Chapter 26
A Subspace-Based Analysis Method for Anomaly Detection in Large and High-Dimensional Network Connection Data Streams

Ji Zhang
University of Southern Queensland, Australia

ABSTRACT

A great deal of research attention has been paid to data mining on data streams in recent years. In this chapter, the authors carry out a case study of anomaly detection in large and high-dimensional network connection data streams using Stream Projected Outlier deTector (SPOT) that is proposed in (Zhang et al. 2009) to detect anomalies from data streams using subspace analysis. SPOT is deployed on the 1999 KDD CUP anomaly detection application. Innovative approaches for training data generation, anomaly classification, and false positive reduction are proposed in this chapter as well. Experimental results demonstrate that SPOT is effective and efficient in detecting anomalies from network data streams and outperforms existing anomaly detection methods.

INTRODUCTION

Great research efforts have been taken by researchers in recent years to study discovery of useful patterns from data streams. One important category of such data streams are the streams collected over the network. Analyzing these network data streams is quite critical in unveiling suspicious patterns that may indicate network intrusions. An intrusion into a computer network can compromise the stability and security of the network, leading to possible loss of privacy, information and revenue (Zhong et al. 2004). To safeguard network security, there are two major classes of approaches for detecting

DOI: 10.4018/978-1-4666-2455-9.ch026

anomalies that may represent the manifestations of intrusions: misuse-based detection (or signature-based detection) and anomaly-based detection.

As far as data format representation is concerned, data streams collected in network environments can be typically, but not necessarily, modeled as continuously arriving high-dimensional connection oriented records. Each record contains a number of varied features to measure the quantitative behaviors of the network traffic. Such data representation is used in the 1999 KDD CUP anomaly detection application. In high-dimensional space, anomalies 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 context. The underlying reason for this phenomenon is the Curse of Dimensionality. The increase in dimensionality will make data to be equally distant from each other. Consequently, 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. This is the major motivation for detecting outliers in subspaces.

We can formulate the problem of detecting projected anomalies from high-dimensional data streams as follows: given a data stream D with ϕ-dimensional data points, each data point pi = {pi1, pi2,..., piϕ} in D will be labeled as either a projected anomaly if it is found abnormal in one or more subspaces. Otherwise, it will be flagged as a regular data. If pi is a projected anomaly, its associated outlying subspace(s) will be presented as well in the result.

Unfortunately, the existing outlier/anomaly detection techniques are mostly limited in identifying anomalies embedded in subspaces. Most are only capable of detecting anomalies in relatively low dimensional and static data sets (stored in databases without frequent changes) (Breuning et al., 2000; Knorr et al., 1998; Knorr et al., 1999; Ramaswamy et al, 2000; Tang et al., 2002). Recently, there are some emerging work in dealing with outlier detection either in high-dimensional data or data streams. However, there have not been any substantial research work so far for exploring the intersection of these two active research areas. For those methods in projected outlier detection in high-dimensional space (Aggarwal et al., 2001; Aggarwal et al., 2005; Boudjeloud et al., 2005; Zhu et al., 2005; Zhang et al., 2006; Guha et al., 2009), 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 (Aggarwal, 2005; Palpanas et al, 2003;, Zhang et al., 2010) rely on full data space to detect outliers and thus the projected outliers cannot be discovered by these techniques.

To detect anomalies from high-dimensional data streams, a new technique, called Stream Projected Outlier deTector (SPOT), is proposed (Zhang et al, 2009). It utilizes a novel subspace analysis method to detect anomalies hidden in the subspaces of the full data space. In this paper, efforts are taken to carry out a real-life case study of SPOT to test its practical applicability. We apply in 1999 KDD CUP anomaly detection application. We have also tackled several important issues, including training data generation, anomaly categorization using outlying subspaces analysis and false positive reduction, for a successful deployment of SPOT in the case study. Experimental evaluates reveals that SPOT is efficient in this application for detecting anomalies.

Overview of SPOT

Before the case study is carried out, it is worthwhile presenting a short description of SPOT. SPOT performs anomaly detection into two stages: the learning and detection stages. SPOT can further support two types of learning, namely offline learning and online learning. In the offline learning,