Information based feature selection for intrusion detection systems

M.S Irfan Ahmed, Riyad A.M, R.L Raheemaa Khan, Mohamed Jamshad K, Shamsudeen E

Abstract— Intrusion detection is the act of detecting harm full attacks towards computer systems and networks. Intrusion detection systems are used to identify and report those intrusions in order to take necessary remedial actions. We use can use real time network data collected from various sensors or devices for analyzing the traffic. Also there are various standardized data sets which contain various types of intrusions such as DARPA and KDD Cup99 data sets. We use NSL-KDD dataset for our analysis as NSL-KDD is an improved version of KDD Cup99 data set. Here, we evaluate the intrusion detection system by with and without applying data pre-processing and feature selection techniques. Thus we analyze the importance of data pre-processing and feature selection applied on the data set using rough sets and information gain.

Index Terms— intrusion detection system; information gain; NSL-KDD; feature selection; RST.

1 INTRODUCTION

Intrusions can be found by tracing the anomalous network activities. Different methods are devised to identify intrusions. We can broadly classify them into misuse detection and anomaly detection. Misuse detection techniques trace the abnormalities through matching anomalous signatures of previous attacks with current patterns. This is more similar to antivirus logic. On the other hand, Anomaly detection techniques catch hold on all activities other than normal ones. Hence, here the normal profiles are well identified first to observe deviations of current activities if any. It is the responsibility of this technique to make sure that the deviations identified is well enough to label as 'intrusion'.

The objective of this paper is to test the impact of data preprocessing and feature selection on the NSL-KDD data set used for classifying intrusions. Here we use rough sets and information gain for feature selection. After applying feature selection, we use J48 classifier to evaluate the performance.

Other sections of this paper are categorized as follows. Section 2 discusses the related works. Section 3 discusses the proposed model for dataset preprocessing. Section 4 discusses the preprocessing of NSL-KDD data set. Section 5 discusses the feature selection of data set. Section 6 discusses the experimental platform. Section 7 discusses performance evaluation. Conclusion of the paper is discussed in the Section 8. Section 9 lists the references.

2 RELATED WORKS

As interest in intrusions increases so does the intrusion detection. The concept of IDS was first proposed by James P Anderson in his technical report in 1980[1]. He introduced the notion that audit trail calls, application logs, file system modifications and other host activities related to the machine. After seven years, Dr. Dorothy Denning published a model which revealed the necessary information for commercial IDS development [2].

Data preprocessing forms a critical part of anomaly based network intrusion detection. Data preprocessing takes a significant amount of effort, and directly impacts on the accuracy and capability of the downstream algorithm according to Lee and Stolfo [3]. Data preprocessing forms a critical part of anomaly based network intrusion detection. According to Lei Yu, Jieping Ye and Huan Liu [4], Dimensionality reduction is an effective approach to downsizing the data. It is a methodology that attempts to project a set of high dimensional vectors to a lower dimensionality space while retaining metrics among them. The machine learning and data mining techniques may not be effective for high-dimensional data because of the curse of dimensionality and efficiency will degrade rapidly as the dimension increases.

Jashan [5] proposed hybrid model for developing the intrusion detection system by combining C4.5 decision tree and Support Vector Machine (SVM) approaches. They collected data set from KDD"99 cup. Reduce the dimensionality of a data set of network traffic is used by selecting features methods. They selected 12 attributes between 41 attributes. They applied the hybrid model to detect normal and abnormal classes. A comparative analysis between single approaches and hybrid approaches are presented. They shared that the hybrid approach gives high accuracy and less time to detect intrusion. Ghanshyam [6] proposed increment SVM with RST approaches to detect intrusion. The authors collected data set from KDD"99 cup. The selection of significant attributes from the network traffic dataset is applied by RST method. Analysis dataset, which already used for training and testing, is used by the increment SVM method. From data analysis, a comparison between incremental SVM and non-incremental SVM is presented. According to the authors, the incremental SVM approach increased performance for intrusion detection. Xin Du [7] used K-means clustering data mining technique for intrusion detection. They collected data set from the MIB network with a period. The authors applied information entropy to select most significant attributes from the entire set of dataset. They selected four attributes destination IP, source IP, destination port, and source port. They in-

serted various types of attacks into the network with interval time. They used K-means clustering data mining to determine normal and abnormal class.

3 PROPOSED MODELFOR DATA SET PRE-PROCESSING

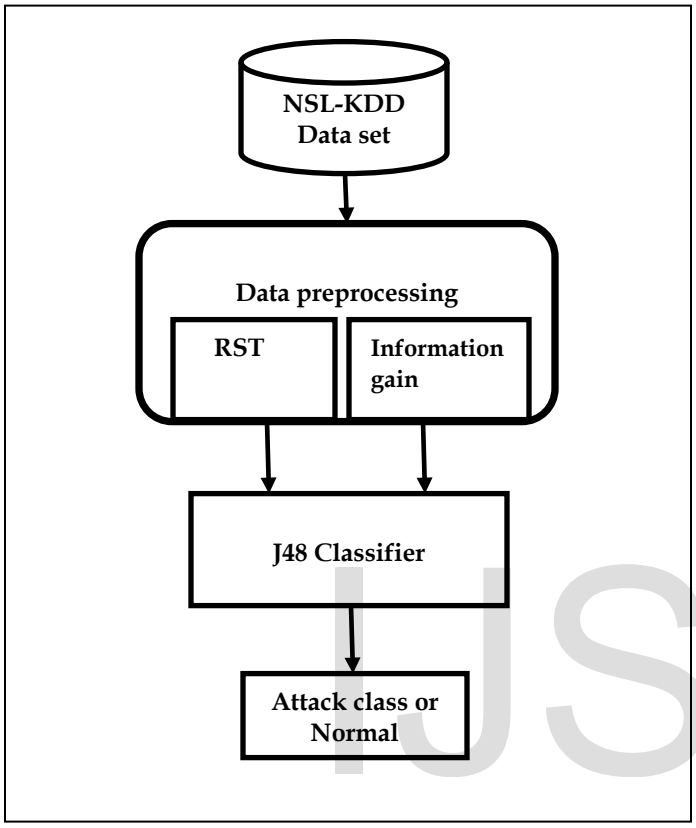

Fig 1.

We use the efficient NSL-KDD data set. The NSL-KDD data set suggested to solve some of the inherent problems of the KDDCUP'99 data set. KDDCUP'99 is the mostly widely used data set for anomaly detection. But Tavallaee et al conducted a statistical analysis on this data set and found two important issues that greatly affected the performance of evaluated systems, and results in a very poor evaluation of anomaly detection approaches. To solve these issues, they proposed a new data set, NSL-KDD, which consists of selected records of the complete KDD data set [8].

The dataset contains 41 features labeled as normal or specific attack type. These 41 features grouped into 4 categories: basic features, content features, time-based traffic features and host-based traffic features. The value of these features mainly based on continuous, discrete and symbolic value. All the feature categories are discussed below:

• Basic features: Among 41 features of NSL-KDD dataset, the features which derived from TCP/IP connection are known as basic features. These features lead an implicit delay in detection (example: duration, protocol_type, service etc).

• Content features: Some features use domain knowledge to access the payload of the original TCP packets are known as content features (example: logged_in, root_shell, is_hot_login etc).

• Time-based traffic features: Time-based traffic features are specially designed to capture one special property of the data-set that is to capture those features which are mature over a 2 second temporal window (example: srv_serror_rate, srv_rerror_rate).

• Host-based traffic features: Among four types of attacks in NSL-KDD dataset, few attacks span longer than 2 seconds intervals. Host-based traffic features are designed to access all attacks which span longer than 2 second intervals that have the same destination host as the current connection (example: serror_rate, rerror_rate).

NSL-KDD dataset mainly contain four types of attacks. They are given below:

• Denial of Service attack (DoS): When an attacker successfully makes computing and memory resources too busy or denies legitimate users access, to a machine is called DoS attack.

• Remote to Local attack (R2L): In R2L attack, an attacker wants to gain the local access as a user of particular machine without any account. To accomplish this, attacker sends packet to a remote machine over a network and exploits some vulnerability.

• User to Root attack (U2R): By using these types of attack, attacker can access normal user account on the system and gain the root access to the system and exploit some vulnerability.

• Probe: In this type of attack, by scanning a network of computer, attacker can gather all necessary information about the target system or can find out known vulnerabilities.

The following are the advantages of the NSL-KDD over the original KDD data set:

First, it does not include redundant records in the train set, so the classifiers will not be biased towards more frequent records.

Second, the number of selected records from each difficulty level group is inversely proportional to the percentage of records in the original KDD data set. As a result, the classification rates of distinct machine learning methods vary in a wider range, which makes it more efficient to have an accurate evaluation of different learning techniques.

Third, the numbers of records in the train and test sets is reasonable, which makes it affordable to run the experiments on the complete set without the need to randomly select a small portion. Consequently, evaluation results of different research works will be consistent and comparable.

4 PRE-PROCESSING OF NSL-KDD DATA SET

Pre-processing of original NSL-KDD dataset is necessary to make it as a suitable input for SVM. Data set pre-processing can be achieved by applying: i. Data set transformation, ii. Data set normalization and iii. Data set discretization.

i) Data set transformation: The training dataset of NSL-KDD consist of approximately 4,900,000 single connection instances. Each connection instance contains 42 features in-

IJSER

cluding attacks or normal. From these labeled connection instances, we need to transform the nominal features to numeric values so as to make it suitable input for classification.

ii) Data set normalization: Dataset normalization is essential to enhance the performance of intrusion detection system when datasets are too large. Here, we have used min-Max method of normalization.

iii) Data set discretization: Dataset discretization technique is used for continuous features selection of intrusion detection and to create some homogeneity between values, which have different data types. Here, we have used range discretization technique for this purpose.

5 FEATURE SELECTION FROM DATA SET

NSL-KDD data set consists of 41 features. When we use all the features, there are possibilities of dimensionality curse. This may lead to incorrect classification and increase in execution time. So, we use two feature selection algorithms such as rough sets and information gain. The techniques are defined below.

5.1 Roughset Theory

Rough set theory (RST) was first proposed by Zdzisaw pawlak in 1982[9]. Rough set theory extends set theory. This theory is applied for making decisions on inconsistent, confused and vague concepts. It is a mathematical tool usually utilized for classification of vague information. Data relations can be easily found by using RST. It is well capable of reducing number of attributes defining an object, hence making feature reduction possible. It helps to find the best features that can represent a class. Three regions are formed by RST known as lower approximation, upper approximation and boundary region. The reduct in RST is the subset of feature set that can represent the whole feature set. Elements in lower approximation definitely belong to the set while elements in upper approximation are likely to be in the set.

The definitions for the rough set theory are given below.

Definition 1:

The information system is defined as S= <U, A, V, f> where U=$\{x_1, x_2, x_3, \ldots . x_p\}$ is the finite set of p objects. A is non-empty finite set of q attributes used for object description, A=$\{a_1, a_2, a_3, \ldots . a_q\}$. V= $\cup_{a \in A}\ V_a$ where $V_a$is the set of values of a^{th}feature. $f : UxA \rightarrow V$ is a total function such that $f(x, a) \in V_a$ for all $a \in A$ and $x \in U$.

If the features in S can be divide into conditions feature set C and decision feature set D ($A=C\ U\ D$ and $C \cap D = \Phi$) the information system can be called as a decision table, DT =<U, C U D, V, f>

Definition 2:

Every $B \subseteq A$ produce an equivalence relation known as indiscernibility relation on U given by $IND_A(B) = \{(x, , x') : \forall a \in B = (x')\}$ a reduct of A is the least $B \subseteq A$ that is equivalent to A up to imperceptibility. ie., $IND_A(B) = IND_A(A)$.

Rough set theory is applied to select significant attributes among 41 attributes from the data set. The rough set theory method was applied by using Rosetta toolkit. It identified eight significant features out of 41 features. All these features are listed in table1.

Feature number	Features
1	Duration
2	Services
3	Dst-bytes
4	Count
5	Same-srv-rate
6	Dst-host-srv count
7	Dst-host-diff-srv rate
8	Dst-host-some-svr-port-rate

Table 1

5.2 Information Gain

To use Information Gain for feature selection an entropy value of each attribute of the data has to be calculated[6]. The entropy value is used for ranking features that affect data classification. A feature which does not have much effect on the data classification has very small information gain and it can be ignored without affecting the detection accuracy of a classifier. First, by calculating feature IG for each specific attack type, we select one feature which plays most significant role in distinguishing this attack and collect all most relevant features for the whole 23 attack types and remove duplicate ones and redundant ones.

IG (C; X) = H(C) –H (C|X)
Where H(C) = -∑ P(C=ci) log2 P(C=ci)

Probability that the class attribute ci occurs, and
H (C|X) = -∑ P(X=xi) H (C|X=xi)
IG(C; X) is information gain of attribute X.

H(C) is entropy of C and H (C|X) is the average conditional entropy of C.

X defines individual input features in the training dataset, and C defines class.

5.3 Feature selection algorithm using information gain

Input: Original Dataset with n dimensions
Output: Reduced Dataset with r dimensions

1. begin
2. For i=1 to N do
3. Gain[i]=information[i];
4. Sort Gain[i] in ascending order
5. end;

We select the features ranked as per the highest information gain score. The dimensionality reduction process conducted using information gain reduces the feature size of the dataset from 42 to 8.

Feature number	Features
1.0529431	Diff_srv_rate
1.0453802	Dst-host-diff-srv-rate
1.0323339	Count
1.0189628	Dst_host_srv_count
1.0145247	Src_bytes
1.0092918	Same_srv_rate
1.0092918	Dst-host-same-diff-srv-rate
1.0092918	Flag

Table 2

5.4 J48 Classifier

For evaluating the performance of rough set and information gain based pre-processing tasks, we used the robust J48 algorithm.

C4.5 decision tree is the most popular tree classifier which is developed by Quinlan [10]. J48 is a Java implementation of C4.5 in Weka environment [11].

C4.5 builds decision trees from a set of training data using the concept of information entropy. The training data is a set $S = s_1, s_2, ...$ of already classified samples. Each sample s_i consists of a p-dimensional vector $(x_{1,i}, x_{2,i}, ..., x_{p,i})$, where x_j represent attributes.

Pseudocode is given below.

1. Check for base cases
2. For each attribute *a*
   i. Find the normalized information gain ratio from splitting on *a*
3. Let *a_best* be the attribute with the highest normalized information gain
4. Create a decision *node* that splits on *a_best*
5. Recur on the sublists obtained by splitting on *a_best*, and add those nodes as children of node [12].

6 EXPERIMENTAL PLATFORM

Experiments are done on windows OS platform with 2.8 GHz Intel core i5 processor and 4 GB RAM. The software used are Rosetta and Weka (Waikato Environment for Knowledge Analysis) 3.5.7 designed by machine learning group at University of Waikato. NSL-KDD with reduced features are used here.

The original data seT contains 18.3 MB data with 1, 25973 instances. In our experiment, we worked with three major attacks namely Neptune, snmpget and Satan. These attacks correspond to 96002 instances out of total 1, 25973 instances.

7 PERFORMANCE EVALUATION

Accuracy= TP+TN/ (TP+TN+FP+FN)
Detection rate = TP/ (TP+FN)
False alarm rate= FP/ (TN+FP)
Where TP = True positive, TN = True negative, FP = False positive, FN = False negative.

The true positives (TP) and true negatives (TN) are correct classifications. A false positive (FP) occurs when the outcome is incorrectly predicted as yes (or positive) when it is actually no (negative). A false negative (FN) occurs when the outcome is incorrectly predicted as negative when it is actually positive.

J48 classifier with full 41 features

Detection rate	**False positive rate**	**Accuracy**	**Duration**
97.58	3.6	96.94	301.2

Table 3

J48 classifier after applying rough set theory pre-processing

Detection rate	**False positive rate**	**Accuracy**	**Duration**
97.56	3.7	96.92	66.43

Table 4

J48 classifier after applying information gain pre-processing

Detection rate	**False positive rate**	**Accuracy**	**Duration**
96.99	4.24	96.22	0.19

Table 5

From the tables above, it can be noticed that there is tremendous decrease in the processing time with reduced feature set. Also, the results produced have good detection rate and accuracy, while keeping the false positives into minimum. It is also noted that, the RST preprocessing has slightly better results when compared to information gain. This encourages using rough set theory method for pre-processing in future systems for getting better results.

8 CONCLUSION

Data preprocessing before classifying detections will definitely improve the results in various dimensions. We used RST and information gain for preprocessing the NSL-KDD data set which reduced 41 features to 8 features. By applying J48 classifier with and without feature selection, we found that very good results are maintained even with less features. Also, processing time was reduced to a big extent.

IJSER

REFERENCES

[1] J. P. Anderson. "Computer security threat monitoring and surveillance", Technical report, James P. Anderson Company, Fort Washington, Pennsylvania, April 1980.

[2] Dorothy E. Denning. "An intrusion-detection model", IEEE Trans. Software Eng., 1987.

[3] Lee W, Stolfo S and Mok K, "A data mining framework for building intrusion detection model", In: Proc. of IEEE symposium on security and privacy , 1999.

[4]Lei Yu Binghamton University, Jieping Ye,Huan Liu ,Arizona State University, "Dimensionality Reduction for datamining-Techniques, Applications and Trends",2007

[5] J. Koshal, M. Bag " Cascading of C4.5 Decision Tree and Support Vector Machine for Rule Based Intrusion Detection System" Computer Network and Information Security, vol. 8, pp. 8-20, 2012.

[6] G. P.Dubey, N. Gupta, R. K Bhujade "A Novel Approach to Intrusion Detection System using Rough Set Theory and Incremental SVM" .International Journal of Soft Computing and Engineering IJSCE 2231-2307, vol. 1, 2011.

[7] X. Du, Y. Yang, X. Kang "Research of Applying Information Entropy and Clustering Technique on Network Traffic Analysis" IEEE 2008.

[8] M. Tavallaee, E. Bagheri, W. Lu, and A. Ghorbani, "A detailed analysis of the KDD CUP 99 data set", 2009 IEEE Int. Conf. Comput. Intell. Security Defense Appl.

[9] P.Ghosh, C. Debnath, D. Metia3, R. Dutt, "An Efficient Hybrid Multilevel Intrusion Detection SysteminCloud Environment", IOSR Journal of Computer Engineering (IOSR-JCE), vol 16, Issue 4, 2014.

[10] Quinlan, J, "C4.5: Programs for Machine Learning", Morgan Kaufmann, San Mateo (1993)

[11] Weka – Data Mining Machine Learning Software, http://www.cs.waikato.ac.nz/ml/weka/

[12] Http://en.wikipedia.org/wiki/C4.5_algorithm

IJSER