Knowledge-Based Systems 61 (2014) 89–97

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Graph-based approach for outlier detection in sequential data and its application on stock market and weather data

Ali Rahmani [a], Salim Afra [a], Omar Zarour [a], Omar Addam [a], Negar Koochakzadeh [a], Keivan Kianmehr [a], Reda Alhajj [a,b,*], Jon Rokne [a]

[a] Department of Computer Science, Global University, Beirut, Lebanon
[b] Department of Computer Science, University of Calgary, Calgary, Alberta, Canada

ARTICLE INFO

Article history:
Received 15 January 2012
Received in revised form 5 February 2014
Accepted 15 February 2014
Available online 4 March 2014

Keywords:
Data mining
Graph-based algorithm
Outlier detection
Weather data
Stock market

ABSTRACT

Outlier detection has a large variety of applications ranging from detecting intrusion in a computer network, to forecasting hurricanes and tornados in weather data, to identifying indicators of potential crisis in stock market data, etc. The problem of finding outliers in sequential data has been widely studied in the data mining literature and many techniques have been developed to tackle the problem in various application domains. However, many of these techniques rely on the peculiar characteristics of a specific type of data to detect the outliers. As a result, they cannot be easily applied to different types of data in other application domains; they should at least be tuned and customized to adapt to the new domain. They also may need certain amount of training data to build their models. This makes them hard to apply especially when only a limited amount of data is available. The work described in this paper tackle the problem by proposing a graph-based approach for the discovery of contextual outliers in sequential data. The developed algorithm offers a higher degree of flexibility and requires less amount of information about the nature of the analyzed data compared to the previous approaches described in the literature. In order to validate our approach, we conducted experiments on stock market and weather data; we compared the results with the results from our previous work. Our analysis of the results demonstrate that the algorithm proposed in this paper is successful and effective in detecting outliers in data from different domains, one financial and the other meteorological.

1. Introduction

Outliers can be informally defined as patterns in data that do not conform to the expected behavior [7] or observations which appear to be inconsistent with the remainder of the given set of data instances [23]. Outlier patterns are also referred to as anomalies. In this paper, we use these two terms interchangeably. We argue that outliers are different from noise, though some authors classify both under the same category as noise. While noise represents instances that do not belong to the specific domain of knowledge, outliers are instances that do belong to the investigated domain but represent exceptional cases that do not fit within the general trend as the rest of the data. For instance, negative or non-numeric value for the age is considered a noise but a value over say 50 years may be considered outlier as the age of a daytime student because normally students are under 35 even those attending the PhD program. On the other hand, evening courses are mostly attended by older students and it is rare to find someone under 20 attending an evening class. So, outliers are context sensitive and hence should be studied from this point of view.

Anomaly detection is an important problem which finds extensive applications in various domains, e.g., network intrusion detection [16], credit card fraud detection [5], industrial damage detection [11], and community-based graph networks and graph streams [34,35], among others. Recently, it has been widely used in the area of spatial data analysis. An example of anomaly detection application in the spatial data analysis area is finding anomalies in meteorological data [18,29]. Finding such anomalies is of vital importance because they are frequently associated with abnormalities like hurricanes, tornados, cyclones, etc. Such disasters are becoming more destructive and even more common with the global warming and its consequences. Timely locating such anomalies in meteorological data could help in taking precautions that would lead to saving lives (which is a vital humanitarian concern) and may be minimize the loss in economy.

In addition to the various interpretations of anomalies in miscellaneous application domains, there are different types of

* Corresponding author. Tel.: +1 403 210 9453.
E-mail addresses: alhajj@ucalgary.ca, rsalhajj@gmail.com (R. Alhajj).

http://dx.doi.org/10.1016/j.knosys.2014.02.008
0950-7051/

anomalies. For instance, Chandola et al. [7] classify anomalies into three categories: point anomalies, collective anomalies, and contextual anomalies. Point anomalies are data instances that are anomalous with respect to the rest of the data. Collective anomalies are collections of data instances that are not anomalies by themselves, but their occurrence together in one collection is anomalous as a group. Contextual anomalies are data instances that are anomalous in their specific context, but not otherwise. In this article, we will focus on contextual anomalies and we will describe the algorithm which we have developed to detect this kind of anomalies. Nonetheless, the algorithm can also be applied with different parameters to detect point anomalies; this has been left as future work.

In order to clarify the notion of context, we should define each data instance as two sets of attributes. The first set contains attributes that determine the context or neighborhood of a data instance. For example, time attribute in time-series data is a contextual attribute. The second set contains behavioral attributes that define the non-contextual characteristics of an instance. For instance, in stock market time series data, stock price is a non-contextual attribute.

The goal of our work described in this paper is to develop an algorithm to find contextual outliers in various types of sequential data. We propose an approach that models data instances as the vertices of a graph and sets up edges between the vertices based on the Euclidean distance between the data instances. Here it is worth noting that the choice of the distance function to be used in the analysis is domain related, i.e., the characteristics of the data to be analyzed dictate whether to use Euclidean or Non-Euclidean distance function. Then, we utilize a *Minimum Spanning Tree* (*MST*) based clustering technique to cluster the nodes of the graph. Finally, we apply a voting scheme to detect contextual outliers in the analyzed data.

One of the distinguished characteristics of the proposed algorithm is that it does not need training data to build its model. In other words, it is a new type of algorithm that does not expect the existence of labeled data for training. This is an essential characteristics that eliminates the need for the domain expertise required to prepare the training data. Furthermore, unlike many other outlier detection techniques, our approach does not detect outliers based on the periodic patterns in data. These characteristics are important especially when no training data is available or no periodic patterns could be found in the sequence of data to be analyzed. Even though training data is not required, parameters of the algorithm should still be set with suitable values. These values can be found either by doing some analysis on the data or by inquiring a domain expert. But setting the parameters does not require the time and effort needed for preparing the training data.

In order to verify the correctness and applicability of our approach to various domains, we have chosen two unrelated domains, namely finance and meteorology. We conducted experiments on weather and stock market data which are characterized by the existence of outliers as vital indicators within the data. We compared the results with our previous work on weather data. We argue that this comparison is enough to demonstrate the effectiveness of the proposed approach because our previous work was illustrated as better approach compared to those described in the literature. Our analysis of the results showed that the performance of our algorithm is satisfactory.

The rest of the paper is organized as follows. Section 2 covers the related work though there is no work in the literature which handles outliers the same way as our proposed algorithm. Section 3 formulates the contextual outlier detection problem. Section 4 presents an exhaustive explanation and running time analysis of our algorithm. The experimental results are fully described in Section 5.2. Section 6 is conclusions and future work.

2. Related work

Outlier detection related research is rooted in statistics, and there is a huge amount of the literature that describes outlier detection algorithms, e.g., [2,4,9,20,30]. Shekhar et al. [23] broadly classify these algorithms into two categories: set-based outlier detection methods and spatial-set-based outlier detection methods. Set-based outlier detection algorithms consider the statistical distribution of attribute values and use statistical outlier detection tests to extract the outliers. They completely ignore the contextual relationships among items. On the other hand, spatial-set-based outlier detection methods consider both contextual and behavioral attributes. They use concepts such as distance, density, and convex-hull depth to identify outliers. Shekhar et al. [24] pointed out that multidimensional approaches do not consider graph structure of the spatial data and do not consider priori information of the statistical distribution of the data.

Chandola et al. [7] provide a more comprehensive classification of the outlier detection techniques. They classify the techniques into six categories: classification-based [15], nearest neighbor-based [27], clustering-based [12], statistical [10], information theory based [17] and spectral theory based [1]. In addition to the mentioned algorithms, other approaches have been adopted to solve the outlier detection problem. For instance, signal processing techniques such as wavelet transform [3] and Fourier transform [21] have been used to detect outlier regions in meteorological data.

Cheng and Li [8] extended spatial outlier mining e.g., [25,26] to capture the semantic and dynamic aspects of spatio-temporal data in multi-scales. They adopted a multi-resolution clustering algorithm based on semantic knowledge of spatio-temporal objects and applied them to find outliers in multi-scale properties of geographic phenomena. Birant and Kut [6] improved the DBSCAN algorithm to capture the temporal aspects of ST-objects and introduced a scale to measure the density of each cluster. Ramachandran et al. [19] developed a flexible framework ADaM (Algorithm Development and Mining) to mine large scientific data for meteorological phenomena detection and feature extraction. The architecture offers a package of a compact data mining system along with the flexibility of porting a user defined complex mining algorithm.

Yu et al. [28] introduced a multidimensional based approach called *FindOut*, which uses wavelet transform to remove the clusters from the original data and then identifies the outliers. They proposed a method named *WaveCluster*, which quantizes the feature space, and spatial data objects are represented in *n*-dimensional feature space. Wavelet transform is applied on the quantized space to find in the feature space dense regions forming clusters. The clusters can be of arbitrary shapes – convex, concave or nested clusters. The filtering property of wavelet transform removes the noise from the feature space, and the multi-resolution property identifies the clusters at different degrees of detail. Yu et al. [28] showed that the whole procedure is very fast and indicated that parallelism can bring further speed up. Hung and Cheung [13] introduced parallelism of the Nested-Loop (NL) algorithm [14] for distance-based outlier mining. They developed an efficient version (ENL) of NL, which reduces the cost by half in terms of computation and disk I/O costs.

Although many techniques have been developed to detect point anomalies, only a few of them support contextual anomaly detection. Two generic approaches have been proposed to solve the problem of contextual anomaly detection. The first approach identifies a context for each data instance using its contextual attributes, and then computes its anomaly score with respect to its behavioral attributes using a point anomaly detection algorithm. The second approach tries to exploit the structure in the data in