International Journal of Electronics Communication and Computer Engineering
Volume 2, Issue 2, ISSN (Online): 2249–071X, ISSN (Print): 2278–4209

Improve Intrusion Detection for Decision Tree with Stratified Sampling

Devendra kailashiya
SATI, Vidisha
dskailashiya@gmail.com

Kanak Saxena
SATI, Vidisha
ks.pub.2011@gmail.com

Abstract: The present paper aims to improve accuracy of intrusion detection for decision tree algorithm. A number of techniques available for intrusion detection. In this paper we have supervised learning with preprocessing step for intrusion detection. The database is generated i using the stratified sampling techniques and the classification algorithm is applied on the samples. The accuracy of proposed model is compared with existing results in order to verify the validity and accuracy of the proposed model.

Index Term: Intrusion Detection, Decision Tree, Stratified Sampling, ID3, preprocessing and classification.

1. INTRODUCTION

There are currently major threats in Network-borne attacks to information security. As an important technique in the Defense-in-depth network security framework. Intrusion Detection has become a widely studied topic in computer networks in recent years [4]. Intrusion detection Technique fall into two Major categories: signature-based detection and anomaly detection. Intrusion Detection Systems (IDS) can also be categorized as Host-Based IDSs and Network-Based IDSs according to the target Environment for detection.

Host-based IDSs usually monitor the host system behavior by examining the Information of the system, such as CPU time, system calls and command sequences.

Network-based IDSs, on the other hand, monitor Network behavior usually by examining the content as well as some statistical attributes of network traffic.

In 1999, Lee et al. [5] constructed 41 attributes from raw traffic Data (i.e., tcpdump files) to build classification models for network based intrusion detection [5]. The raw traffic Data was collected at MIT Lincoln Laboratory for the 1998 DARPA Intrusion Detection Evaluation program. The 41 attributes have been shown effective for network intrusion detection and the attribute sets of The network traffic have also been used as KDD Cup 1999 data (The 1999 Knowledge Discovery and Data Mining Tools Competition) used *Ripper* to mine some detection rules from the attribute sets and to build misuse detection[6] models used unsupervised methods, namely, cluster based Estimation, k-Nearest Neighbor (kNN) and one class Support Vector Machines (SVM) for network intrusion Detection. Jin et al. utilized the covariance matrices of sequential samples to detect multiple network attacks. We used Principal Component identify some important on the Analysis (PCA) for network intrusion Detection [13] based on the KDD Cup 1999 data. Data attributes based involved in current computer networks increases very fast and is naturally massive. In experiments carried out by MIT Lincoln Lab for the 1998 DARPA evaluation, for example, network traffic over 7 weeks contains four gigabytes of compressed binary tcpdump data that were processed into about five million connection records. Practical IDS, therefore, should have the capacity of fast processing large amounts of network data so that actions for response can be taken as soon as possible before substantial damage is done. Most existing network intrusion detection methods detect intrusions by using all the 41 attributes constructed from network traffic data. However, some of the attributes may be redundant or even noise and therefore decrease the detection system's performance. Empirical evidence from the attribute selection literature also shows that redundant information as well as irrelevant attributes should be eliminated for efficient classification tasks. Sung and Mukkamala used ANN and SVM to performance comparison. For example, an attribute is identified as important if the detection accuracy decreases and/or computation time increases after the attribute is deleted from the training set. We used different criteria to select key attributes. Filter (e.g., Information Gain) and Wrapper [13] (with Bayesian Networks and decision trees classifiers) based attribute selection methods are used to select some key subsets from the 41

attributes. The subsets of attributes are then used for fast intrusion detection. This largely simplifies the detection problem because only a smaller set of attributes is required to extract from raw network traffic and to process in detection step [16]. Detection accuracy becomes better compared with that of using all the 41 attributes with both Bayesian Networks (BN) and Decision trees classifiers. To further validate our methods, we used the 10 key attributes that were identified for DoS attacks based on KDD Cup 1999 data to detect DoS attacks in a real network.

II. DECISION TREE

Decision trees are a popular structure for supervised learning. Its construction process is top down, divide and conquer, and also a greedy algorithm. The basic ID3 algorithm works well for limited number of records in data set and it cannot handle missing values and also when the data set size is increased the tree is not accustomed to the changes [1].

One of the greatest advantages of decision tree classification algorithm is that: It does not require user to know a lot of background knowledge in the learning process [13].

The ID3 algorithm uses the Information Gain to select a splitting attribute and then construct the tree. There are some other algorithms that consider the gini index and gain ratio to select the splitting criterion and attribute.

For ID3 decision tree, concept used to quantify information is called entropy. Entropy is used to measure the amount of uncertainty in a set of data. When all data in a set belongs to a single class the entropy is zero that is there is no Uncertainty [14].

The following Step is done recursively [15].

1) Computing the Information Gain for each Attribute.
2) The attribute with the highest information gain, is selected as a splitting attribute.
3) If the selected attribute is discrete (categorical), the node is branched with all possible values. If the attribute is continuous, a cut point with the highest Information gain is selected.
4) After splitting, consider whether or not these new nodes are leaves; otherwise, new nodes are the root of the sub tree.
5) Repeat Step 1 to 4.

Given the probabilities P1, P2, . . . , Pn

Where $\sum_{i=1}^{n} P_i = 1$,

Entropy I is calculated as

$$I(P1, P2 \ldots Pn) = \sum_{i=1}^{n} (P_i \log (1/P_i)) \quad (1)$$

$$\text{Information Gain } (D,G) = I(D) - \Sigma p(Di) I(Di) \quad (2)$$

Where D= Select attribute

The Proposed algorithm makes use of Information gain to construct the decision tree.

III. DATA SAMPLING

Sampling is the process of selecting representative which indicates the whole data set by examining a part. Sampling is needed in order to make abstraction of complex problem as well as it is used to acquire a sub set that is inferring a larger data set. It is widely accepted that a fairly modest-sized sample can sufficiently characterize a much larger population [18].

A Sampling Technique

Simple Random Sampling : Each data record has the same probability of being included in the sample.

Weighted Sampling: In which the inclusion probabilities for each element of the population is not uniform, each element in the population has a different probability of being selected in the sample according to a defined criteria [19].

Stratified Sampling : In stratified sampling, one or more categorical variables are specified from the input data table to form strata (or subsets) of the total population by dividing the area up into a number of strata such that within each of the strata the values of the variable of interest are expected to be relatively similar[18].

B. Stratified sampling

```
Stratified(S,P,Q )
{
finalEstimate = 0
for i = 1 .. P
{
if P.W>X
{
curEstimate = 0
for j = 1 .. Q
{
```

```
curEstimate += S(
uniformRandRange( (i-1)/S, i/S ) );
}
finalEstimate += curEstimate/Q;
}
}
return finalEstimate;
}
```

Where X= Weight of define at an instance

C. Improve Decision Tree Algorithm

PROCEDURE Build Tree (D, ATTR)
{
Build (D);
IF (all Risk Class values of sample data in Data are The same)
THEN Return Q as a leaf node;
ELSE
{
FOR (each attribute in ATTR)
{
IF (the attribute of the node hasn't been used to be a classification attribute before) THEN
Compute the information gain of the
Attribute of the node;
}
IF (the attribute whose information gain is the Biggest (>0) is marked as ATT) THEN
{
Mark node as the node which needs to be divided
Next step according to ATT ;
Divide Q into Qk, and generate each branch of the node Q;
}
ELSE
{
Return the node as a leaf node;
}
FOR (each branch Qk) Build Tree (Q, ATTR);
}
}

Where D=Data, ATTR= ATTRIBUTE

IV. KDD CUP 1999 DATASET

The Experimental data comes from KDD CUP 1999 dataset [17]. It is test set widely used in Intrusion detection field. It includes about 4.9 million simulative attack records and 22 types of attack. Because the entire data set is to large. We select 10% subset as training dataset.

Specifically, the four broad classes of attack type defined in IDS as: DoS, Probe, R2L and U2R.

Denial-of-Service (DoS): These are attacks designed to make some service accessible through the network unavailable to legitimate users.

Probe: A Probe is reconnaissance attack designed to uncover information about the network, which can be exploited by another attack.

Remote-to-Local (R2L): This is where an attacker with no privileges to access a private network attempts to gain access to that network from outside, e.g. over the internet.

User-to-Root (U2R): The attacker has a legitimate user account on the target network. However, the attack is designed to escalate his privileges so that one can perform unauthorized actions on the network.

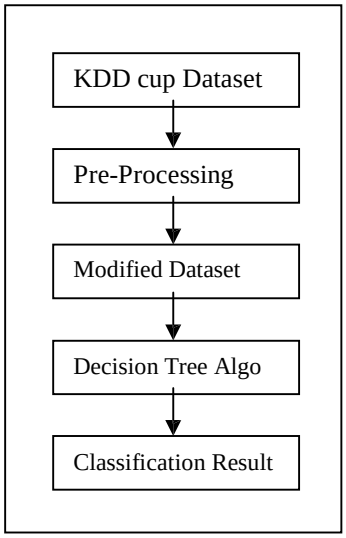
Attack Class	Attack Type
DoS	apache2, back, land, mailbomb, neptune , pod, processtable, smurf, teardrop,dpstrom
Probe	ipsweep,mscan,nmap, portsweep, saint
R2L	ftp_write, guess_passwd, imap, multihop, named, phf, sendmail, spy, snmpgetattack, snmpguess , war ezclient, warez master, worm, xlock, xsnoop
U2R	buffer_over flow, httptunnel, loadmodule, perl, ps, rootkit, sqlattack, xtern

Table1. Various types of Attack described in Four Major category.

V. PROPOSED APPROACH

A. System Architecture

Proposed Intrusion detection technique is represented in flowchart 1. Data preprocessing is done to convert the non-numeric value to numeric value [20].The information obtained by KDD Cup"99 can be a combination of many system calls. A system call is a text base record. Every text record in the database has 41 features as listed in table 2 of [16].



Flowchart 1 Proposed Decision tree Approach for Intrusion detection

B. Preprocessing Step to KDD Data

An integrated decision system that consists of three phases was proposed in this paper [3]:

Data Preprocessing Phase, Fusion Decision Phase and Data Callback Phase.

Data Preprocessing (DP) Phase, in order to reduce data as much as possible without any information loss, two data reduction strategies for IDS is performed efficiently. The first step, we call it DP1, is attribute selection, which extracts some attributes from relatively corresponding attributes. Different from dealing with attributes individually, techniques dealing with vectors using "ad hoc" techniques that incorporate correlation information with criteria tailored for scalar attributers were performed. The second step, we call it DP2, is sample reduction, which classifies two samples as one class and removes either of the two samples meanwhile if their similarity exceeds a fixed threshold. Distance measure is often used to assess dissimilarities based on the attribute values describing the samples. If part or all attributes of an attribute vector are represented by a string, we may replace these strings with some real number.

Fusion Decision Phase, in order to filter false rates and improves detection rates, a dynamic and a fusion classification technology is designed and performed. Data Callback Phase, some undetermined samples need to be updated to the testing date pool. This strategy ensures the availability of our performance. Our experiment demonstrated that our integrated decision system is availability. This strategy is useful for our integrated decision system to discover suspicious or intrusion.

C. Analysis of Preprocceing Step

Intrusion detection system requires carefully planning, preparation, prototyping, testing, and Specialized training [3]. Most researches are conceptual deficient and mutually Independent. Some researches provide sufficient data formats, and some others are conspicuous on analysis.

Each intrusion pattern analysis method has its own limitation, so an integration of these methods seems better than redesigning a new analysis framework. The main purposes of our system analysis are to filter false rates and improve detection rates. So, some issues are derived from these purposes:

(1) How to provide an optimal and efficient computing data for IDS.

(2) How to filter false rates and improve detection rates.

(3) How to discover attack patterns and display appropriate data types for administrators to make policies.

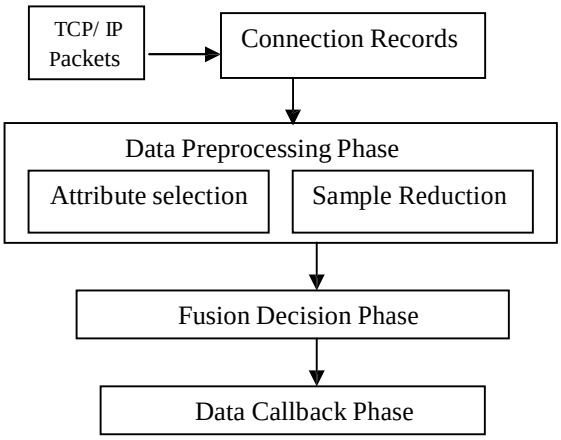


Figure1. Framework of Integrated Decision System [3]

In connection records stage, the term "connection" refers to a sequence of data packets related to a particular service, e.g., the transfer of a web page via the HTTP protocol

VI. Conclusion and Future Work

In this paper, we propose preprocessing step to improve decision tree algorithm, which is classified in to the phase, data preprocessing phase, fusion decision phase and data call back phase. These strategies ensure the availability of our performance in terms of accuracy, error rate and false positive rate. This paper uses Stratified weighted sampling techniques to generate the samples from the original datasets and then apply the improved decision tree algorithm which overcomes the limitations of the ID3 algorithm. Hence the proposed method can be implemented for various datasets where size of data is large and result are very accurate with less Error rate than existy algorithm. When compared to two random samples.

For Future work, It can be reduce the number of attribute of KDD Data Set. Attribute Reduction concept take less time for classification.

Acknowledgement

The authors sincerely thank the anonymous reviewers whose comments have greatly helped clarify and improve this paper.

References

[1] Huang Ming, Niu Wenying, Liang Xu".An improved decision tree classification algorithm based on ID3" and the application in score analysis.
[2] Mahesh V. Joshi, George Karypis, Vipin Kumar Scal ParC: "A New Scalable and Efficient Parallel Classification Algorithm for Mining Large Datasets" Department of Computer Science University of Minnesota, Minneapolis.
[3] Wang Ling, Xiao Haijun"An Integrated Decision System for Intrusion Detection " 2009 International conference on Multimedia Information Networking and Security.
[4] Wei Wang, SylvainGombault, Thomas guyet "Towards fast Detecting Intrusion: Using key attribute of network traffic" 3rd International conference on Internet monitoring & Protection.
[5] W.Lee, S. Stalfo, K. Mok,"A Datamining Frame-work for building Intrusion Detection model" IEEE Symposium on Security & Privacy.
[6] E. Eskin, A. Arnold, M. Prarun , L. Portnoy,
S. Stalfo"Geometric frame work for Unsupervised anomaly detection "Application of data Mining in computer Security,2002.
[7] UM Fayyad, G Piatetshy-Shapiro, P Smyth, and R Uthurusamy, Advances in Knowledge Discovery and Data Mining AAAI/MIT Press, 1996
[8]D.Denning, "An Intrusion Detection Model", IEEE Transaction on Software Engineering, 13(2), 1987.pp.222-232.
[9].M.Adams and D.J.Hand, "Improving the practice of the classifiers performance assessment", Neural Computation, vol.12, issue.2, MITPress, 2000, pp.305-312.
[10] D. Moore, C.Shannon, "Code-red: A Case Study on The Spread And Victims of An Internet Worm", Proceedings of the 2002 ACM SICGOMM internet measurement workshop, Marseille, France, November, 2002, pp. 273–84.
[11] A. Lazarevic, A. Ozgur, L. Ertoz, J. Srivastava, V. Kumar A Comparative Study of Anomaly Detection Schemes in Network Intrusion Detection.
[12]Haijun Xio, Fan Hong, Zhaoli Zhang, Junguo lio"Application of PSVM and data processing for intrusion detection" journal dynamics of Continuous, discrete and impulsive system,2007.
[13]Juan Wang, Qiren Yang, Dasen Ren"An Intrusion Detection algorithm based on decision tree technology" 2009 Asia pacific conference on information processing.
[14] T. Jyothirmayi, Suresh Reddy "An algorithm of Better decision tree" International journal of computer science and Engineering.
[15]Mohammadreza Ektefa, Sara Memar, Fatimah sidi , Lilly Suriani Affendey" Intrusion Detection using Data Mining Technique"IEEE 2010.
[16] R. Shanmugavadivu, Dr. N. Nagrajan "Network Intrusion Detection System Using fuzzy Logic" Indian Journal of computer Science and Engineering.
[17]http://kdd.ics.uci.edu/database/kddcup99/kddcup99.html.
[18]Prof. Punam V. Khandar, Prof. Sugandha V. Dani "Knowledge Discovery and Sampling Techniques with Data Mining for Identifying Trends in Dataset" IJCSE 2010 Special Issue.
[19] Saar-Tsechansky, M. and F. Provost. "Active Sampling for Class Probability Estimation and Ranking." Machine Learning 54:2 2004.
[20]Shailendra k. Shrivastav, Preeti Jain" Effective Anomaly based Intrusion Detection using Rough Set Theory and Support Vector Machine" IJCSE Vol. 18, No-3, 2011