Information Sciences 318 (2015) 91–110

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Detecting anomalies from big network traffic data using an adaptive detection approach

Ji Zhang [a,*], Hongzhou Li [b], Qigang Gao [c], Hai Wang [d], Yonglong Luo [e,*]

[a] *University of Southern Queensland, Toowoomba, QLD 4350, Australia*
[b] *Guilin University of Electronic Technology, Guilin City, Guang Xi Province 541004, China*
[c] *Dalhousie University, Halifax, Nova Scotia B3H 2Z3, Canada*
[d] *Saint Mary's University, Halifax, Nova Scotia B3H 3C3, Canada*
[e] *Anhui Normal University, Wuhui, Anhui Province 24100, China*

ARTICLE INFO

Article history:
Received 22 January 2014
Received in revised form 19 July 2014
Accepted 29 July 2014
Available online 7 August 2014

Keywords:
Anomaly detection
Big data
Outlier detection

ABSTRACT

The unprecedented explosion of real-life big data sets have sparked a lot of research interests in data mining in recent years. Many of these big data sets are generated in network environment and are characterized by a dauntingly large size and high dimensionality which pose great challenges for detecting useful knowledge and patterns, such as network traffic anomalies, from them. In this paper, we study the problem of anomaly detection in big network connection data sets and propose an outlier detection technique, called Adaptive Stream Projected Outlier deTector (A-SPOT), to detect anomalies from large data sets using a novel adaptive subspace analysis approach. A case study of A-SPOT is conducted in this paper by deploying it to the 1999 KDD CUP anomaly detection application. Innovative approaches for training data generation, anomaly classification and false positive reduction are proposed in this paper as well to better tailor A-SPOT to deal with the case study. Experimental results demonstrate that A-SPOT is effective and efficient in detecting anomalies from network data sets and outperforms existing detection methods.

1. Introduction

We have witnessed in recent years a tremendous research interest sparked by the explosion of big data, many of which are collected and transferred over the network. Such data sets, which capture the real-time network traffic situation, are typically very large in size and well beyond human's capability of comprehension without the use of proper intelligent computer technology to process them efficiently. One of the important kinds of patterns that can be detected from these big network-based data sets are network traffic anomalies which are suspects/candidates of malicious network intrusions. An intrusion into a computer network would compromise the stability and security of the network, leading to possible loss of privacy, information and revenue [23]. To safeguard network security, there are two major classes of approaches for detecting anomalies that may carry the manifestations of intrusions: *misuse-based detection* (or signature-based detection) and *anomaly-based detection*. The misuse-based detection approaches extract features from audit streams and compare these features with known signatures, formulated as patterns or rules, that are provided by domain experts. An intrusion is

[*] Corresponding authors.
E-mail addresses: ji.zhang@usq.edu.au (J. Zhang), ylluo@ustc.edu.cn (Y. Luo).

http://dx.doi.org/10.1016/j.ins.2014.07.044
0020-0255/© 2014 Elsevier Inc. All rights reserved.

detected if the features violate one or more given signatures. The misuse-based detection methods are relatively simple and accurate in detecting known types of intrusions. Yet, due to limited knowledge of domain experts, the misuse-based detection methods cannot effectively detect the previously unseen intrusions. In contrast, the anomaly-based detection methods build the models or profiles for the normal data and consider the data with noticeable deviations from these models or profiles as intrusions. These methods can effectively detect new intrusions. However, they typically have a high False Positive Rate (FPR) and most of them do not have relevant mechanisms to deal with false positives and rely entirely on human (such as the security officers) to further screen the detected anomalies. This is time-consuming and error prone. As the ability to detect new attacks is usually considered more important than low false positive rate, the anomaly-based detection methods are thus more appealing and receiving more and more attentions. The anomaly-based detection is very similar in spirit to the problem of outlier detection. Consequently, the problem of anomaly detection can be largely solved by employing outlier detection methods that have been proposed in recent years.

In many cases, network data sets can be modeled as high-dimensional connection oriented records each of which contains a number of features to measure the quantitative behaviors of the network traffic, as in the 1999 KDD CUP anomaly detection application. A salient characteristics of high-dimensional data is that almost all the anomalies are embedded in some lower-dimensional subspaces (spaces consisting of a subset of attributes) due to the so-called Curse of Dimensionality. These anomalies are termed *subspace anomalies* in the high-dimensional space context. As the dimensionality of data goes up, data tend to be equally distant from each other. As a result, the difference of data points' outlier-ness will become increasingly weak and thus undistinguishable. Only in moderate or low dimensional subspaces can significant outlier-ness of data be observed. The problem of detecting subspace anomalies from high-dimensional data streams can be formulated as follows: given a data stream $\mathcal{D}$ with a potentially unbounded size of φ-dimensional data points, each data point $p_i = \{p_{i1}, p_{i2}, \ldots, p_{i\varphi}\}$ in $\mathcal{D}$ will be labeled as either a subspace anomaly if it is found abnormal in one or more subspaces. It will be flagged as a regular data otherwise. If p_i is a subspace anomaly, its associated outlying subspace(s) will be presented as well in the result.

Most of the conventional outlier/anomaly detection techniques that are capable of detecting anomalies are only applicable to relatively low dimensional and static data sets (stored in databases without frequent changes) [5,11,12,17,19]. Recently, there are some emerging work in dealing with outlier detection either in high-dimensional data or data streams. However, there lacks substantial research work on exploring the intersection of these two active research areas to efficiently deal with big network-originated data. For those methods in subspace outlier detection in high-dimensional space [1,3,6,10,18,20–22,24], their measurements used for evaluating points' outlier-ness are not incrementally updatable and many of the methods involve multiple scans of data, making them incapable of handling fast data streams. The techniques for tackling outlier detection in data streams [2,16,25] rely on full data space to detect outliers and thus the subspace outliers cannot be discovered by these techniques. There are some recent research work on outlier/anomaly detection using graph/diagram-based approaches [13,15] and visualization method [26], but they are not able to efficiently deal with high-dimensional data streams either.

To detect anomalies from big high-dimensional data sets that can be readily applied to networking environment, we have developed a new technique, called Adaptive Stream Projected Outlier deTector (A-SPOT). A-SPOT constructs Sparse Subspace Template (SST), a set of subspaces where anomalies are more likely to be detected. A-SPOT uses multiple criteria, called *Projected Cell Summaries (PCS)*, to measure the outlier-ness of data and draw on Multi-Objective Genetic Algorithm (MOGA) to search for the subspaces where subspace anomalies exist in order to obtain SST. To test the applicability of A-SPOT, we apply it in 1999 KDD CUP anomaly detection application. More specifically, the scientific contributions of this paper are summarized as follows:

- A-SPOT is able to deal with high-dimensional data streams in the network environment, where most existing approaches cannot. As discussed above, most of the existing approaches can either handle streaming data or high-dimensional data, but not both.
- A-SPOT is able to effectively and efficiently explore subspaces to detect anomalies embedded in subspaces where most of anomalies exist for high-dimensional data.
- A-SPOT can adapt to the dynamics of data streams by using a dynamic set of detecting subspaces, which is able to not only remarkably speed up the detection process, but also reduce the false positives in the detection result.
- A few of innovative techniques are also developed to particularly dealing with the case study using 1999 KDD data set including training data generation, anomaly categorization using outlying subspaces analysis and false positive reduction. These techniques are utilized to enable our method to detect anomalies more efficiently and effectively.
- The experimental evaluation demonstrates that A-SPOT outperforms the existing approaches in terms of effectively detecting anomalies in subspaces for big high-dimensional streaming network traffic data.

2. Overview of A-SPOT

Our technique for anomaly detection in data streams, also known as A-SPOT, can be broadly divided into two stages: the learning and detection stages. The learning stage of A-SPOT can further support two types of learning paradigms, namely offline learning and online learning. In the offline learning, Sparse Subspace Template (SST) is constructed using either