweigh network carry will be discarded: noisy, when network is unstable, data information may get changed in the transportation of message [Liu W.]. It can be seen that these data is in accordance with the feature of data mining, naturally data mining need to be applied to Intrusion Detection System.

The various detection models need log data as training collection whose accuracy will largely influence Intrusion Detection System. Because of the density and accuracy of visit network, it is difficult to acquire completely no attack action. In addition, it is also uneasy to log attack behavior. Data mining technology can resolve this problem, in the analysis of general network visit, isolated point is an invasive behavior to reduce the difficulty of acquisition of training data.

Intrusion Detection System is a passive method in the security field, it monitors information system and sends out warning when it does detect intrusion, but data mining technology can analyze these data when network message is acquired, it can forecast for visit on its own initiative, thus reduce the frequency of matching, thus achieve the function of active defense. Data mining technology, for instance, Clustering, Classification, Feature Summary, Association Rules can be applied in the intrusion detection system. It has been proved that data mining technology improves the property of Intrusion Detection System and the processing rate.

II. DATA MINING AND INTRUSION DETECTION

Data Mining is assisting various applications for required data analysis. Recently, data mining has become an important component in Intrusion Detection System. Different data mining approaches like classification, clustering, association rule, and outlier detection are frequently used to analyze network data to gain intrusion related knowledge. This section will elaborate the data mining technique used in the context of intrusion detection.

A.The Classification Terminology

The classification task evaluates a function or makes a relationship between a dependent variable and contingent independent variables by mapping the data points. A classification problem is simply stated as identification of an object as belonging to a given class. The classes in a classification problem are made predefined & non-overlapping before the application of any algorithm.

Classification algorithms always find a rule or a set of rules to organize data into classes. A classification rule or formula for making any decision may be developed from a part of the historical data and tested on the remaining data. The part of data is used to develop a model is called the training data and the part that is used to test the model is called the test data. A too good model is developed with the training data that captures unimportant details; when that model is applied to new instances, the unimportant functionalities may distort the results, leading to poor performance. There are many methods available for analyzing the data using the decision tree technique of classification. The two decision tree methods used in this work are classification and regression trees (CART) and J48.

B. Classification And Regression Trees (CART)

The CART methodology proposed by Breiman and associates Berry and Linoff in the year 1997 is perhaps best known and most widely used. CART uses cross-validation or a large independent test sample of data to select the best tree from the sequence of trees considered in the pruning process. The basic CART building algorithm is a greedy algorithm in that it chooses the locally best discriminatory feature at each stage in the process. This is suboptimal but a full search for a fully optimized set of question would be computationally very expensive. The CART approach is an alternative to the traditional methods for prediction [Breiman et al.] [Steinberg and Colla(1995)] [Steinberg and Colla (1997)]. In the implementation of CART, the dataset is split into the two subgroups that are the most different with respect to the outcome. This procedure is continued on each subgroup until some minimum subgroup size is reached.

C. Decision tree J48

The decision tree J48 [Quinlan (1992)] implements Quinlan's C4.5 algorithm [Quinlan (1993)] for generating a pruned or unpruned C4.5 tree. C4.5 is an extension of Quinlan's earlier ID3 algorithm. The decision trees generated by J48 can be used for classification. J48 builds decision trees from a set of labeled training data using the concept of information entropy. It uses the fact that each attribute of the data can be used to make a decision by splitting the data into smaller subsets. J48 examines the normalized information gain (difference in entropy) that results from choosing an attribute for splitting the data. To make the decision, the attribute with the highest normalized information gain is used. Then the algorithm recurs on the smaller subsets. The splitting procedure stops if all instances in a subset belong to the same class. Then a leaf node is created in the decision tree telling to choose that class. But it can also happen that none of the features give any information gain. In this case J48 creates a decision node higher up in the tree using the expected value of the class.

III. RESULT AND EXPERIMENTAL DETAILS

The experimentation detailed in this section was carried out within the Waikato Environment for Knowledge Analysis (WEKA) suite for machine learning. This software, developed in the Java programming language, offers a powerful testing harness for analysis of various data mining concepts and implementations.
The performance of classification algorithms is evaluated by using the tcpdump dataset. In the first step, the data is applied into the preprocessing techniques. After this preprocessing, the data is applied into the classification algorithm. Then the performance accuracy per attack class is computed by using correctly predicted instances (true positives and true negatives) as:

Accuracy= (TP+TN)/ (TP+TN+FP+FN)

In the above formula, TP, TN, FP and FN stand for true positive and true negative false positive and false negative respectively.

Table I shows the performance of the two classification algorithms in terms of accuracy for the tcpdump dataset with correctly and incorrectly classified instances.

Table I Classification Accuracy for tcpdump data.

Classifier Name	Total number of instances	Incorrectly Classified Instances	Correctly Classified Instances	Accuracy in terms of Correctly Classified
Simple CART	295	9	286	96.95%
J48	295	5	290	98.31%

The graphical representation of the performance of the two classifiers is represented in figure 1 below:

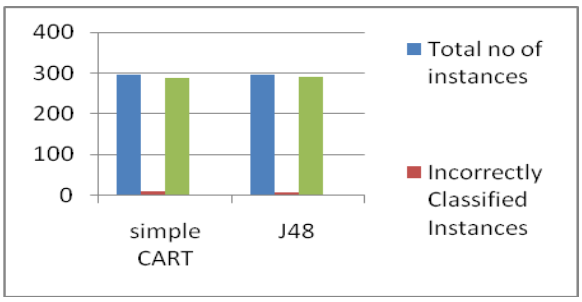

Figure 1 Performance Comparison of the two Classifiers.

The performance of the two classifiers in terms of accuracy for each attack class on the tcpdump dataset is computed by making use of the correctly predicted instances.
The results are summarized in Table II below:

Table II Classification Accuracy per attack class for tcpdump data.

Class	Attack Name	J48	Simple CART
a	-	100	96.94
b	phf	99.66	99.66
c	guess	99.32	98.64
d	rcp	98.98	99.66
e	rlogin	99.66	99.66
f	rsh	98.98	99.32
g	port-scan	100	100

The accuracy comparison for each attack class in the tcpdump dataset for both the classifiers is shown in figure 2 below:

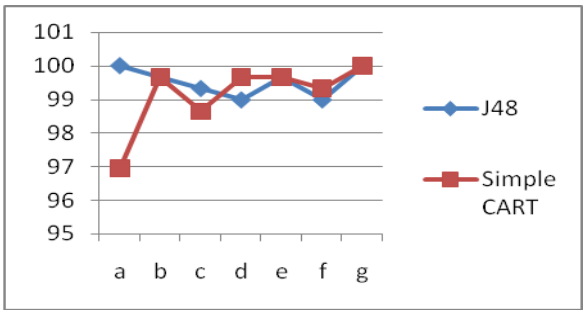

Figure 2 Performance Comparison per Attack Class.

IV. CONCLUSION

Decision tree based classifiers classify the dataset into certain classes. Rules can be generated according to the classification result. The result demonstrates that the J48 classification methodology provides better results over the simple CART methodology. It is found that the accuracy given by the J48 classifier for the tcpdump data is 98.31% whereas that of the simple CART is 96.95%. The accuracy per attack class achieved by the J48 is better than the simple CART as the number of incorrectly classified instances is less in case of the J48 classifier. The results are mainly used for considering the most accurate tool to be considered in detecting intrusions or attacks so that actions can be taken against the misclassified types of attacks. This information helps to build Intrusion Detection Systems which can correctly identify various attacks and thus helps to reduce the rate of misreporting.

References

[1] Breiman L., Friedman J., Olshen R., Stone C., "Classification and Regression Trees", 1984.
[2] Liu W., "Research of Data Mining in Intrusion Detection System and the uncertainty of the attack", 978-1-4244-5273-6/09, IEEE, 2009.
[3] Quinlan R.J., "Learning with Continuous Classes", 5th Australian Joint Conference on Artificial Intelligence, Singapore, 1992, pp. 343-348.
[4] Quinlan R., "C4.5: Programs for Machine Learning", Morgan Kaufmann Publishers, 1993.
[5] Steinberg D., and Colla P.L., "CART: Tree-Structured Nonparametric Data Analysis", Salford Systems, 1995.
[6] Steinberg D., and Colla P.L., "CART-Classification and Regression Trees", Salford Systems, 1997.