Information Sciences 348 (2016) 243–271

Contents lists available at ScienceDirect

Information Sciences

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

A multi-step outlier-based anomaly detection approach to network-wide traffic

Monowar H. Bhuyan[a,*], D.K. Bhattacharyya[b], J.K. Kalita[c]

[a] Department of Computer Science & Engineering, Kaziranga University, Koraikhowa, Jorhat 785006, Assam, India
[b] Department of Computer Science & Engineering, Tezpur University, Napaam, Tezpur 784028, Assam, India
[c] Department of Computer Science, University of Colorado at Colorado Springs, CO 80933-7150, USA

ARTICLE INFO

Article history:
Received 4 June 2014
Revised 1 February 2016
Accepted 8 February 2016
Available online 15 February 2016

Keywords:
Anomaly detection
Network-wide traffic
Clustering
Reference point
Outlier score

ABSTRACT

Outlier detection is of considerable interest in fields such as physical sciences, medical diagnosis, surveillance detection, fraud detection and network anomaly detection. The data mining and network management research communities are interested in improving existing score-based network traffic anomaly detection techniques because of ample scopes to increase performance. In this paper, we present a multi-step outlier-based approach for detection of anomalies in network-wide traffic. We identify a subset of relevant traffic features and use it during clustering and anomaly detection. To support outlier-based network anomaly identification, we use the following modules: a mutual information and generalized entropy based feature selection technique to select a relevant non-redundant subset of features, a tree-based clustering technique to generate a set of reference points and an outlier score function to rank incoming network traffic to identify anomalies. We also design a fast distributed feature extraction and data preparation framework to extract features from raw network-wide traffic. We evaluate our approach in terms of detection rate, false positive rate, precision, recall and *F*-measure using several high dimensional synthetic and real-world datasets and find the performance superior in comparison to competing algorithms.

1. Introduction

There is growing need for efficient algorithms to detect exceptional patterns or anomalies in network-wide traffic data. Network-wide traffic anomalies disrupt normal network operations. Therefore, detection of anomalies in network-wide traffic has been of great interest for many years. Wide-area network traffic is voluminous, high dimensional and noisy, making it difficult to extract meaningful information to discover anomalies by examining traffic instances. Change of traffic characteristics over time is a crucial characteristic of network-wide traffic. Network-wide traffic data contain both categorical and numeric attributes [29,57]. The task of outlier discovery in such data has five subtasks: (a) dependency detection, (b) class identification, (c) class validation, (d) frequency detection and (e) outlier or exception detection [32,36]. The first four subtasks consist of finding patterns in large datasets and validating the patterns. Techniques for association rules, classification and data clustering are used in the first four subtasks. Outlier detection focuses on a very small percentage of data objects,

* Corresponding author. Tel.: +91 94353 88234; fax: +91 3762351318.
E-mail addresses: monowar.tezu@gmail.com (M.H. Bhuyan), dkb@tezu.ernet.in (D.K. Bhattacharyya), jkalita@uccs.edu (J.K. Kalita).

http://dx.doi.org/10.1016/j.ins.2016.02.023
0020-0255/

which are often ignored or discarded as noise in normal analysis. Outlier detection techniques focus on discovering infrequent pattern(s) in the data, as opposed to many traditional data mining techniques such as association analysis or frequent itemset mining that attempt to find patterns that occur frequently.

Outliers may represent aberrant data that may affect systems adversely by producing incorrect results, incorrect models and biased estimation of parameters. Outlier detection enables one to identify such aberrant data prior to modeling and analysis [38]. There are many significant applications of outlier detection. For example, in the case of credit card usage monitoring or mobile phone monitoring, a sudden change in usage pattern may indicate fraudulent usage such as stolen cards or stolen phone airtime. Outlier detection can also help discover critical entities such as in military surveillance where the presence of an unusual region in a satellite image in an enemy area could indicate enemy troop movement. Most outlier detection algorithms make the assumption that normal instances are far more frequent than outliers or anomalies. Generally, network intrusion detection techniques are of two types: *signature-based* and *anomaly-based* [18,22,45,51]. *Signature-based* detection aims to detect intrusions or attacks from known intrusive patterns. It cannot detect new or unknown attacks. *Anomaly-based* detection looks for attacks based on deviations from established profiles or signatures of normal activities. Events or records that exceed certain threshold scores are reported as anomalies or attacks. It can detect unknown attacks based on the assumption that the attack data deviate from normal data behavior. However, a drawback of anomaly-based systems is high false alarm rates. Lowering the percentage of false alarms is the main challenge in anomaly-based network intrusion detection. Outlier-based anomaly detection is an effective method in detecting network anomalies with desirable accuracy.

Outlier detection techniques [15,32,47] are usually developed based on distance or density computation or a combination of both. Outlier detection can use soft computing as well as statistical measures. Several outlier detection techniques have been developed and applied to network anomaly detection [46,49,58]. A general outlier detection technique, when tuned for network intrusion detection, does not perform well in high dimensional large datasets. A possible cause is that the classes are embedded inside a subspace. In many applications, the data organization has inherently overlapping clusters. This includes network-wide traffic and gene expression data. Hence, in such applications, some instances may be allowed to be members of two or more clusters based on the subspace which is more appropriate than forcing them to belong to a single cluster. Allowing cluster overlap also reduces the cost in computing similarity. Additionally, in case of score-based outlier detection, the score values may not vary commensurately as candidate objects change during testing. In such cases, it is very difficult to reliably assign a label as normal or outlier for a candidate object.

To address such issues, we develop an efficient multi-step outlier-based technique to analyze high dimensional voluminous network-wide traffic data for anomaly detection. Our method has several redeeming features including: (i) It selects a subset of relevant features that reduces the computational cost during anomaly detection. (ii) It works with any proximity measure and can identify both disjoint and overlapping clusters hierarchically for any dataset. (iii) The proposed score function can identify network anomalies or outliers with few false alarms. (iv) It performs exceptionally well for DoS, probe and R2L attacks when applied to network-wide traffic datasets. Specifically, the technical contributions of this paper include the following.

- We propose MIGE-FS, a mutual information and generalized entropy based feature selection technique to select a subset of relevant features that make detection faster and more accurate.
- A difficulty in clustering high dimensional large network-wide traffic datasets is handling of mixed type data and arranging the data in computationally efficient structures for analysis. For example, *protocol* is categorial and *byte count* is numeric. Another key issue is coming up with a distance function that incorporates subspaces to find meaningful clusters. In this paper, we propose TCLUS, an effective tree-based clustering algorithm based on relevant subspace computation, to identify compact as well as overlapping clusters.
- We develop an outlier score function and use it to detect anomalies or outliers efficiently. The score function uses the TCLUS algorithm for generating reference points for each cluster. We apply this approach in network-wide traffic anomaly detection and obtain excellent results using several real-life network-wide traffic datasets.
- We extract various features including basic, content-based, time-based and connection-based features at both packet and flow levels from captured raw network-wide traffic datasets, which are generated using our TUIDS (Tezpur University Intrusion Detection System) testbed [14] by using a fast distributed feature extraction framework. We prepare a feature-based intrusion dataset for training and evaluating the proposed approach.

The rest of the paper is organized as follows: Section 2 discusses related work on outlier-based techniques with an application to network-wide traffic anomaly detection while Section 3 provides the problem formulation. In Section 4, we introduce the foundations of our proposed approach. Section 5 discusses the proposed approach in three parts, which are feature selection, clustering and anomaly detection while Section 6 presents empirical evaluation of our approach using synthetic and several real-life datasets. Finally, Section 7 contains concluding remarks and future work.

2. Related work

In the recent past, a good number of anomaly and outlier detection techniques have been published [16,19,27,30,53,55] in the literature. We broadly classify these techniques into four types: (a) statistical, (b) distance-based, (c) density-based and (d) soft computing.