Feature Selection for Intrusion Detection using NSL-KDD

Hee-su Chae*, Byung-oh Jo*, Sang-Hyun Choi**1, Twae-kyung Park***
*Department of Information Security Management
**Department of Management Information System
Chungbuk National University in Korea
***SAMMI INFORMATION SYSTEMS CO., LTD.
52 Naesudong-ro, Heungdeok-gu, Cheongju Chungbuk 361-763
Korea
enigma0724@gmail.com, joecho86@gmail.com, chois@cbnu.ac.kr, tkpark@sammicomputer.co.kr

Abstract: These days, network traffic is increasing due to the increasing use of smart devices and the Internet. Amount of the intrusion detection studies focused on feature selection or reduction because some of the features are irrelevant and redundant which results lengthy detection process and degrades the performance of an intrusion detection system (IDS). The purpose of this study is to identify important selected input features in building IDS that is computationally efficient and effective. For this we evaluate the performance of standard feature selection methods; CFS(Correlation-based Feature Selection), IG(Information Gain) and GR(Gain Ratio). In this paper, we propose a new feature selection method using feature average of total and each classes. We apply one of the efficient classifier decision tree algorithm for evaluating feature reduction method. We compare between proposed method and other methods.

Key-Words: - Data Mining, Preprocessing, Feature selection, Feature Reduction, Intrusion detection system, NSL-KDD

1 Introduction

In recent year, due to the growing use of smart devices and the Internet, network traffic is rapidly increasing. A Cisco report found the following : "Global IP traffic in 2012 stands at 43.6 exabytes per month and will grow threefold by 2017, to reach 120.6 exabytes per month" [1].

Intrusions are defined as attempts or action to compromise the confidentiality, integrity or availability of computer or network [2]. Intrusion detection systems (IDSs) are software or hardware systems that automate the process of monitoring the events occurring in a computer system or network, analyzing them for signs of security problems [3].

Feature selection is the process of removing features from the original data set that are irrelevant with respect to the task that is to be performed. So not only the execution time of the classifier that processes the data reduces but also accuracy increases because irrelevant or redundant features can include noisy data affecting the classification accuracy negatively [4].

In this paper, we suggest a new feature selection method that uses the attributed average of total and each class data. The decision tree classifier will be evaluated with the NSL-KDD dataset to detect attacks on four attack categories: Dos, Probe, R2L, and U2R. Feature reduction is applied using three standard feature selection methods Correlation-based Feature Selection (CFS), Information Gain (IG), Gain Ratio (GR) and the proposed method. The decision tree classifier's results are computed for comparison of feature reduction methods to show that our proposed model is more efficient for network intrusion detection. The remainder of the paper is organized as follows: Section 2 give an overview of feature selection methods and NSL-KDD. The experimental study is discussed in section 3, and section 4 presents the result. Finally the paper is concludes with their future work in section 5.

2 Related work

2.1 Feature Selection

Feature selection is important to improving the efficiency of data mining algorithms. It is the process of selecting a subset of original features according to certain criteria, and is an important and frequently used technique in data mining for dimension reduction. Most of the data includes irrelevant,

[1] Corresponding author

redundant, or noisy features. Feature selection reduces the number of features, removes irrelevant, redundant, or noisy features, and brings about palpable effects on applications: speeding up a data mining algorithm, improving learning accuracy, and leading to better model comprehensibility [5].

There are two common approaches to feature reduction : a wrapper uses the intended learning algorithm itself to evaluate the usefulness of features, while a filter evaluates features according to heuristics based on general characteristics of the data. The wrapper approach is generally considered to produce better feature subsets but runs much more slowly than a filter [6].

In introduction, we explain that network traffic data is increasing rapidly. In order to detect intrusion from large traffic data, not only detection algorithm, but also feature selection method have to more efficient. These three feature selection methods use a complex calculation. For this reason, these methods is inefficient for amount of large data. In this paper, we propose simple and efficient feature selection method.

2.2 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 [7].

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.

The NSL-KDD data includes 41 features and 5 classes that are normal and 4 types of attacks: Dos, Probe, R2L, and U2R. Denial of Service Attack (DoS) is an attack in which the attacker makes some computing or memory resource too busy or too full to handle legitimate requests, or denies legitimate users access to a machine. Probing Attack is an attempt to gather information about a network of computers for the apparent purpose of circumventing its security controls. User to Root Attack (U2R) is a class of exploit in which the attacker starts out with access to a normal user account on the system (perhaps gained by sniffing passwords, a dictionary attack, or social engineering) and is able to exploit some vulnerability to gain root access to the system. Remote to Local Attack (R2L) occurs when an attacker who has the ability to send packets to a machine over a network but who does not have an

Table 1 Type of features

Type	Features
Nominal	Protocol_type(2), Service(3), Flag(4)
Binary	Land(7), logged_in(12), root_shell(14), su_attempted(15), is_host_login(21),, is_guest_login(22)
Numeric	Duration(1), src_bytes(5), dst_bytes(6), wrong_fragment(8), urgent(9), hot(10), num_failed_logins(11), num_compromised(13), num_root(16), num_file_creations(17), num_shells(18), num_access_files(19), num_outbound_cmds(20), count(23) srv_count(24), serror_rate(25), srv_serror_rate(26), rerror_rate(27), srv_rerror_rate(28), same_srv_rate(29) diff_srv_rate(30), srv_diff_host_rate(31), dst_host_count(32), dst_host_srv_count(33), dst_host_same_srv_rate(34), dst_host_diff_srv_rate(35), dst_host_same_src_port_rate(36), dst_host_srv_diff_host_rate(37), dst_host_serror_rate(38), dst_host_srv_serror_rate(39), dst_host_rerror_rate(40), dst_host_srv_rerror_rate(41)

account on that machine exploits some vulnerability to gain local access as a user of that machine.

41 attributes are consisted three features: Basic features, Content features, and Traffic features. Table 1 shows Features name and type of features.

3 Experimental Study

We explained above that network traffic data is increasing rapidly. In order to detect intrusion from large traffic data, detection algorithm, and feature selection method have to more efficient. The above three feature selection methods use a complex calculation. For this reason, these methods is inefficient for large scale data. In this paper, we propose a simple and efficient feature selection method.

3.1 Descriptive Statistics of NSL-KDD

NSL-KDD data has three features types : Numeric, Nominal, and Binary. Features 2, 3, and 4 are nominal, features 7, 12, 14, 15, 21, and 22 are binary, and the rest of the features are numeric type. Table 3 shows the average of feature 23 which is numeric type. The total average is bigger than normal, R2L, and U2R average and less than the Dos and Probe average. Table 3 shows frequency of feature 12 for each class and total. Feature 12 is binary type consisting of 0 and 1.

We propose a feature selection method using the Attribute Ratio (AR). AR is calculated by mean and frequency of features.

Table 2 The average of feature 23 in total and each classes data

Class	Mean
Total	0.16459408
Dos	0.348512787
Normal	0.044066495
Probe	0.150787218
R2L	0.002539183
U2R	0.011365423

Table 3 Frequency of feature 12

	Dos	Normal	Probe	R2L	U2R	Total
0	44970	19486	11573	86	6	76121
1	957	47857	83	909	46	49852

3.2 Proposed Method

In section 5, we explain NSL-KDD data which has three attribute types. We use attribute average and frequency for each class calculate the AR from numeric and binary type. AR can be calculated as :

$$\mathbf{AR(i) = MAX\big(CR(j)\big)} \tag{7}$$

Class Ratio (CR) is attribute is ratio of each class for Attribute i. CR is calculated by two methods according to the type of attributes. CR can be calculated as for numeric:

$$\mathrm{CR(j)} = \frac{\mathrm{AVG(C(j))}}{\mathrm{AVG(total)}} \tag{8}$$

CR can be calculated as for binary.

$$\mathrm{CR(j)} = \frac{\mathrm{Frequency(1)}}{\mathrm{Frequency(0)}} \tag{9}$$

After calculating AR(i), Features rank ordering larger AR. Table 4 shows the rank of features with a calculated AR. We did not use nominal type features to calculate AR.

3.3 Experimental Setup

We used WEKA 3.7 a machine learning tool [9], to compute the feature selection subsets for CFS, IG, and GR, and to evaluate the classification performance on each of these feature sets. We chose the J48 decision tree classifier [3] with full training set and 10-fold cross validation for the testing purposes. In 10-fold cross-validation, the available data is randomly divided into 10 disjoint subsets of approximately equal size. One of the subsets is then used as the test set and the remaining nine sets are used for building the classifier. The test set is then used to estimate the accuracy, and the accuracy estimate is the mean of the estimates for each of the classifiers. Cross-validation has been tested extensively and has generally been found to work well when sufficient data is available[8].

4 Results

We used the three standards and our proposed method for feature selection. The feature selection was performed on 41 features ; we used selected features and all nominal features.

To evaluate the results of the classifier, we used accuracy.

$$\text{Accuracy} = \frac{(\text{TP+TN})}{(\text{TP+FN+FP+TN})} \times 100 \qquad (10)$$

We calculated the accuracy for the accumulation of the number of features using the AR ranker and the accuracy of AR, CFS, IG, and GR for the accumulation of the number of features and Full data. The result show the inverse correlation between accuracy and AR up to 22 features. It is clear that the highest accuracy is 99.794% at 22 features. The accuracy of full data is 99.763%. The highest CFS accuracy was 99.781% with 25 features, IG was 99.781% with 23 features, and GR was 99.794% with 19 features.

5 Conclusion and Future Work

In this paper, we have proposed feature selection methods using AR and compared it with three feature selectors CFS, IG, and GR.

The experiment shows that between accuracy and AR value is inverse correlation in our feature selection method and the highest accuracy is 99.794% using 22 features. The accuracy of our method is higher than the accuracy of full data and is also as highly as accuracy of other methods. Future work will include a comparison of calculation time for our method and other methods. Also. we will calculate the True Positive Rate(TPR), False Positive Rate(FPR), and accuracy for each attack type.

Acknowledgement

This research was supported by the MSIP (Ministry of Science, ICT & Future Planning), Korea, under the "Employment Contract based Master's Degree Program for Information Security" supervised by the KISA (Korea Internet Security Agency) (H2101-13-1001). This research was supported by the MSIP(Ministry of Science, ICT and Future Planning), Korea, under the ITRC(Information Technology Research Center)) support program (NIPA-2013-H0301-13-4009) supervised by the NIPA(National IT Industry Promotion Agency)

References:

[1] Cisco, *Cisco Visual Networking Index: Forecast and Methodology*, 2012-2017, Cisco, 2013.

[2] Shilpa lakhina, Sini Joseph, Bhupendra verma, Feature Reduction using Principal Component Analysis for Effective Anomaly–Based Intrusion Detection on NSL-KDD, *International Journal of Engineering Science and Technology*, Vol. 2(6), 2003, pp.1790-1799.

[3] R.Bace and P. Mell, *NIST Special Publication on Intrusion Detection Systems*, 2001.

[4] Y Yang, JO Pedersen, A comparative study on the effect of feature selection on classification accuracy, *Procedia Technology 1*, 2012, pp.323 – 327.

[5] Liu H ,Setiono R, Motoda H, Zhao Z, Feature Selection: An Ever Evolving Frontier in Data Mining, *JMLR: Workshop and Conference Proceedings 10*, 2010, pp. 4-13

[6] Y. Kim, W. N. Street, F. Menczer, and G. J. Russell, *"Feature selection in data mining" in Data Mining*: Opportunities and Challenges: J. Wang, Ed. Hershey, PA: Idea Group Publishing, 2003, pp. 80–105.

[7] 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.*, 2009, pp. 53–58.

[8] Saurabh Mukherjeea, Neelam Sharmaa, Intrusion Detection using Naive Bayes Classifier with Feature Reduction, *Procedia Technology 4*, 2012, pp.119-128.

[9] http://www.cs.waikato.ac.nz/~ml/weka/