Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 96 (2016) 139 – 146

Procedia Computer Science

20th International Conference on Knowledge Based and Intelligent Information and Engineering Systems

The Entropy and PCA Based Anomaly Prediction in Data Streams

Daocheng Hong[a], Deshan Zhao[a]*, Yanchun Zhang[a,b]

[a]*Shanghai Key Laboratory of Intelligent Information Processing & Shanghai Key Laboratory of Data Science, School of Computer Science, Fudan University, Shanghai, China*

[b]*Centre for Applied Informatics, Victoria University, Melbourne, Australia*

Abstract

With the increase of data and information, anomaly management has been attracting much more attention and become an important research topic gradually. Previous literatures have advocated anomaly discovery and identification ignoring the fact that practice needs anomaly detection in advance (anomaly prediction) but anomaly detection with post-hoc analysis. Given this apparent gap, this research proposes a new approach for anomaly prediction based on PCA (principle component analysis) and information entropy theory, and support vector regression. The main idea of anomaly prediction is to train the historical data and to identify and recognize outlier data according to previous streams patterns and trends. The explorative results of SO_2 concentration of exhaust gas in WFGD (Wet Flue Gas Desulfurization) demonstrate a good performance (efficient and accurate) of the target data prediction approach. This robust and novel method can be used to detect and predict the anomaly in data streams, and applied to fault prediction, credit card fraud prediction, intrusion prediction in cyber-security, malignant diagnosis, etc.

Peer-review under responsibility of KES International

Keywords: data streams analysis; anomaly prediction; data management

1. Introduction

In this modern era of Internet + with the online and offline depth fusion, the data assimilation process has changed significantly into the form of data streams. The data stream is a continuous, unbounded sequence of data

* Corresponding author. Tel.: +86+21+65654549; fax: +86+21+65654253.
E-mail address: 15210240108@fudan.edu.cn, hongdc@fudan.edu.cn

Peer-review under responsibility of KES International
doi:10.1016/j.procs.2016.08.115

points accompanied special characteristics, such as transiency, uncertainty, dynamic data distribution, multidimensionality, and dynamic relationship. The arrival rate of data stream is usually high and the distributions of data stream often change over time. The most but not the last point, these multiple data streams are not independent. Actually, these data streams frequently demonstrate high correlations with each other in the latent layer. Furthermore, much useful information and knowledge is lost if each stream only is analyzed individually. Outlier or anomaly detection refers to automatic identification and recognization of unforeseen or abnormal phenomena embedded in a large amount of normal data [1].

Predicting anomalies or outliers can be more useful than finding common patterns in data streams, even it's really hard to predict anomalies exactly. It is desirable to find particular aspects of the current stream which are indicative of events of significance to the future event. One of most attractive application scenarios of anomaly prediction is when multiple data streams are targeted since it provides significant information for these applications. For instance, anomaly prediction in medical stream data, for example the electrocardiogram (ECG) signals, can save lives, identify outliers in disease event provides useful information about the possibility of occurring outbreaks. And in industrial processes, anomaly prediction may help to diagnose the incident faults and promote the performance for the organization.

Well known abnormalities can be modeled and detected according to the literatures, but unforeseen problems are not defined and hence are much harder to detect even in a single data stream. In this particular study, we understand that for many scenarios, it is more meaningful to predict abnormal for multiple data streams instead of finding anomaly in individual stream. More specifically, our goal is to monitor multiple data streams so that we observe the value of each stream at every time-tick and to automatically detect anomalies for targeted data stream. The main contributions to the data streams and anomaly management include:

(1) We propose a new approach for anomaly prediction based on the main idea of training the historical data and to identify and recognize outliers according to previous streams patterns and trends which provide a novel and integrated perspective both for academia and industry practice.

(2) We design the features extraction algorithm of anomaly detection based on PCA (principle component analysis) and information entropy theory which is applied to anomaly prediction with support vector regression in data streams.

(3) Within the proposed method, we conduct an experimental study (cooperative research between Fudan University and ABC Company) on the SO_2 concentration prediction of exhaust gas in WFGD (Wet Flue Gas Desulfurization) of power plant which shows a good performance (efficient and accurate), and illustrates that the new approach of outlier prediction can be extended to other anomaly management.

After the introduction, we will present the theoretical concepts such as definitions of the theoretical terms and relevant research. And then the novel approach of anomaly prediction is delineated with strong and rigorous logic. After that we will apply this new method to the SO_2 concentration prediction of exhaust gas in WFGD of power plant and verify the effectiveness. Finally, the future research directions and conclusions are drawn for both academia and industry practice.

2. Related Work

2.1. Anomaly Analysis Techniques

Anomaly analysis aims to detect a small set of observations that deviate considerably from other observations in data [1]. Anomaly detection has been applied to various applications, such as fault detection, credit card fraud detection, intrusion detection of cyber space, and malignant diagnosis [2-5]. Practically, there is only a small amount of labeled data available for the real world applications. Therefore, many researchers of data mining and machine learning communities have been interested in and devoted to the detecting anomaly of unseen data [6-9]. Many anomaly analysis techniques have been proposed in the related works [9–17]. These studies can be divided into three groups roughly: statistical, distance and density-based techniques for anomaly detection. The statistical techniques presume that the data has some predetermined distributions, and then they use the deviation of instance from these predetermined distributions to find the anomaly. However, most distribution models are presumed