Information Sciences 285 (2014) 162–180

Contents lists available at ScienceDirect

Information Sciences

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

Dealing with temporal and spatial correlations to classify outliers in geophysical data streams

Annalisa Appice [a,*], Pietro Guccione [b], Donato Malerba [a], Anna Ciampi [a]

[a] Dipartimento di Informatica, Università degli Studi Aldo Moro di Bari, via Orabona, 4, 70125 Bari, Italy
[b] Dipartimento di Ingegneria Elettrica ed Informazione, Politecnico di Bari, via Orabona, 4, 70125 Bari, Italy

ARTICLE INFO

Article history:
Received 28 February 2013
Received in revised form 15 October 2013
Accepted 14 December 2013
Available online 24 December 2013

Keywords:
Time series forecasting
Temporal correlation
Spatial correlation
Outlier detection and classification

ABSTRACT

Anomaly detection and change analysis are challenging tasks in stream data mining. We illustrate a novel method that addresses both these tasks in geophysical applications. The method is designed for numeric data routinely sampled through a sensor network. It extends the traditional time series forecasting theory by accounting for the spatial information of geophysical data. In particular, a forecasting model is computed incrementally by accounting for the temporal correlation of data which exhibit a spatial correlation in the recent past. For each sensor the observed value is compared to its spatial-aware forecast, in order to identify the outliers. Finally, the spatial correlation of outliers is analyzed, in order to classify changes and reduce the number of false anomalies. The performance of the presented method is evaluated in both artificial and real data streams.

1. Introduction

The widespread dissemination of sensor networks has led data stream research to focus on the deployment of network-integrated surveillance systems. These systems should gather continuous, evolving data from several sensors, recognize and, possibly, adapt a behavioral model and deal with outliers when data do not conform the normal model.

Outlier detection is a challenging task, extensively investigated in stream data mining. Its main difficulty is in distinguishing between outliers and change points [39]. The former are isolated and exceptional cases, while the latter are the watershed of a changing data distribution. A model should be adapted when change points are met, so to account for new emerging behaviors of a system. On the contrary, anomalies defined by outliers should not be included in a general model.

Most of the studies on outlier or change detection consider only the temporal dimension of the data. In this paper, we investigate geophysical applications where data (observations of random fields) are characterized by both a temporal and a spatial dimension. In particular, we consider the scenario of fixed-to-ground sensors which routinely and continuously sample data for a geophysical numeric field. Some recent studies have faced the problem of detecting outliers or changes in spatio-temporal data, so they seem to fit the specific task at hand. Nevertheless, they all neglect the distinction between anomalies and changes. Therefore, they are not appropriate for evolving phenomena. Our approach to the problem, as proposed in this paper, takes a holistic view where both the involved dimensions (spatial and temporal) and the type of abnormal behavior (due to anomalies or changes) are equally taken into account. This view is based on the concept of autocorrelation.[1]

* Corresponding author. Tel.: +39 080 5443262; fax: +39 080 5443269.
E-mail addresses: annalisa.appice@uniba.it (A. Appice), guccione@poliba.it (P. Guccione), donato.malerba@uniba.it (D. Malerba), aciampi@di.uniba.it (A. Ciampi).

[1] In the present study we do not consider multivariate data, i.e. distinct fields can be dealt with separately. For univariate data, the distinction between different forms of correlations is blurred, and we can indifferently use either the terms spatial/temporal autocorrelation or the terms spatial/temporal correlation.

0020-0255/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ins.2013.12.009

We begin observing that the spatial location of a sensor inspires inferences on the *spatial correlation* of the data. This is a measure of how many data, taken at a relatively close location, behave similarly to each other. Similar behavior, which is very common in the geophysical field, is well described by the first law of Geography of Tobler [41], according to which "everything is related to everything else, but near things are more related than distant things". The time extent of the sampling operation also encourages the consideration of the *temporal correlation* of the data. This is a measure of how many future observations can be predicted from past behavior. Inferences based on temporal correlation require a stability (a property known as stationarity) of the statistical properties of the random field. However, in the geophysical context, data are frequently subject to the temporal variation of such properties. In this case, the distribution of a field can undergo a time change [17] to be taken into account in the modeling phase.

In this paper, we carry out the idea of modeling simultaneously the influence that both spatial and temporal correlation may have on a geophysical field. We illustrate a semi-supervised outlier detection method, called SWOD (Sliding Window Outlier Detector), that builds a spatio-temporal model of data obtained from the sliding window history and uses this model to detect and classify outliers. This model-based approach guarantees a forecasting service, in addition to outlier detection and classification.

The method includes an incremental modeling phase and an outlier detecting phase. The *modeling phase* updates (on-line) a cluster model of the sampling network. The clustering phase is data-driven by a spatial contiguity constraint on the sliding windowed data. Each cluster collects sensors whose observations are correlated in space in each row of the sliding window. Simultaneously, the behavior of a cluster is not necessarily stable over time, as observations can change over consecutive rows. The time series of the cluster prototypes (e.g. arithmetic mean) at each row describes the trend according to which intra-cluster data change over time. The temporal correlation in the trend time series is the knowledge used to estimate incrementally the coefficients of a forecasting model that is associated to the whole cluster. The *detection phase* is inspired by the idea that temporal correlation governs data. A cluster, whose data evolved similarly in the recent past, measures data that are somewhat similar at the present time. Therefore, similarly to the existing time series analysis methods [36,29,5], an outlier can be detected when data deviate greatly from its prediction. However, differently from these methods, spatial correlation is also accounted for as the classification of these deviations is spread on the set of nearby sensors.

This spatial-aware analysis of outliers is motivated by a common understanding of the nature of outliers and data changes in geophysical applications. While spatially isolated outliers (even when prolonged in time) are realistically anomalies because of incorrect hardware design, improper calibration or low battery level [32,36], data changes rarely involve single sensors. On the other hand, anomalies, also when they occur in groups, are not expected to be correlated in space. Based upon these considerations, we can assume that spatially correlated outliers pinpoint a pervasive change of the data, while uncorrelated (even if spatially close) outliers are anomalies.

We underline that this paper is based on the concept of trend clusters, which appears in several recent studies [7,8]. Nevertheless these studies concern two data mining tasks, namely summarization [7] and interpolation [8], which are different from the outlier classification task investigated in this study. Summarization compresses data and interpolation reconstructs past (unobserved) data. These studies resort to a count-based model to process the stream, so that trend clusters are discovered in non-overlapping windows without resorting to an incremental learning approach. The algorithm for incremental trend cluster discovery is originally introduced in [10]. However, none of these previous studies investigates the combination of trend cluster discovery and time series analysis for outlier detection and classification. This is a new contribution in this paper.

The paper is organized as follows. In Section 2 related works are revised. In Section 3 some basic definitions are reported. In Section 4, the proposed method is illustrated and in Section 5 experiments with artificial and real data are presented. Finally, conclusions are drawn.

2. Background and contribution

Several studies in recent literature address the problem of outlier detection, change detection and a combination of both problems.

2.1. Outlier detection

Outlier detection has been studied for decades in time series analysis (see [19] for a recent survey). Several methods use the semi-supervised learning schema. Fewer methods resort to the supervised or unsupervised schema. This is due to the fact that the supervised schema (e.g. [22,3]) comprises a training phase with labeled data, in order to learn the data model which classifies outliers and inliers. So it pays for the need to label "manually" the training outliers. This can be an expensive operation with massive data.

An alternative could be to build a fully labeled training data set by injecting "artificial" outliers into a normal data set [1]. However, in both cases, the labeling phase must be periodically repeated, in order to fit evolving data. In contrast, the unsupervised schema overlooks the request for labeled data by giving away any training phase (e.g. [21,32]) with the computation of an informative model of the data behavior. In this scenario, the semi-supervised schema stands out as it has a training