Environmental Modelling & Software 25 (2010) 1014–1022

Contents lists available at ScienceDirect

Environmental Modelling & Software

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

Anomaly detection in streaming environmental sensor data: A data-driven modeling approach

David J. Hill [a,*], Barbara S. Minsker [b]

[a] Department of Civil and Environmental Engineering, Rutgers University, 623 Bowser Rd, Piscataway, NJ 08854, USA
[b] Department of Civil and Environmental Engineering, University of Illinois Urbana-Champaign, 205 N. Mathews Ave., Urbana, IL 61801, USA

ARTICLE INFO

Article history:
Received 9 March 2009
Received in revised form
25 August 2009
Accepted 25 August 2009
Available online 24 October 2009

Keywords:
Coastal environment
Data-driven modeling
Anomaly detection
Machine learning
Real-time data
Sensor networks
Data quality control
Artificial intelligence

ABSTRACT

The deployment of environmental sensors has generated an interest in real-time applications of the data they collect. This research develops a real-time anomaly detection method for environmental data streams that can be used to identify data that deviate from historical patterns. The method is based on an autoregressive data-driven model of the data stream and its corresponding prediction interval. It performs fast, incremental evaluation of data as it becomes available, scales to large quantities of data, and requires no pre-classification of anomalies. Furthermore, this method can be easily deployed on a large heterogeneous sensor network. Sixteen instantiations of this method are compared based on their ability to identify measurement errors in a windspeed data stream from Corpus Christi, Texas. The results indicate that a multilayer perceptron model of the data stream, coupled with replacement of anomalous data points, performs well at identifying erroneous data in this data stream.

© 2009 Published by Elsevier Ltd.

1. Introduction

In-situ environmental sensors are sensors that are physically located in the environment they are monitoring. Through telemetry, the time-series data collected by these sensors can be transmitted continuously to a repository as a data stream. Recently, there have been efforts to make use of streaming data for real-time applications (e.g., Bonner et al., 2002). For example, draft plans for the Water and Environmental Research Systems (WATERS) Network, a proposed national environmental observatory network, have identified real-time analysis and modeling as a significant priority (NRC 2006).

Because *in-situ* sensors operate under harsh conditions, and because the data they collect must be transmitted across communication networks, the data can easily become corrupted. Undetected errors can significantly affect the data's value for real-time applications. Thus, the NSF (National Science Foundation), 2005 has indicated a need for automated data quality assurance and control (QA/QC). Anomaly detection is the process of identifying data that deviate markedly from historical patterns (Hodge and Austin, 2004). Anomalous data can be caused by sensor or data transmission errors or by infrequent system behaviors that are often of interest to scientific and regulatory communities. In addition to data QA/QC, where data anomalies may be the result of sensor or telemetry errors, anomaly detection has many other practical applications, such as adaptive monitoring, where anomalous data indicate phenomena that researchers may wish to investigate further through increased sampling, and anomalous event detection, where anomalous data signal system behaviors that require other actions to be taken, for example in the case of a natural disaster. These applications require that data anomalies be identified in near-real time; thus, the anomaly detection method must be rapid and be performed incrementally to ensure that detection keeps up with the rate of data collection.

Traditionally, anomaly detection has been carried out manually with the assistance of data visualization tools (Mourad and Bertrand-Krajewski, 2002), but manual methods are unsuitable for real-time detection in streaming data, since they necessitate an operator to be performing analysis 24 h a day, 7 days a week. More recently, researchers have suggested automated statistical and machine learning approaches, such as minimum volume ellipsoid (Rousseeuw and Leroy, 1996), convex pealing (Rousseeuw and Leroy, 1996), nearest neighbor (Tang et al., 2002; Ramaswamy et al., 2000), clustering (Bolton and Hand, 2001), neural network classifier

* Corresponding author. Tel.: +1 217 714 3490.
E-mail addresses: ecodavid@rci.rutgers.edu (D.J. Hill), minsker@illinois.edu (B.S. Minsker).

1364-8152/$ – see front matter © 2009 Published by Elsevier Ltd.
doi:10.1016/j.envsoft.2009.08.010

(Kozuma et al., 1994), support vector machine classifier (Bulut et al., 2005), and decision tree (John, 1995). These methods are faster than manual methods, but they have drawbacks that make them unsuitable for real-time anomaly detection in streaming data. Minimum volume ellipsoid and convex peeling require all of the data to have accumulated before anomalies can be identified. Nearest neighbor, clustering and support vector machines are computationally intractable for large quantities of data and neural network classifier, support vector machine classifier and decision tree require pre-classified (anomalous/non-anomalous) data, which characterize all anomalies that may be encountered. Since real-time sensors collect data continuously, and the data are to be used in real time, decisions based on the totality of the data cannot be made. Instead, anomaly classifications must be made based on the data that have been collected up to the current point in time.

Several researchers have suggested anomaly detection methods specifically designed for real-time detection in streaming data. These methods are often referred to as analytical redundancy methods because they employ a model of the sensor data stream as a simulated redundant sensor whose measurements can be compared with those of the actual sensor. The classification of a measurement as anomalous is based on the difference between the model prediction and the sensor measurement. Early work on such methods (e.g., Upadhyaya et al., 1990; Belle et al., 1983) employed multivariate autoregressive (MAR) models to predict the next measurement in the sensor data stream, using historical values. More recently, other regression approaches, such as artificial neural networks (ANNs), have been suggested (Nairac et al., 1999; Fantoni and Mazzola, 1996; Silvestri et al., 1994). These methods were designed for use in manufacturing/power plants, and the only researchers to have suggested a method of threshold selection (Belle et al., 1983) for classifying data as anomalous/non-anomalous required detailed process knowledge that is not generally available for natural systems. Krajewski and Krajewski (1989) present an analytical redundancy method for streamflow data that employs model error standard deviations to set the threshold. This method, however, relies on a physically-based real-time model of the natural system – a tool which may not always be readily available.

This study develops a real-time anomaly detection method that employs a data-driven univariate autoregressive model of the data stream and a prediction interval (PI) calculated from recent historical data to identify streaming data anomalies. The method used to calculate the PI in this study accounts for uncertainty in the data and in the parameters of the data-driven model. A data-driven time-series model is employed because it is simpler to develop than a physics-based time-series model and it can rapidly produce accurate short forecast horizon predictions. Data are classified as anomalous/non-anomalous based on whether or not they fall outside a given PI. Thus, the method provides a principled framework for selecting a threshold. This method does not require any pre-classified examples of data, scales well to large volumes of data, and allows for fast incremental evaluation of data as it becomes available. The Section 2 describes this method in detail. Next, it is tested through a case study, in which several types of data-driven models are used to identify erroneous measurements in a windspeed data stream from the WATERS Network Corpus Christi Bay testbed, provided by the Shoreline Environmental Research Facility (SERF) (http://www.serf.tamus.edu). Finally, the results of the different instantiations of the anomaly detection method are compared, and implications of the different modeling strategies are discussed.

2. Methods

This study proposes a new analytical redundancy method for anomaly detection that uses a moving window of q sensor measurements (or their expected values as explained shortly) $D^t = \{x_{t-q+1}, \ldots, x_t\}$ to classify the next chronologically sequential sensor measurement x_{t+1}. A measurement is classified as anomalous if it deviates significantly from the one-step-ahead prediction of its value calculated using D^t as input. Upon initialization, the method fills the window with the q most recent sensor measurements and commences classification with the next measurement taken by the sensor.

In brief, the method consists of the following steps beginning at time t: (1) use a one-step-ahead prediction model that takes D^t as input to predict $\bar{x}_{t+1}$, the expected value of the sensor measurement at time $t+1$; (2) calculate the upper and lower bounds of the range within which the sensor measurement should lie (i.e., the prediction interval) with probability p; (3) when the measurement at time $t+1$ arrives from the sensor, compare the sensor measurement with this range, and if it falls outside the range, classify it as anomalous, otherwise classify it is non-anomalous; (4a) under the anomaly detection and mitigation (ADAM) strategy, if the measurement is classified as anomalous modify D^t by removing x_{t-q+1} from the back of the window and adding $\bar{x}_{t+1}$ to the front of the window to create D^{t+1}; (4b) under the anomaly detection (AD) only strategy, modify D^t by removing x_{t-q+1} from the back of the window and adding x_{t+1} to the front of the window to create D^{t+1}; (4) repeat steps 1–4. This process is illustrated with a flow chart in Fig. 1. The remainder of this section describes these steps in detail.

2.1. Step 1: One-step-ahead prediction

In the first step of the anomaly detection process, a univariate autoregressive model of the sensor data stream is used to predict the next measurement in the data stream. Univariate autoregressive models are models that predict future measurements in a sensor data stream using only a specified set of previous measurements from the same sensor; they are used because they simplify the anomaly detection process in several ways. First, using only previous values of the same data stream avoids complications caused by different sampling frequencies that can arise if a heterogeneous set of sensor data streams were used. Second, because of the frequency with which embedded environmental sensors go offline, a significant number of gaps of undefined duration exist within each of the sensor data streams, and when comparing the streams, these gaps usually do not correspond with the same time periods. Since time-series models require a defined set of input variables, it is unclear how to make predictions if one or more of the input variables is not available; thus, the use of autoregressive models reduces the number of predictions that cannot be made due to insufficient data. Finally, since anomalous data cannot be expected to produce reasonable predictions when used as inputs into a model, if data from more than one sensor are used for prediction of a particular data stream, then anomalous data from different sensors must be detected in a particular order, such that all of the independent variables of a model have already been processed. For example, if the model for data stream A requires the most recent measurements from data stream B, then anomaly detection must first be performed on data stream B, before it can be performed on data stream A. The use of autoregressive models allows anomaly detection of several sensor data streams to take place rapidly and in an arbitrary order.

The moving window $D^t = \{x_{t-q+1}, \ldots, x_t\}$ is used as input to the autoregressive model of the data stream to predict the next (i.e., one-step-ahead) sensor measurement:

$$\bar{x}_{t+1} = M(D^t) + \varepsilon \tag{1}$$

where $M()$ is the model and ε is a random variable representing the model error. This method assumes that the behavior of the

Download English Version:

https://daneshyari.com/en/article/568704

Download Persian Version:

https://daneshyari.com/article/568704

Daneshyari.com