Information Sciences 429 (2018) 315–331

Contents lists available at ScienceDirect

Information Sciences

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

Incremental anomaly detection using two-layer cluster-based structure

Elnaz Bigdeli[a], Mahdi Mohammadi[b], Bijan Raahemi[b,*], Stan Matwin[c,d]

[a] School of Electrical Engineering and Computer Science, University of Ottawa, 55 Laurier Ave., East, Ottawa, Ontario, Canada, K1N6N5
[b] Knowledge Discovery and Data Mining Lab, University of Ottawa, 55 Laurier Ave., East, Ottawa, Ontario, Canada, K1N6N5
[c] Faculty of Computer Science, Dalhousie University, 6050 University Ave, Halifax, NS, Canada, B3H4R3
[d] Institute of Computer Science, Polish Academy of Sciences, Jana Kazimierza 5, 01-248 Warszawa, Poland

ARTICLE INFO

Article history:
Received 14 February 2017
Revised 13 November 2017
Accepted 15 November 2017
Available online 15 November 2017

Keywords:
Anomaly detection
Incremental clustering
Noise resilience
Gaussian mixture model

ABSTRACT

Anomaly detection algorithms face several challenges, including processing speed, adapting to changes in dynamic environments, and dealing with noise in data. In this paper, a two-layer cluster-based anomaly detection structure is presented which is fast, noise-resilient and incremental. The proposed structure comprises three main steps. In the first step, the data are clustered. The second step is to represent each cluster in a way that enables the model to classify new instances. The Summarization based on Gaussian Mixture Model (SGMM) proposed in this paper represents each cluster as a GMM. In the third step, a two-layer structure efficiently updates clusters using GMM representation, while detecting and ignoring redundant instances. A new approach, called Collective Probabilistic Labeling (CPL) is presented to update clusters incrementally. This approach makes the updating phase noise-resistant and fast. An important step in the updating is the merging of new clusters with existing ones. To this end, a new distance measure is proposed, which is a modified Kullback–Leibler distance between two GMMs.

In most real-time anomaly detection applications, incoming instances are often similar to previous ones. In these cases, there is no need to update clusters based on duplicates, since they have already been modeled in the cluster distribution. The two-layer structure is responsible for identifying redundant instances. Ignoring redundant instances, which are typically in the majority, makes the detection phase faster.

The proposed method is found to lower the false alarm rate, which is one of the basic problems for the one-class SVM. Experiments show the false alarm rate is decreased from 5% to 15% among different datasets, while the detection rate is increased from 5% to 10% in different datasets with two-layer structure. The memory usage for the two-layer structure is 20 to 50 times less than that of one-class SVM. The one-class SVM uses support vectors in labeling new instances, while the labeling of the two-layer structure depends on the number of GMMs. The experiments show that the two-layer structure is 20 to 50 times faster than the one-class SVM in labeling new instances. Moreover, the updating time of the two-layer structure is two to three times less than for a one-layer structure. This reduction is the result of using two-layer structure and ignoring redundant instances.

* Corresponding author.
E-mail address: braahemi@uottawa.ca (B. Raahemi).

https://doi.org/10.1016/j.ins.2017.11.023
0020-0255/

1. Introduction

Anomaly detection systems must not only find previously known anomalies, but also new unknown anomaly patterns [6]. Most anomaly detection approaches are signature-based, a method which is not able to detect new attacks. An issue in anomaly detection is a lack of labeled data. In all applications, what is mainly known are normal behaviors, along with a limited number of anomalous behaviors. Because of this limitation, supervised methods are not applicable; instead, unsupervised and semi-supervised methods are better alternatives.

In cluster-based approaches for anomaly detection, normal behaviors are modeled as a set of clusters, without requiring any previous knowledge of anomalous instances. The challenge is determining whether the new instances should be classified as normal or anomaly. Also, in many applications, finding a rigid boundary for a normal class is difficult, and current approaches do not preserve the exact shape or even a good approximation of the cluster. Moreover, normal patterns change over time, and a roster of recognized normal patterns may not be valid in the future. This introduces a requirement for an incremental structure to able to update normal patterns. Another challenge in this area is the presence of noise. Although anomalous instances differ from normal ones, they do not show the random behavior of noise instances. Noise is random behavior in data that does not follow any pattern, while anomalous behaviors are not random and they follow specific patterns [12]. A major responsibility of anomaly detection methods is to mitigate the effect of noise on system performance.

In this paper, a two-layer cluster-based structure is presented to deal with the aforementioned problems. The structure of the paper is as follows. In Section 2, anomaly detection methods and recent research advances in this area are reviewed. The methods are grouped into six categories, and the most influential approaches in each category are described in detail. In Section 3, the general structure and the components of two-layer cluster-based anomaly detection method are introduced. In Section 4, the Summarization based on Gaussian Mixture Model (SGMM) is explained in detail. Section 5 presents the collective labeling step, done to label and update clusters, which is one of the main parts of the two-layer structure. Collective labeling is the GMM-based incremental approach to update clusters. Issues and problems associated with collective labeling are also discussed in this section. Section 6 synthesizes the components of the two-layer structure, and presents the general two-layer structure based on the proposed approaches from previous sections. The capability to ignore redundant instances, and remove noise effects, are also discussed in this section. The experimental results presented in Section 7 cover extensive testing of the two-layer structure. Finally, the conclusion and proposals for future work are presented in Section 8.

2. Literature review

There are six main categories of anomaly detection algorithms: rule-based, statistical, proximity-based, Artificial Immune System (AIS), supervised and unsupervised methods.

2.1. Rule-based methods

In rule-based methods, first, a set of rules for detecting anomalous behavior is extracted. If newly-captured behavior fits one of the rules, it is considered to be an anomaly [22]. A major drawback is that these methods are not capable of detecting previously unseen anomalies. Moreover, the rule-based methods mainly rely on an expert's opinion, which may not be accurate, due to limited knowledge about any new anomalous behaviors. In the detection of new anomalous instances, a database of all normal rules has to be searched to find a matching case, which can take considerable time.

2.2. Statistical methods

The assumption of statistical methods is that normal behaviors are generated based on a stochastic model. With statistical models, anomalies are in the low-probability regions. Thatte et al. use packet-size statistics and traffic rates to build a statistical model, and then employ the sequential probability ratio test (SPRT) in the detection phase [25]. To find a statistical model for data, two sets of techniques are applied to the dataset: parametric techniques, non-parametric. The Gaussian-based model is a parametric statistical techniques. Using a Gaussian distribution for data, a simple statistical test for anomaly detection is the box plot test [15]. Non-parametric techniques do not have any predefined assumptions on the dataset. These methods work with the original data, with no predefined distribution model such as histogram-based methods. Another non-parametric technique is the Kernel function-based method, which is based on Parzen window estimation. Probability Distribution Function (PDF) for normal instances is estimated using Kernel functions. Any new instance in the low probability area is deemed anomalous [7].

2.3. Proximity-based methods

Proximity-based methods assume that normal instances are in a dense area of data, while anomalous instances are in sparse regions, and far from dense regions. Proximity-based methods can be classified into two groups: distance-based and density-based methods.

Distance-based methods assume that abnormal instances are remote from normal instances, based on a distance measure. Distance-based methods are of two main types: K-Nearest-Neighbors-based (KNN)[20] and grid-based methods. KNN is time-consuming and higher-level structures like hyper-grid have been tried to make it faster [28].