Anomaly Detection in High-dimensional Network Data Streams: A Case Study

Ji Zhang, Qigang Gao
Faculty of Computer Science
Dalhousie University
Halifax, Nova Scotia, Canada
{jiz, qggao}@cs.dal.ca

Hai Wang
Sobey School of Business
Saint Mary's University
Halifax, Nova Scotia, Canada
hwang@smu.ca

Abstract

In this paper, we study the problem of anomaly detection in high-dimensional network streams. We have developed a new technique, called Stream Projected Ouliter deTector (SPOT), to deal with the problem of anomaly detection from high-dimensional data streams. We conduct a case study of SPOT in this paper by deploying it on 1999 KDD Intrusion Detection application. Innovative approaches for training data generation, anomaly classification and false positive reduction are proposed in this paper as well. Experimental results demonstrate that SPOT is effective in detecting anomalies from network data streams and outperforms existing anomaly detection methods.

1 Introduction

In recent years, we have witnessed a tremendous research interest sparked by the explosion of data collected and transferred in the format of streams over the network. An intrusion into a computer network can compromise the stability and security of the network leading to possible loss of privacy, information and revenue [8]. In many cases, network data streams can be modeled as high-dimensional vectors where each of them contains a number of varied features to measure the quantitative behaviors of the network traffic, as in the 1999 KDD Intrusion Detection application. For high-dimensional data, we notice that most, if not all, anomalies existing in high-dimensional data streams are embedded in some lower-dimensional subspaces (spaces consisting of a subset of attributes). These anomalies are termed *projected anomalies* in the high-dimensional space.

Most of the conventional outlier detection techniques that are capable of detecting anomalies are only applicable to relatively low dimensional data sets [2][5]. For those methods in projected outlier detection in high-dimensional space [9][7][6], 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 [1] rely on full data space to detect outliers and thus projected outliers cannot be discovered by these techniques.

To detect anomalies from high-dimensional data streams, we have developed a new technique, called Stream Projected Outlier deTector (SPOT). SPOT uses multiple criteria, called *Projected Cell Summaries (PCS)*, to measure the outlier-ness of data and draws on Multi-objective Genetic Algorithm (MOGA) to search for the subspaces where the projected anomalies exist. As the major contributions of this paper, we have tackled several important issues, including training data generation, anomaly categorization and false positive reduction, in an effort to successfully apply SPOT in 1999 KDD intrusion detection application.

2 Overview of SPOT

SPOT can be divided into two stages: learning and detection stages. In learning stage, Sparse Subspace Template (SST) is constructed by supervised and/or unsupervised learning process. SST casts light on where projected anomalies are likely to be found in the high-dimensional space. Based upon SST, SPOT screens projected anomalies from constantly arriving data in the detection stage.

In SPOT, SST consists of the following three subspace subsets, *Fixed SST Subspaces (FS), Clustering-based SST Subspaces (CS)* and *Outlier-driven SST Subspaces (OS)*, respectively. Fixed SST Subspaces (FS) contains all the subspaces in the full lattice whose maximum dimension is $MaxDimention$, where $MaxDimention$ is a user-specified parameter. Clustering-based SST Subspaces (CS) consists of the sparse subspaces of the top training data that have the highest overall outlying degree. The selected training data are more likely to be considered as anomalies that can be potentially used to detect more subsequent anomalies in the stream. The overall outlying degree of training data is computed by employing clustering method. The average size of clusters a point belongs to provides an useful insight into its accurate outlying degree. The sparse subspaces of the top training data, obtained by MOGA, will become CS of SST. In some cases, a few anomaly examples may be provided by domain experts. MOGA is applied on each of these outliers to find their top sparse subspaces. There subspaces are called Outlier-driven SST Subspaces

ISI 2008, June 17-20, 2008, Taipei, Taiwan
1-4244-2415-3/08/$20.00 ©2008 IEEE

Authorized licensed use limited to: UNIVERSITY OF SOUTHERN QUEENSLAND. Downloaded on August 6, 2009 at 20:01 from IEEE Xplore. Restrictions apply.

(OS). Based on OS, we can effectively detect more anomalies that are similar to these anomaly examples.

The detection stage performs anomaly detection for incoming stream data using SST. As streaming data arrive continuously, the data synapses. i.e. PCS, are first updated dynamically in order to capture new information of arrived data. Then, we retrieve PCS of each subspace in SST to see whether or not this data is a projected anomaly. These steps can be performed very fast so that the time criticality requirement posed by the data streams can be met.

3 Network Anomaly Detection using SPOT

Our case study is anomaly detection from 1999 KDD intrusion detection data stream. This data stream contains a wide variety of intrusions simulated in a military network environment. Each instance in this data set is a vector of extracted feature values from a connection record obtained from the raw network data gathered during the simulated intrusions. Each connection was labeled as either normal or as exactly one specific kind of attacks. The simulated attacks fall into one of the following four categories: DoS, R2L, U2R and Probing. A training data set is available in data set for model building purpose. The goal of this application is to category a large number of unlabeled network connections into the above-mentioned four classes.

In this case study, we need to address several important issues in order to successfully deploy SPOT. These issues include generation of new training data sets, incorporation of anomaly categorization function into SPOT and development of new mechanism to noticeably reduce false positives.

3.1 Training Data Generation

The training data set available in this case study cannot be directly used by SPOT and other anomaly-based detectors. This is due to the high proportion of attacks instances in this training data set; as high as 91% of the samples in this training data are attacks. Normal data behavior is needed in quantitizaton of fitness function of MOGA in the first step of SPOT outlined above.

In this case study, we adopt a multiple-sample generation strategy to meet the learning needs of SPOT. The basic idea is that, for each intrusion class, multiple training samples are generated and MOGA is applied on each of these samples to produce OS for each class. Mathematically, let D_T be the original training data set available. D_T consists two parts, the normal and intrusion samples, denoted by D_N and D_I, respectively, where D_I consists intrusion samples of the four different classes, i.e. $D_I = \cup_{i=1}^{4} D_I^i$. In our work, we generate multiple new training samples with replacement from D_I^i for $i \in [1, 4]$, each such new sample can be expressed as $D_{T'}^{i,j} = D_N \cup D_I^{i,j}$, where j is the number of samples generated from D_I^i. By applying MOGA on $D_{T'}^{i,j}$, we can obtain OS for class $i \in [1, 4]$, denoted as OS^i, as $OS^i = \cup_j MOGA(D_{T'}^{i,j})$. The complete OS is simply the union of OS^i for $i \in [1, 4]$.

3.2 Intrusion Classification

Intrusion classification mainly involves the categorization of detected anomalies into one of the known intrusion classes or the class of false positive. To do this, we need to construct the signature subspaces for each intrusion classes. The signature subspaces for a target class are those subspaces that can be used to identify intrusions for this particular class. To construct the signature subspaces for a particular class, we collect the outlying subspaces of those anomalies belonging to this class.

Within a class, different signature subspaces have varying weights as to correctly identify anomalies for this class. The weight of a signature subspace s with respect to c represent the discriminating ability of s towards c. In our work, we borrow the idea of tf-idf (*term frequency-inverse document frequency*) weighting method, a commonly used technique in the domain of information retrieval and text mining, to measure the weight of signature subspaces for each class.

Similarity measure needs to be properly defined before the anomalies can be classified. The similarity between an anomaly o and class c is defined as their average inner product, i.e. $Sim(o, c) = \frac{o \cdot c}{|OS(o)|}$, where $|OS(o)|$ denotes the number of outlying subspaces of o.

3.3 Handle False Positives

False positives are those anomalies that are erroneously detected as the attacks by the system. They consume a fair amount of human efforts and consequently makes it almost impossible for security officers to really concentrate only on the real attacks. It is much desirable to quickly screen out these false-positives in order to allow closer attention to be paid towards the real harmful attacks.

In this case study, there are no available false-positive exemplars in the training data set. Therefore, unlike the attack classes, it is not easy to directly create the signature subspaces for the false-positive class. However, there are a large amount of normal data in the training data set. If any of them are found abnormal, i.e. having some outlying subspaces, then they will be considered as false positives. The outlying subspaces of these false positives will be used to construct the signature subspaces for false-positive class.

4 Experimental Evaluation

We conduct experimental evaluation on SPOT in 1999 KDD Intrusion Detection application. Comparative study is also performed to investigate the detection accuracy and false positive rate between SPOT and other existing anomaly detection methods. The methods we would like to compare with SPOT in this case study include the anomaly

Authorized licensed use limited to: UNIVERSITY OF SOUTHERN QUEENSLAND. Downloaded on August 6, 2009 at 20:01 from IEEE Xplore. Restrictions apply.

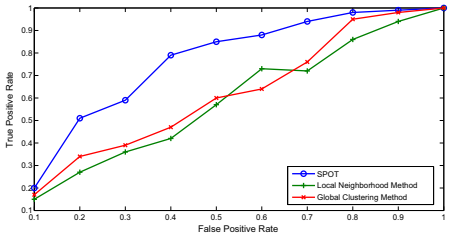

Figure 1. ROC curves of different methods

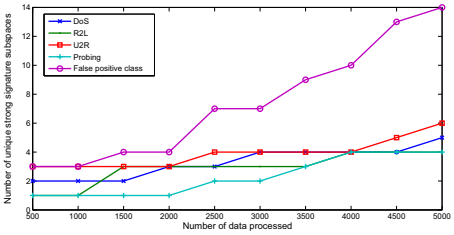

Figure 2. Effect of number of samples for each anomaly class

detection method proposed in [4] and [3]. The first method defines the outlier-ness of a data point using the distance between it and its k^{th} neighborhood (called local neighborhood method) [4]. The second method uses clustering method to identify those data that are far from the dense regions of the data points as anomalies (called global clustering method) [3].

ROC Analysis. Receiver Operating Characteristic (ROC) analysis is a commonly used technique for performance evaluation of detection methods by plotting its True Positive Rate (TPR) and False Positive Rate (FPR) in the same ROC space. True Positive Rate refers to the percentage of hits (correctly identified anomalies) in the whole set of anomalies existing in the data set and False Positive Rate represents the percentage of erroneously labeled anomalies in the whole set of normal data. Since these two metrics are conflicting to achieve, we thus need to consider these two rates simultaneously. In Figure 1, we plot the ROC curves for SPOT and the two competitive methods. We can see from this figure that the ROC curve of SPOT progresses much closer to the upper-left coner of the plot than the curves of the other two methods, indicating that SPOT achieves a much better detection performance. SPOT performs subspace exploration in anomaly detection, thus it has a better anomaly detection ability that leads to better detection accuracy. In addition, the false positive classification enable SPOT to identify false positives in an automated fashion while the competitive methods cannot.

Signature Subspace Analysis. We are also interested in studying the diversity of signature subspaces of false-positives, as compared with those of the attack classes. We record the number of unique strong signature subspaces for anomalies in different classes (including the false-positive class) as the number of data we evaluated increases. We plot the results in Figure 2. It is clear that the number of unique signature subspaces for the false-positive class is significantly higher than any other attack class by a factor of three or four. This means that the signature subspaces for the false-positive is far more diverse that those of the attack classes. This finding offers an important guidance to us in creating the set of signature subspaces of the false-positive class. We need to collect a relatively large pool of signature subspaces in the detection process for achieving accurate detection of false-positives.

5 Conclusions

In this paper, we investigate anomaly detection problem in high-dimensional data streams. We carry out a case study using 1999 KDD Intrusion Detection data set. As the major contributions, several important issues, including training data generation, anomaly categorization using outlying subspaces analysis and false positive reduction, have been successfully tackled in this paper for rendering SPOT applicable in this study. The experimental evaluations show that SPOT is not only more effective in detecting anomalies but also produces a significantly lower level of false positives than the existing anomaly detection methods.

References

[1] C. C. Aggarwal. On abnormality detection in spuriously populated data streams. In *SDM*, 2005.

[2] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander. Lof: Identifying density-based local outliers. In *SIGMOD Conference*, pages 93–104, 2000.

[3] H. Cui. Online outlier detection detection over data streams. In *Master thesis, Simon Fraser University*, 2002.

[4] E. Eskin, A. Arnold, M. Prerau, L. Portnoy, and S. Stolfo. A geometric framework for unsupervised anomaly detection: Detecting intrusions in unlabeled data. In *Data Mining for Security Applications*, 2002.

[5] E. M. Knorr and R. T. Ng. Finding intensional knowledge of distance-based outliers. In *VLDB*, pages 211–222, 1999.

[6] J. Zhang, Q. Gao, and H. Wang. A novel method for detecting outlying subspaces in high-dimensional databases using genetic algorithm. In *ICDM*, pages 731–740, 2006.

[7] J. Zhang and H. Wang. Detecting outlying subspaces for high-dimensional data: the new task, algorithms, and performance. *Knowl. Inf. Syst.*, 10(3):333–355, 2006.

[8] S. Zhong, T. M. Khoshgoftaar, and S. V. Nath. A clustering approach to wireless network intrusion detection. In *ICTAI*, pages 190–196, 2005.

[9] C. Zhu, H. Kitagawa, and C. Faloutsos. Example-based robust outlier detection in high dimensional datasets. In *ICDM*, pages 829–832, 2005.

Authorized licensed use limited to: UNIVERSITY OF SOUTHERN QUEENSLAND. Downloaded on August 6, 2009 at 20:01 from IEEE Xplore. Restrictions apply.