An approach to detect TCP/IP based attack

Ugtakhbayar.N[#1], Usukhbayar.B[#2,] Nyamjav.J[#3]
#National University of Mongolia Ulaanbaatar, Mongolia

Abstract
Intrusion Detection Systems have become an indispensable in computer networking security of all network types such as wired, wireless. In the last years, the system needs to identify new intrusion in large datasets in a timely manner because internet to instantly access information at anytime from anywhere. That is a massive increasing of data traffic and internet nodes. Therefore, to refine IDS's performance and false alarm is a one of the important challenges in intrusion detection and prevention fields. In this work we propose an approach to detect TCP connection based attacks using some data mining algorithms. We gather raw network traffic and classify it into normal and abnormal traffic by Bro IDS and Backtrack security operation system. First, we extract features in TCP/IP headers of the packets such as sequence and acknowledge numbers, window size, control flags, and an event which is the time between neighbour segments from our collected traffic. Next, we evaluate the worth or merit of a feature in novel attacks and select valuable subset of features using Markov Blanket and Pearson correlation. Finally, we are training our machine with the KDD 99 dataset and the selected features are given to learn the classifiers: J-48, Naïve Bayes. By adopting the concepts of machine learning and data-mining, we could detect about 74% of novel attacks with 19 features.

Keywords
weka, data mining, learning algorithms, IDS, intrusion detection.

1. Introduction

Information security is still quickly developing in any information technology fields. In the last few years, due to the growing use of computer networks, network traffic immediately increases. There are several private as well as business sectors, government organizations that store valuable data over the computer network. Cause, new threats are showing up on quickly, while older often abide relevant. Therefore, more dynamic mechanisms such as Intrusion Detection Systems are should also be utilized. A Cisco report found the following: "Global IP traffic in 2012 stands at 43.6 exabyte per month and will grow threefold by 2017, to reach 120.6 exabyte per month, by 2019, there will be 24 billion networked devices and connections globally." [1].
Intrusion detection is the process of monitoring the events occurring in a computer system or network and analyzing them for signs of possible attacks [2, 4]. Intrusion detection mechanism is divided in two; anomaly detection and misuse detection. Misuse detection is an approach where each suspected attack is compared to a set of known attack signatures [3]. It is in an exclusive manner the attacks in that database that can be detected, this method does not can for detection of unknown attacks. Unknown attack can be most zero day attacks. The role of anomaly detection is the identification of data points, substance, event and observations or attacks that do not conform to the expected pattern of a give collection [5].
Network traffic speeds and volume are increasing at an exponential rate. The conventional approach of tuning the hardware and software of the NIDS platform to maximize its performance can yield considerable improvements, but falls short in supporting next-generation networks operating at gigabits per second and faster.
This paper worked for data mining as a data processing technique, it can increase detection ratio using data mining algorithms and decreasing processing time using feature selection methods. We are using KDD 99 dataset and our university gateway traffic in this research. The KDD 99 dataset has been used for evaluating the most eminent available in the literature for feature selection and classification. KDD 99 dataset consist of nearly 5 million training connection records labeled as an intrusion or not an intrusion, and separate testing dataset consists of seen and unseen attacks [6]. In Methodology section, we presented a Markov Blanket feature selection and J48, Naïve Bayes in our collected and KDD 99 datasets. In the finally section we are summarizing our paper and give final conclusion.

2. Related Work

In the last years, network security has been the subject of many researchers. There are many works in the literature that discuss about information security, Intrusion detection system, using artificial intelligence and data mining in intrusion detection system. Intrusion detection and prevention systems used to detect and prevent the known and unknown attacks made by intruders. Moradi and Zulkernine, who are publishers of [8]. In this paper, there are presented an IDS that uses IDS for effective intrusion detection. One of the disadvantage of their approach is that it increases the time in training. In the literature [9] et al proposed a new method based on Continuous random function for selecting appropriate feature sets to perform network intrusion detection. Also,

Manuscript received April 5, 2016
Manuscript revised April 20, 2016

Liu and Gu have used Learning Vector Quantisation neural networks to detect attacks, that is a supervised version of quantization, which can be used for pattern recognition, multi-class classification and data compression tasks [10]. In paper [11] has written a highly referenced article about intrusion detection using neural networks. In the article, he studies in detail the advantages and disadvantages of neural networks for this application. In the conclusion of this article that neural networks are very suitable for Intrusion detection system. The [12] have used a neural network to detect the number of zombies that have been involved in DDoS attacks. The objective of their work is to identify the relationship between the zombies and in sample entropy. Shon and Moon used a genetic algorithm to extract optimized information from raw internet packets [13]. Ruchi Jain and Nasser S. Abouzakhar applied J48 decision tree algorithm to determine significant features from the KDDCUP 1999 dataset for anomaly intrusion detection [14]. And experimental results demonstrate that the Hidden Markov model is able to classify network traffic with approximately 76% to 99%. Most proposed techniques utilize characteristics of network traffics to identify abnormalities absolutely. But, performing the real time network traffic detection with maintaining higher accuracy is restricted due to the complex nature of networks.

3. Methodology

This research focuses on solving the issues in Intrusion Detection methods that can help the network and system administrators to make pre-processing, classification of network traffic. Most of the attacks can be identified only after it happens. Data mining approaches have been implemented by many researchers to solve the abnormal detection problem. In this section, we are explaining the proposed methodology for anomaly intrusion detection. We concentrated on data mining such as J48 algorithm, Naïve Bayes classifiers, because data mining approaches use strong statistical foundations for enhancing the dynamic and accurate learning that gives better accuracy, reduce false alarm rates, performance improvements, ability to detect novelty, protection against zero-day exploits.

The entire framework of proposed methodology shown in figure 1, we are collecting our university's internet and intranet traffic using Bro IDS by a sensor. Our method consists of two stages. In stage 1, collecting data with real time intrusion detection analyzer with Bro IDS system. In stage 2, in the figure with red frame, feature is selected in KDD 99 dataset and train to data mining algorithms. In the prepossessing section, we clean duplicated dataset from KDD 99 dataset. After we are using the Markov blanked model [15, 16] and Pearson correlation for feature selection from the KDD 99 dataset in TCP/IP.

Markov blanket is a novel idea for significant feature selection in large dataset [16]. The formal definition of Markov blanket is: [17] the Markov blanket of a feature T, MB(T) of a BN. The set of parents, children, and parents of children of T. MB(T) is the minimal set of features conditioned on which all other features are independent of T. Shown as (1), the formal theory about Markov blanket.

$$P(T| MB(T), S) = P(T| MB(T)) \quad (1)$$

In our approach, the dataset is divided into training and testing datasets. Data for the research paper originated from two sources. First, training data sets include KDD 99's labeled datasets. The labeled datasets are applied to J-48, Naïve Bayes, classifier and the model are generated.

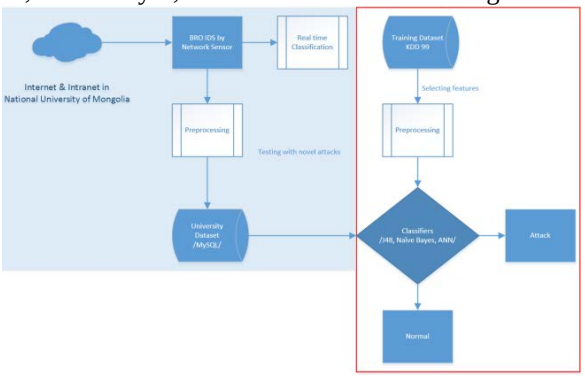

Fig. 1 The workflow of the proposed anomaly detection

The second set of data was from our collecting datasets from National University of Mongolia's gateway router. The router located inside of the firewall. Our sensor system Bro IDS is running with the specifications of 2nd generation Intel Atom Dual core processor, 2GB DDR3 RAM disk, 128GB SSD hard disk. Show as figure 2, the sensor location and our collector. We are using 4 sensors in the university network. Also, we are collected testing dataset with novel attacks using Backtrack system collected by Netflow, tcpdump. In the preprocessing section, our system is designed by applying feature extraction and feature selection. The dataset contains divers attack types that could be classified into four main categories. The dataset has 41 features for each connection record. The features divided into three categories that are host features, service features, traffic features.

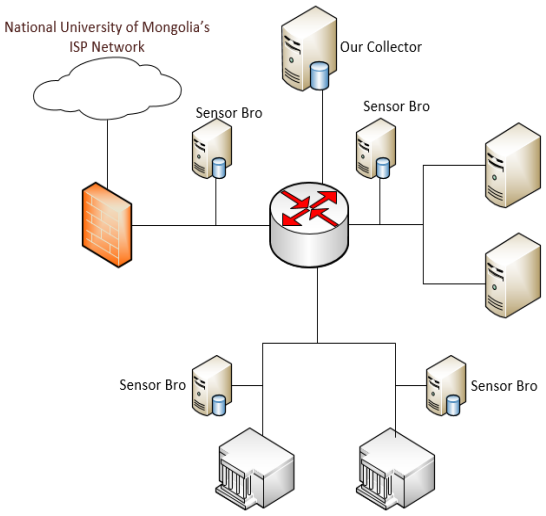

Fig. 2 Sensors and Collector network topology

Using the sample algorithm 1, we are removing the duplicated connection record before insert into database.

```
INPUT network_connections_with_features
IF (exist in database)
        ITERATION: next connection
ELSE
        INSERT into database
```

Algorithm. 1 The code for preventing duplicated data

KDD 99 dataset has been used in this research work of which 100% is treated as training data. We are implemented the proposed feature selection method in Matlab and classification in the Weka data mining tool. These tools are contained the tools required for the analysis and programmable.

Our computing environment for this paper included an Ubuntu operating system that ran on a Dell desktop. The system's hardware consists of an Intel I5 processor, 16GB of memory and 2TB hard disk space.

We found that 16 features of the dataset from the Markov blanket. These 16 features are "duration", "protocol-type", "service", "src_bytes", "land", "wrong_fragment", "num_failed_logins", "root_shell", "num_file_creations", "num_outbound_cmds", "is_guest_login", "srv_count", "serror_rate", "srv_serror_rate", "diff_srv_rate", "dst_host_count". But these 16 features are not shown good result. Therefore, we chose the collision of Markov blanket and Pearson correlation shown as figure 3. We found and remove every Fx using Markov blanket. And we used Pearson correlation in all Fx. We chose 3 features from pearson correlation from other 25 that are "num_file_creations", "count", "dst_host_srv_count".

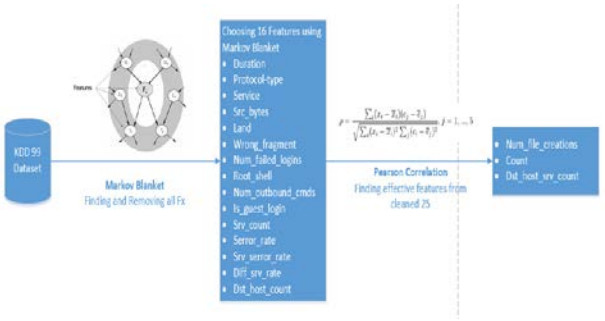

Fig. 3 Feature selection methodology

We calculated the classification accuracy and false alarm rate that are the percentage of the general number of correctly classified and percentage of the number of normal connections incorrectly classified.

Another important method is cross-validation that is a method for assessing how the results of a statistical analysis will generalize to an independent dataset. It is one way of measuring the accuracy of a learning method.

4. Result of Simulation

In this work we used 3 fold cross-validation. We have summarized our collected, labeled data and 10% of KDD 99. The summarized data is divided randomly into 3 parts. One partition is chosen for testing, while the remaining 2 are used for training. This is repeated 3 times so that each partition is used for only once. The accuracy of the feature selection and classifiers are given in table 1. The results show that, on classifying the dataset with all 41 features, the average accuracy rate of 73% and 70.3% is obtained for J-48 and Naïve Bayes. When our selected features are used, accuracy rate of J-48 and Naïve Bayes are increased significantly to 75.16% and 73.6%. In the table 2, on classifying the dataset with selected features, the average true positive detection rate of about 74% and 65% is obtained for J-48 and Naïve Bayes.

TABLE I the average of 3 times Accuracy with selected 19 feature dataset

Attack class	J-48		Naïve Bayes	
	Selected features	**All features**	**Selected features**	**All features**
Normal	76.4%	77.8%	56.1%	60.2%
Probe	89.2%	83%	93.6%	84.7%
DoS	98.2%	95.6%	95.1%	92%
U2R	45.7%	48.2%	33.2%	35.1$
R2L	66.3%	60.2%	90.4%	79.6%
Overall	**75.2%**	73%	**73.6%**	70.3%

Table 2 True positive with selected 19 feature dataset

Attack class	J-48	Naïve Bayes
Normal	75.1%	53.2%
Probe	88.3%	92.1%
DoS	98.9%	96.2%
U2R	45.2%	10.6%
R2L	63.8%	73.4%
Overall	74.3%	65%

5. Conclusions

In this paper, we present an intrusion detection system using J-48 and Naïve Bayes for classification and combination of Markov blanket and Pearson correlation for feature selection. To implement and classify of our system we used KDD 99 dataset and our University's traffic. The principal challenge in intrusion detection is to obtain high detection rate and reduce false alarm rate with novel attacks. In our experimental result shown as single classifier is not sufficient to obtain the high result and feature selection is the most important to detection ratio also showed that the effectiveness of J-48 is comparable to the Naïve Bayes. In the future, we will further improve feature selection with Markov blanket and Pearson correlation and investigate the use of new approaches as a classification algorithm in this area of intrusion detection.

Acknowledgment

We are thankful to Mongolian Foundation for Science and Technology for their support in the research.

References

[1] Cisco, Cisco Visual Networking Index: Forecast and Methodology, 2012-2017, 2013.
[2] IDPS, Recommendations of the National Institute of Standards and Technology, Technology Administration U.S. Department of Commerce. NIST Special Publication 800-94.
[3] K. Pathan, "The state of the Art in Intrusion Prevention and Detection," CRC press, 2014.
[4] Li Hanguang, Ni Yu, "Intrusion Detection Technology Research Based on Apriori Algorithm", 2012 International Conference on Applied Physics and Industrial Engineering, Physics Procedia 24 (2012), p. 1615 – 1620.
[5] M.Naga Surya Lakshmi, Dr. Y. Radhika "A complete study on intrusion detection using data mining techniques" Volume IX, IJCEA Issue VI, June 2015.
[6] Miroslav Stampar "Artificial Intelligence in Network Intrusion Detection"
[7] Senthilnayaki Balakrishnan, Venkatalakshmi K, Kannan A "Intrusion detection system using Feature selection and Classification technique" IJCSA Volume 3, Issue 4, p. 144-151, 2014
[8] Moradi M and Zulkernine M "A Neural Network based System for Intrusion Detection and Classification of Attacks", Proceedings of IEEE International Conference on Advances in Intelligent Systems – Theory and Applications, Luxembourg, Vol. 148, p. 1-6, 2004.
[9] Wang Jianping, Chen Min and Wu Xianwen, "A Novel Network Attack Audit System based on Multi-Agent Technology", Physics Procedia, Elsevier, Vol. 25, p. 2152 – 2157, 2012
[10] J.Li, Y.Liu and L.Gu "DDoS attack detection based on neural network": Proceedings of the 2nd International Symposium on Aware Computing (ISAC), Tainan, 1–4 Nov.2010, p.196–199.
[11] James Cannady. "Artificial neural networks for misuse detection". In Proceedings of the 1998 National Information Systems Security Conference (NISSC'98) October 5-8 1998. Arlington, VA., p. 443–456, 1998.
[12] B.B. Gupta, C.Joshi and M.Misra "ANN based scheme to predict number of zombies in a DDoS attack", International Journal of Network Security. 13(3) (2011)216–225.
[13] T. Shon and J. Moon, "A hybrid machine learning approach to network anomaly detection," Inf. Sci., vol. 177, no. 18, p. 3799–3821, Sep. 2007.
[14] R. Jain and N. Abouzakhar, "A comparative study of hidden markov model and support vector machine in anomaly intrusion detection," Journal of Internet Technology and Secured Transactions (JITST), vol. 2, no. 1/2/3/4, p. 176–184, 2013.
[15] Cho S-B, Park H-J, "Efficient anomaly detection by modeling privilege flows with hidden Markov model." Computers and Security, 22(1), p. 45-55, 2003.
[16] Tsamardinos I, Aliferis CF, Statnikov A, "Time and sample efficient discovery of Markov blankets and direct causal relations." 9th ACM SIGKDD international conference on knowledge discovery and data mining, ACM press, pp. 673-678, 2003.
[17] Dzeroski S, Zenko B, "Is combining classifiers better than selecting the best one.", ICML 2002. pp. 123-130, 2002