Intrusion Detection Combining Multiple Methods

Dai Hong[1], **Li Haibo**[2]

1. *Institute of Software, Liaoning University of Science & Technology, Anshan 114051, Liaoning ,China*
2. *Department of Graduate, Liaoning University of Science & Technology, Anshan 114051, Liaoning, China*

1.dear_red9@163.com, 2.hi_bird@sina.com

Abstract: The Intrusion detection system deals with huge amount of data which contains irrelevant and redundant features. An appropriate feature set got by feature selection can help to build lightweight intrusion detection system. This paper presents a method of Chi Square combining multiple decision trees based on enhanced C4.5 algorithm to select minimal feature set. We have examined the feasibility of our approach by using KDD 1999 CUP dataset. The experiment results show that the method not only has more higher detection rate, but also has more lower false positive rate to DoS and Probe attacks while retaining training and testing time.

Keywords: intrusion detection; Chi Square; enhanced C4.5 algorithm; multiple sub-decision trees

1. Introduction

Network Intrusion Detection Systems (NIDS) plays a vital role in network security. As an active defense technology, it attempts to identify existing attack patterns and recognize new intrusions. Due to deal with huge amount of data which contains irrelevant and redundant features causes slow training and testing process, higher resource consumption as well as lower detection rate. In the case of anomaly detection, people select statistical character to build IDS mainly depending by their intuition and experience[1]. Anomaly detection usually has high false positive rate. In the case of misuse detection, the knowledge base needs to update regularly. This updating has to be done by experts or the designers of the system[2]. Misuse detection cannot detect unknown intrusions.There has been an increased interest in data mining approaches to building detection models for IDS, which can be generated in a quicker and more automated method than manually encoded models. Regardless of the detection technique by IDS used, one of the first challenges has to be resolved is the feature selection process. This decision will later influence both the performance of the system as well as the types of attacks that the system will target.

We propose a feature classification scheme for NIDS that is intended to provide a better understanding of the large number of attributes that can be extracted from network packets, their relationships, as well as their usefulness in detecting different types of attacks.

The subsequent sections are organized as follows. In section 2, describes the systems architecture of the new CS-EC4.5(Chi Square/χ^2 Statistic - Enhanced C4.5 Algorithm) classifier and gives a brief introduction on both feature selection based on Chi Square and Enhanced C4.5 Algorithm that are designed and implemented in this research work. In Section 3, we discuss the experimental results and its possible implications. In Section 4, gives conclusion and plans for future works.

2. Proposed Approach

2.1 System Architecture

The overall architecture of our approach is depicted in Figure 1.

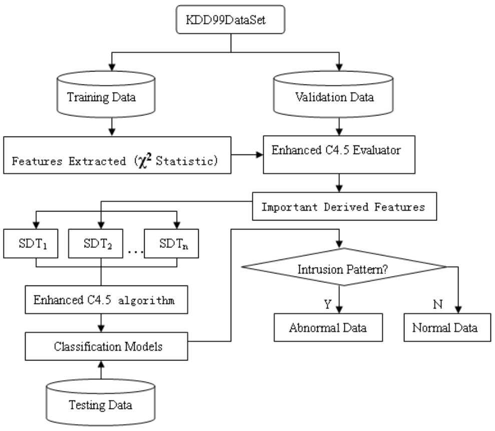

Figure 1. Overall Architecture of Proposed Approach

The experiment used data from the third international knowledge discovery and data mining tools competition (KDDcup’99) to train and test the feasibility of this proposed model. We experiment with the CS-EC4.5 by using the 10% subset of KDDcup’99 data[3]. In the training process, training data is preprocessed and passed to our feature-selection engine using Chi-Square(χ^2 Statistic) approach. We get a set of extracted features with the order of importance. Afterwards, using enhanced C4.5 algorithm test over validation set and then derive importance features. In next phase, the order importance derived features are used to build decision tree. In contrast to using only a single tree classifier to classify intrusions, the basic idea of the model is to divide a large decision tree into several sub-decision trees to design NIDS, mine on sub-decision trees by enhanced C4.5 algorithm. In the last process, testing data is sent to our intrusion detection model to detect. We can detect the data is attack data if the intrusion pattern is found, otherwise the data is normal.

2.2 Feature Selection Based on Chi-Square

Feature construction is needed to extract a set of features which can detect intrusions effectively. In the attributes set of network transaction and system behavior, some attributes are critical, some attributes just give the reference information to the intrusion detection. In probability theory and statistics, the Chi-Square distribution (also χ^2 distribution) is one of the most widely used theoretical probability distributions in inferential statistics, e.g., in statistical significance tests[4]. A Chi-Square approach is a simple and general algorithm that can automatically select a proper χ^2 value, statistic to discretize numeric features repeatedly until some inconsistencies are found in the data, and achieves feature selection via discretization[5]. The measure is defined to be:

$$\chi^2 = \sum_{m=1}^{2}\sum_{n=1}^{k}\frac{(A_{mn}-E_{mn})}{E_{mn}}^{2} \qquad (1)$$

where:

A_{mn} = no. patterns in the m th interval, n th class,

E_{mn} = expected frequency of $A_{mn} = \frac{1}{N}R_m * C_n$

R_m = no. patterns in the m th interval = $\sum_{n=1}^{k} A_{mn}$

k= number of (no.) classes,

N = total no. patterns = $\sum_{n=1}^{2} R_m$

C_n = no. patterns in the n th class= $\sum_{n=1}^{2} A_{mn}$

If either R_m or C_n is 0, E_{mn} is set to 0.1. The degree of freedom of the χ^2 statistic is one less the number of classes. Therefore, in this paper, we will use Chi-Square based on the above discussions to fulfill feature selection task to utmost distinguish five classes, including normal, DoS(Denial of Service), Probe, U2R(User to Root), R2L(Remote to Local). By using only part of features out of the 41 features provided by the KDD 99 constructs a small subset of features as input data for the next phase work. Section 3 will give the detailed results.

2.3 Enhanced C4.5 Algorithm

The decision tree classifier by Quinlan[6] is one of most well-known machine learning techniques. The C4.5 algorithm[7] developed by are probably the most popular ones. The C4.5 algorithm to construct one decision tree where each node of the tree specifies a test on an attribute, and each branch of the node corresponds to one of its values. The leaves are the classification results. The top-most node in a tree is the root node. The tree is a model generated by the classification algorithm. The C4.5 algorithm builds a decision tree, from the root node, by choosing one remaining attribute with the highest information gain as the test for the current node. The notion of information gain introduced earlier tends to favor attributes that have a large number of values. The splitting criterion is very important in the process of building the tree, because it determines if we must attach a node or a leaf as next element in the tree. In the paper, we introduce an information gain ratio instead of information gain in order to express the proportion of useful information generated by split in an effective manner.

Split info is defined by SplitInfo(X)

$$\text{SplitInfo(X)}= -\sum_{i=1}^{n}\frac{|T_i|}{|T|} * Log_2\left(\frac{|T_i|}{|T|}\right) \qquad (2)$$

Split info is the information due to the split of T on the basis of the value of categorical attribute D.

The gain ratio, expresses the proportion of useful information generated by split and it is computed as:

$$GainRatio(\mathrm{F},T) = \frac{Gain(\mathrm{F},T)}{SplitInfo(\mathrm{F},T)} \qquad (3)$$

The idea is to partition the training set in such a way that the information needed to classify a given example is reduced as much as possible.

Based on the previous feature set, experiment uses enhanced C4.5 to evaluate data validation of KDD99 dataset and features extracted with Chi-Square.

Table 1 Feature Selection Results Based on Chi-Square

Rank	Feature
NO.1	dst_host_rerror_rate
NO.2	src_bytes
NO.3	dst_bytes
NO.4	dst_host_count
NO.5	srv_count
NO.6	dst_host_srv_rerror_rate
NO.7	srv_rerror_rate
NO.8	count
NO.9	service
NO.10	dst_host_same_src_port_rate
NO.11	dst_host_srv_diff_host_rate
NO.12	diff_srv_rate

Table 2 Experimental results based on multiple trees

Class	true detection rate	false positive rate
Normal	99.15%	0.85%
DoS	96.86%	1.31%
Probe	95.52%	1.93%
U2R	69.25%	10.63%
R2L	56.71%	12.21%

Table 3 Experimental results based on a single tree

Class	true detection rate	false positive rate
Normal	98.95%	1.05%
DoS	95.56%	1.52%
Probe	94.23%	2.39%
U2R	67.12%	11.81%
R2L	50.83%	13.21%

Important derived features are applied to multiple sub-decision trees and use enhanced C4.5 algorithm to generate classification models to distinguish intrusions from legal behaviors efficiently. The lower the classification error rate, the better the fitness of the feature set.

3. Experimental Results and Analysis

We have used open source WEKA library for Chi-Square and C4.5 algorithm. The main aim of features selection is to determine a minimal feature subset from a problem domain while retaining a suitably high accuracy in representing the original features. From Table 1 shows using the Chi-Square to feature selection results. The top 12 is the most important features with Chi-Square. These features are really significant in classifying the data.

After important features selected based on Chi-Square, We applied the selected features to multiple sub-decision trees by the enhanced C4.5 algorithm. The rules are then applied to the testing data. The testing data is thus classified using the rules generated during the training phase. Network intrusion detection systems built using our approach on 10% KDD 99 training datasets named CS-EC4.5.

A number of observations and conclusions are drawn from the results reported: ①Using CS-EC4.5 based on multiple sub-decision trees, we have achieved more higher detection rate for DoS and Probe attacks. However, the low number of instances for U2R and R2L in “10% KDD Cup 99”dataset makes the lower detection rate respectively in this work. ②Table2 and Table3 experimental results show that the detection performance of multiple sub-decision trees is better than that of individual decision trees. ③ Chi-Square approach can be successfully used to compute maximum importance attributes for detecting attacks. ④The classification results provided by this system can be useful for both misuse detection and anomaly detection, but it is more commonly used for misuse detection.

4. Conclusions and Future Works

In this paper, we propose a new framework of CS-EC4.5 to filter out the irrelevant and redundant features combing results from multiple sub-decision trees in order to build lightweight NIDS. The experiment results show that our method not only has more higher detection rate, but also has more lower false positive rate while retaining training and testing time.

In our future research, we will investigate the feasibility of implementing the technique in real time network intrusion detection environment as well as characterizing type of attacks such as U2R and R2L which enhance the capability and performance of NIDS.

References

[1] Hengtai Ma, Dangen Ren, Sihan Qing. A Survey of Intrusion Detection Research on Network Security[J]. Journal of Software, 2000,11(11):460-1466.

[2] E.Biermann, E.Cloete, E.M.Venter. A comparison of intrusion detection systems[J]. Computers & Security,200137(3):676-683.

[3] University of California Irvine[EB/OL]. http://archive.ics.uci.edu/ ml/databases/kddcup99/kddcup99.html. 2008,11.

[4] Mood, Alexander; Franklin A. Graybill, Duane C. Boes. Introduction to the Theory of Statistics (Third Edition).1974: 241-246.

[5] H. Liu, Setiono and R. Chi2: feature selection and discretization of numeric attributes[A]. In Proc of the Seventh International Conference on Tools with Artificial Intelligence[C].1995: 388 - 391.

[6] J. R. Quinlan. C4.5: Programs for Machine Learnin[M]. Morgan Kaufmann Publishers, 1993.

[7] R. J. Henery, “Classification,” Machine Learning Neural and Statistical Classification, 1994.