

A Lightweight Online Network Anomaly Detection Scheme Based on Data Mining Methods

Yang Li

Institute of Computing Technology, Chinese Academy of Sciences, PH.D.

Kexueyuan South Rd. 6, Beijing China

liyang@software.ict.ac.cn

Bin-Xing Fang

Institute of Computing Technology, Chinese Academy of Sciences, Professor

Kexueyuan South Rd. 6, Beijing China

bxfang@software.ict.ac.cn

1. INTRODUCTION

Network IDS look for known or potential malicious activities in network traffic and raise an alarm whenever a suspicious activity is detected. Anomaly detection applied to intrusion detection and computer security has been an active area of research since it was originally proposed by Denning in 1980s [1]. Current anomaly detection algorithms [2,3] are particularly effective in detecting network threats such as DoS, DDoS, Worm, spam, etc. However, they are developed very slowly for the following three reasons:

- (a) High false positives attributed to the complex network environment and the poor adaptivity of current anomaly detection methods is difficult to avoid;
- (b) Effective data mining and machine learning based anomaly detection methods are always coming with high computational cost, which is mainly attributed to large scale and high dimension dataset for training.

2. KEY CONTRIBUTIONS

The previous work mainly focus on the sole anomaly detection algorithm while often ignoring the closely related problems, such as the quality of training dataset which highly influences the detection performance, the high dimension data which often leads to “curse of dimensionality”. Aiming at these unresolved problems, we consider the anomaly detection in a comprehensive view. Our work mainly consists of the following contributions: we first propose an efficient unsupervised network anomaly detection scheme based on TCM-KNN (Transductive Confidence Machines for K -Nearest Neighbors) data mining algorithm in this paper, it has higher true positive rate, lower false positive rate than the traditional anomaly detection methods [4]. Secondly, based on the preliminary work, we mainly address the optimizations for TCM-KNN in two aspects. On one hand, Genetic Algorithm (GA) based instance selection method is employed to limit the scale of training set

and select the most qualified instances to ensure the quality of dataset for training. On the other hand, we also adopt filter-based feature selection method to extract the most necessary and relevant features to form the training set for TCM-KNN. The two optimization tasks would boost the performance of TCM-KNN and reduce the computational cost, thereby ensure the effectiveness and efficiency of TCM-KNN for anomaly detection. By doing these, we hope to sufficiently optimize our TCM-KNN algorithm as a lightweight anomaly detection mechanism.

3. OUR METHODS

Transductive Confidence Machines for K -Nearest Neighbors (TCM-KNN) introduced the computation of the confidence using algorithmic randomness theory [2]. Unlike traditional methods in data mining, transduction can offer measures of reliability to individual points, and uses very broad assumptions except for the iid assumption (the training as well as new (unlabelled) points are independently and identically distributed). It introduces an important measure to help determine the data point in handle as an anomaly or not: *strangeness*. It serves as a measure of how well the data fits the current classes. If the ratio of the points in the training dataset, whose strangeness is higher than that of the point to be diagnosed, exceeds a preset threshold, then we can claim the point is an anomaly with respect to the training dataset.

For feature selection, Information Gain (IG) is calculated for class labels by employing a binary discrimination for each class. That is, for each class, a dataset instance is considered in-class, if it has the same label; otherwise, it will be considered out-class. Consequently, as opposed to calculating one information gain as a general measure on the relevance of the feature for all classes, we calculate an information gain for each class. Thus, this signifies

how well the feature can discriminate the given class (i.e. normal or abnormal type) from other classes.

As the instance selection for TCM-KNN, training dataset is denoted as *TR* with instances. The search space associated with the instance selection of *TR* is constituted by all the subsets of *TR*. Then, the chromosomes should represent subsets of *TR*. This is accomplished by using a binary representation. A chromosome consists of genes (one for each instance in *TR*) with two possible states: 0 and 1. If the gene is 1, then its associated instance is included in the subset of *TR* represented by the chromosome. If it is 0, then this instance does not occur. After running GA algorithm, the selected chromosomes would be the reduced training dataset for TCM-KNN. We employ four well-known GAs, GGA (Generational Genetic Algorithm), SGA (Steady-State Genetic Algorithm), CHC (heterogeneous recombination and cataclysmic mutation) adaptive search algorithm, PBIL (Population-Based Incremental Learning), to fulfill the instance selection tasks.

4. EXPERIMENTAL RESULTS

To verify the effectiveness and availability of our work, we apply it to anomaly detection for web server. We setup a web server located in the college running apache http service (version 2.2) on Linux platform (Red Hat Enterprise Ver 4, kernel 2.6.9-42.0.2.EL). We conducted many experiments over several days during busy hours and with background traffic generated from more than 5000 hosts of the college. In the experiments, the attackers can access the victim web server and they launched well-known DDoS attacks using a series of DDoS tools such as Stacheldraht and TFN2K. We performed many flooding attacks with spoofed IP's like SYN Floods, UDP and ICMP attacks. We first took use of TCM-KNN to detect these traffic anomalies, then adopting feature selection and instance selection mechanism discussed above to optimize it (we finally selected 5 features (from 20) and 5,600 data points (from 98,000) for training), Table 1 shows the detailed experimental results. We found that we could determine the anomalous points with accuracy of 100% (2,600 abnormal points are all correctly diagnosed) and only 1.28% false positives (only 194 out 15,120 normal data points are misjudged) in the real network environment. And, even after the optimizations, the TP (true positive rate) keeps high (99.38%) and the FP (false positive rate) is still manageable (1.87%) in real

network environment. The most important inspiring result we got is the detection time for an anomaly is rather short (sharply reduced from 0.4164s to 0.1397s for diagnosing each statistical data point), thus the system based on our optimized TCM-KNN could deal with the large amounts of anomalies in the network and make corresponding countermeasures to mitigate them. Hence, all the results (see Table 1, which demonstrates the detail optimization results) substantially evident that our TCM-KNN algorithm could be effectively optimized as a lightweight anomaly detection method suitable for near real-time detection and is suitable for real large-scale network traffic environment.

Table 1. Experimental Results

	Building Time	Detection Time	TP	FP
TCM-KNN (original)	22218.62 Sec.	0.4164 Sec.	100%	1.28%
TCM-KNN (optimized)	363.86 Sec.	0.1397 Sec.	99.38%	1.87%

5. CONCLUSIONS

This paper presents our preliminary work in network anomaly detection. The experimental results demonstrate an inspiring and promising trend for lightweight on-line network anomaly detection, which is rather meaningful for the ever-increasing network traffic and the accompanied network threats. In our future work, we will further verify and optimize our methods in terms of the concrete applications, as well as deploying it in our national backbone network to detect anomalies such as DoS, DDoS, probe, spam, etc.

6. REFERENCES

- [1] D.E. Denning. An intrusion detection model, IEEE Transactions on Software Engineering, SE-13, 1987, 222-232.
- [2] E. Eskin, A. Arnold, et al. A geometric framework for unsupervised anomaly detection: detecting intrusions in unlabeled data. Applications of Data Mining in Computer Security, Kluwer, 2002.
- [3] W. Lee, S. J. Stolfo. Data mining approaches for intrusion detection. In Proceedings of the 1998 USENIX Security Symposium, 1998.
- [4] Y. Li, B.X. Fang, et al. Network Anomaly Detection Based on TCM-KNN Algorithm. In Proceedings of the 2nd ACM symposium on Information, computer and communications security (ASIACCS' 2007), ACM, 2007, 13-19.