Performance Comparison of Traffic Classification Techniques for Detecting Malicious Network Traffic

Jintae Choi+, Sinh-Ngoc Nguyen+, Jeongnyeo Kim*, Guee-Sang Lee+, Kyungbaek Kim+

+Dept. Electronics and Computer Engineering,
Chonnam National University,
Gwangju City, Republic of Korea
* The Electronics and Telecommunications Research Institute,
Daejeon City, Republic of Korea
jefron1100@gmail.com+, sinhngoc.nguyen@gmail.com+, jnkim@etri.re.kr*, gslee@jnu.ac.kr+, kyungbaekkim@jnu.ac.kr+

Abstract

The quality of internet services is damaged seriously by malicious network traffics. Detecting these threats has attracted a lot of attentions. Among these, one of the most potential techniques is machine learning, which the need for training and testing data is very necessary. Recently, KDD network dataset has been used in a lot of classification researches. The researchers tried different methods, with different groups of attributes to train and test, in order to classify the malicious traffics with high performance. However, to the best of our knowledge, there is no study regarding the comparisons of performance between various of classification methods using KDD dataset. In this paper we surveyed about them, which includes SVM, KNN and Naïve Bayes. Moreover, with each technique, we make evaluations with different groups of attributes to find which is the best one. Experimental results show that KNN classification technique is the best one for detecting malicious network traffic.

Keywords-Intrusion Detection; KDD; Attribute groups; SVM; KNN; Naive Bayes; Detection Rate

I. Introduction

Due to the rapid growth of the internet network services, it is very convenient to use internet network service. However, these internet networks are constantly being exploited by malwares. They make it impossible to use internet service and to guarantee the quality of network services. To solve this problem, the IDS is being introduced and used to detect these malicious traffics on the network. However, most of the current IDS is signature-based detection, which detects malicious traffic by collecting and analyzing them directly from the network traffic by experts experience. This is a good way to detect malware but requires a lot of time and human efforts to analyze. Besides, it is difficult to detect new malicious traffics. For this reason, many studies have recently emerged using data mining or machine learning tools for detecting outliers.

Dong-hyuk et al. [1] used k-means clustering to detect DDoS attacks and Worm Attack traffics. For this purpose, they experiment by using various combinations of a parameters. For selecting effective parameter(e.g., Different Src IP & Same Dst IP pair Number, Total Packet Byte, Total Packet Number) of best performance [1]. Han et al. [2] detected the malicious traffics by using X-means clustering technique, which can optimize a number of clusters by itself. The traffic was analyzed for the statistical characteristics of the traffic and defined metrics for clustering [2]. Aggarwal et al. [3] compared detecting rate of a combination of attribute classes of malicious traffic on 15 attribute classes. Through this, they found the good combination of attribute classes by using random tree techniques [3]. Mohammad Khubeb Siddiqui et al. did an analysis of 10% of KDD cup'99 training dataset based on intrusion detection. They focused on establishing a relationship between the attack types and the protocols used by the hackers by using clustered data. Analysis of data is performed using k-means clustering. The investigation revealed many interesting results about the protocols and attack types preferred by the hackers for intruding the networks [4].

In this paper, we evaluated techniques performance by using SVM, KNN, and Naïve Bayes for comparing performance about each technique and each attribute group. For this work, we applied these techniques to KDD data set. Through experiments, we found that KNN technique is the most efficient technique among other techniques. In each group, group B(Basic features of individual TCP connections) and H(attributes that designed to assess attacks which last for more than two seconds) show good detecting rate. And group B doesn't detect the probe attack traffics. Also, we found the problem with these techniques. So, we purpose that to solving the problem and new method for detecting the malicious traffics.

II. Background

A. SVM

SVM was on the statistical learning theory, which analyzes the data use for classifiction. This method is made based on the features of data, to find the largest margin between the objects. In our research, the dataset includes many different types of attribute, it is hard to evaluate all of them at a time. However, by spliting this dataset into several groups of attributes and using SVM to evaluate the performance of detection, we can see the effect of each group to the performance of detection. [5].

B. KNN

K-nearest neighbors (KNN) technique assume that if a record has similar features with another one, they would be in the same class. The model is trained with some training samples, then testing data is used, which is classified into the class of k nearest record. When k=1, we can find nearest traning data, and we can label the testing records with the same class of the nearest traning data class. If k=3, we can choose class that the most number of class in nearest three tranning data.

C. Naïve Bayes

In machine learning, naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features, which is not like the situations in the real world. However, assumptions about this can still be made. Consequently, an exponential number of training data is not needed because all features are supposed to be independent and identically distributed.

D. *KDD Cup 1999 Data*

KDD Cup 1999 data is the data set used for The Third International Knowledge Discovery and Data Mining Tools Competition. In this data set, there are “bad” connections, called intrusions or attacks, and “good” connections, which are normal. Also, this dataset contains a wide variety of intrusions simulated in a military network environment. Lincoln Labs set up an environment to acquire nine weeks of raw TCP tcpdump data for a local-area network (LAN) simulating a typical U.S. Air Force LAN. They operated the LAN as if it were a true Air Force environment. The raw training data was about four gigabytes of compressed binary tcpdump data from network traffic collected in seven weeks. There were about five million collected connection records. Similarly, the test data collected in two weeks has around two million connection records. A connection is a sequence of TCP packets starting and ending at some defined times, from a source IP address to a target IP address under some well-defined protocol. Each connection is labeled as normal or attack, with exactly one specific attack type. Each connection record consists of about 100 bytes. There are four categories of attack categories as shown below [6].

Attacks fall into four main categories:

- DoS: denial-of-service, e.g. syn flood
- R2L: unauthorized access from a remote machine, e.g. guessing password
- U2R: unauthorized access to local superuser (root) privileges, e.g., various ``buffer overflow" attacks
- probing: surveillance and another probing, e.g., port scanning.

Also, there are 42 attributes in KDD dataset, which are grouped into 4 groups. Table I below shows the distribution of attributes by each group. Description of each group is as follow [6].

- Group B: Basic features of individual TCP connections. It has 9 attributes.
- Group C: Content features within a connection suggested by domain knowledge. It has 13 attributes.
- Group T: Traffic features computed using a two-second time window. It has 9 attributes.
- Group H: attributes that designed to assess attacks which last for more than two seconds. It has 10 attributes.

Table II. KDD Dataset Attack Traffic Components

Class	Record Count
normal	558,227
dos	419,436
r2l	1,125
probe	21,188
u2r	24

Table I. Attributes for Each Group of KDD Cup 1999 Data

Group	Attribute Name	Group	Attribute Name	Group	Attribute Name	Group	Attribute Name
B	duration	C	hot	T	count	H	dst_host_count
B	protocol_type	C	num_failed_logins	T	serror_rate	H	dst_host_srv_count
B	service	C	logged_in	T	rerror_rate	H	dst_host_same_srv_rate
B	src_bytes	C	num_compromised	T	same_srv_rate	H	dst_host_diff_srv_rate
B	dst_bytes	C	root_shell	T	diff_srv_rate	H	dst_host_same_src_port_rate
B	flag	C	su_attempted	T	srv_count	H	dst_host_srv_diff_host_rate
B	land	C	num_root	T	srv_serror_rate	H	dst_host_serror_rate
B	wrong_fragment	C	num_file_creations	T	srv_rerror_rate	H	dst_host_srv_serror_rate
B	urgent	C	num_shells	T	srv_diff_host_rate	H	dst_host_rerror_rate
		C	num_access_files			H	dst_host_srv_rerror_rate
		C	num_outbound_cmds			-	class
		C	is_hot_login				
		C	is_guest_login				

III. Machine Learning Based Malicious Traffic Detection on KDD Dataset

A. Dataset

In this paper, we used the KDD Cup 1999 data set for comparing each machine learning technique and each attribute group. KDD Cup dataset composed 4,898,432 records. But it is too for running on the weka (which is introduced below). So, we separated the KDD dataset from full KDD dataset to 1 million KDD dataset. Also, there is not enough attack data, so we got the attack data from other parts. The following Table II shows the number and percentage of records for each class. R2L and U2R attack traffics are not enough. However, They look like the real life data.

B. Weka

Weka is a collection of machine learning techniques for data mining tasks. It contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. Weka is an open source software issued under the GNU General Public License. The version used in this study is 3.8.1. This tool accepts the data file either in comma separated value (csv) or attribute-relation file format (arff) file format. And there is k-fold cross-validation functions. So, we used cross-validation because of we did use smaller data than all of KDD dataset [7].

C. Setting Parameters Used for Classifier

When we conduct experiments using the SVM technique, we used default value C in weka, which is used for choosing decision boundary shape. If C value is small, decision boundary look like linear. Also, If C value is big, we can get more detail curved line. C of default SVM parameter is 1. If we configure C parameter higher we can get the accurate result but this decision boundary is not linear. So we can select suitable value. For KNN, we used 1, 3 and 5 for k value. k should be chosen appropriately because if k is too small, the classifying result is not much trustful, and if k is too large, the performance is not good.

D. Metrics

Intrusion detection metrics help evaluate the performance of an intrusion detection system. We will use True Positive, False Positive, Precision, Recall, F-Measure as measuring tools. These metrics were derived from the metrics below.

- True Positive (TP): Number of instances correctly predicted attacks.
- False Positive (FP): Number of instances wrongly Predicted as attacks.
- True Negative (NT): Number of instances correctly predicted as non-attacks.
- False Negative (FN): Number of instances wrongly predicted as non-attacks.

IV. Discussion about the detection rates using different technique and training attribute groups

In each classification method, we employ 2 evaluation phases. The first phase is using different groups of attributes. And the second phase is using all attributes from all groups. The results are shown as follow.

Table III, IV and V below show the results of the experiment using each group of attributes and all attributes By using SVM technique. In SVM experiment, firstly, we choose different group of attributes to evaluate the efficiency. Among all kinds of attack, it is difficult to detect the r2l and u2r case becasue training data for them is not much. So SVM technique can't detect them. However, in the second evalutation, all attributes are chosen as features for training, both r2l and u2r attacks are detected with high rate, even if data is not adequate. Therefore, the more attributes are employed, the more effective SVM model is. The only drawback of using SVM with many chosen features is it takes long time for training, which is over a day.

Table VI, VII and VIII below show the results of the experiment using each group of attributes and all attributes By using Naïve Bayes technique. In Naïve Bayes, the detecting rates for r2l and u2r are not high in both of evaluations with different groups of attributes as well as all attributes. Although with other classes, the detected rate is quite high, it is still not high enough to apply to practical applications with low rate for detecting r2l and u2r. These rates can be improved if the training data for these 2 classes is created more.

Table III VI. SVM result that applies attributes of Group B

Class	TP	FP	Precision	Recall	F-Measure
normal	0.993	0.007	0.994	0.993	0.994
dos	0.995	0.003	0.995	0.995	0.995
r2l	0.021	0	0.421	0.021	0.041
probe	0.979	0.002	0.906	0.979	0.941
u2r	0	0	0	0	0

Table IV. SVM result that applies attributes of Group C

Class	TP	FP	Precision	Recall	F-Measure
normal	0.911	0.205	0.849	0.911	0.879
dos	0.797	0.115	0.834	0.797	0.815
r2l	0.03	0	0.667	0.03	0.058
probe	0	0	0	0	0
u2r	0	0	0	0	0

Table V. SVM Full attribute Result

Class	TP	FP	Precision	Recall	F-Measure
normal	0.999	0.004	0.997	0.999	0.998
dos	0.997	0	1	0.997	0.998
r2l	0.838	0	0.887	0.838	0.862
probe	0.994	0	0.994	0.994	0.994
u2r	0.667	0	0.727	0.667	0.696

Table VI. Naïve Bayes result that applies attributes of Group B

Class	TP	FP	Precision	Recall	F-Measure
normal	0.729	0.005	0.994	0.729	0.841
dos	0.995	0.255	0.739	0.995	0.848
r2l	0.085	0.001	0.103	0.85	0.093
probe	0.861	0.006	0.744	0.861	0.798
u2r	0.167	0	0.033	0.167	0.055

Table VII. Naïve Bayes result that applies attributes to Group T

Class	TP	FP	Precision	Recall	F-Measure
normal	0.945	0.029	0.976	0.945	0.96
dos	0.994	0.017	0.977	0.994	0.986
r2l	0.001	0	0.005	0.001	0.002
probe	0.494	0.021	0.341	0.494	0.395
u2r	0	0.002	0	0	0

Table VIII. Naïve Bayes Full attribute Result

Class	TP	FP	Precision	Recall	F-Measure
normal	0.921	0.005	0.996	0.921	0.957
dos	0.994	0.014	0.982	0.994	0.988
r2l	0.371	0.004	0.095	0.371	0.151
probe	0.946	0.024	0.462	0.946	0.621
u2r	0.883	0.011	0.002	0.833	0.004

Table IX. KNN (k=1) the result that applies attributes of Group B

Class	TP	FP	Precision	Recall	F-Measure
normal	0.999	0	1	0.999	1
dos	1	0.001	0.999	1	0.999
r2l	0.974	0	0.986	0.974	0.896
probe	0.986	0	0.983	0.986	0.893
u2r	0.333	0	0.381	0.333	0.381

Table X. KNN (k=1) result that applies attributes of Group H

Class	TP	FP	Precision	Recall	F-Measure
normal	0.999	0.002	0.998	0.999	0.999
dos	0.999	0.001	0.999	0.999	0.999
r2l	0.684	0	0.799	0.684	0.737
probe	0.992	0	0.994	0.992	0.993
u2r	0.25	0	0.429	0.25	0.316

Table XI. KNN (k=1) Full attribute Result

Class	TP	FP	Precision	Recall	F-Measure
normal	1	0	1	1	1
dos	1	0	1	1	1
r2l	0.962	0	0.983	0.962	0.972
probe	0.998	0	0.999	0.998	0.999
u2r	0.75	0	0.72	0.75	0.735

Table XII. KNN (k=3) Full attribute Result

Class	TP	FP	Precision	Recall	F-Measure
normal	1	0	1	1	1
dos	1	0	1	1	1
r2l	0.946	0	0.973	0.946	0.959
probe	0.996	0	0.999	0.996	0.998
u2r	0.708	0	0.773	0.708	0.739

Table XIII. KNN (k=5) Full attribute Result

Class	TP	FP	Precision	Recall	F-Measure
normal	1	0.001	1	1	1
dos	1	0	1	1	1
r2l	0.935	0	0.971	0.935	0.953
probe	0.995	0	0.999	0.995	0.997
u2r	0.667	0	0.842	0.667	0.744

In KNN evalutation, we make experiments with different k and attributes group, which is shown in Table IX, XI, XII and XIII As being seen from these tables, with k equals 1, 3, and 5, the high detecting rate could be achieved. Overall, this technique showed high F-measure rates for all traffics, which includes group B with not much attributes. It also showed high detection rates for r2l and u2r, which do not have much training data. The first reason for this is the characteristics of KNN, which does not require much features for training. Furthermore, the second reason is the similarity between testing data and training data, which could be exploited well in KNN. When all attributes are used for training, KNN outstands other methods with very high accurate detecting rate, which proves that KNN is the most suitable technique. However, this technique also has problems with training processes, which takes over 2 days.

V. Conclusion & Future work

This paper analyses the performance comparisons between the traffic classification techniques to detect network malicious traffic and the impact on the detection rate of each attribute group. For this purpose, 5-fold cross-validation was performed using one million records of the 5 million KDD Cup 1999 Data Set. The classification techniques included SVM, KNN, and Naïve Bayes. The experimental results showed that the extreme high rate of detection was very high, but the very slow processing rate was problematic. In the future, we plan to conduct a study that applies the corresponding classification techniques to distributed processing systems to improve these problems. Also, We plan to conduct a study of traffic classification techniques by using a CNN technique using deep learning techniques to normalize KDD Cup 1999 Data Set traffic records using a GPU.

Acknowledgments

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Science, ICT & Future Planning(NRF-2017R1A2B4012559). This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIP) (R0110-16-1001, Secure hardware containers technology to protect IoT devices from Denial of Service attack). This research was supported by Chonnam National University(Grant number: 2016-2884).

References

[1] Dong Hyuk Shin, Kwang Kue An, Sung Chune Choi and Hyoung-Kee Choi, "Malicious Traffic Detection Using K-means", *The Journal of The Korean Institute of Communication Sciences,* 41(2), pp. 277-284, Febrary 2016.

[2] Myoungji Han, Jihyuk Lim, Junyong Choi, Hyunjoon Kim, Jungjoo Seo, Cheol Yu, Sung-Ryul Kim and Kunsoo Park, "A Malicious Traffic Detection Method Using X-means Clustering", *Journal of KIISE,* 41(9), pp. 617-624, September 2014.

[3] Aggarwal, Preeti, and Sudhir Kumar Sharma. "Analysis of KDD dataset attributes-class wise for intrusion detection." *Procedia Computer Science,* 57, pp. 842-851, 2015.

[4] Siddiqui, Mohammad Khubeb, and Shams Naahid. "Analysis of KDD CUP 99 dataset using clustering based data mining." *International Journal of Database Theory and Application 6.5* (2013): 23-34.

[5] “Understanding Support Vector Machine technique from examples (along with code)” (2017, Nov.) [Online]. Available: https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/

[6] KDD Cup 1999 Data. (2017, Nov.) [Online]. Available: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

[7] “Waikato environment for knowledge analysis (weka) version 3.8.1.” Available: http://www.cs.waikato.ac.nz/ml/weka/, October 2017