Performance Evaluation 67 (2010) 1059–1075

Contents lists available at ScienceDirect

Performance Evaluation

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

Online anomaly detection for sensor systems: A simple and efficient approach

Yuan Yao [a,*], Abhishek Sharma [b], Leana Golubchik [a,b], Ramesh Govindan [b]

[a] Department of Electrical Engineering-Systems, USC, Los Angeles, CA 90089, United States
[b] Department of Computer Science, USC, Los Angeles, CA 90089, United States

ARTICLE INFO

Article history:
Available online 13 August 2010

Keywords:
Anomaly detection
Sensor systems
Real-world deployments

ABSTRACT

Wireless sensor systems aid scientific studies by instrumenting the real world and collecting measurements. Given the large volume of measurements collected by sensor systems, one problem arises—an automated approach to identifying the "interesting" parts of these datasets, or *anomaly detection*. A good anomaly detection methodology should be able to accurately identify many types of anomaly, be robust, require relatively few resources, and perform detection in (near) real time. Thus, in this paper, we focus on an approach to *online* anomaly detection in measurements collected by sensor systems, where our evaluation, using real-world datasets, shows that our approach is accurate (it detects over 90% of the anomalies with few false positives), works well over a range of parameter choices, and has a small (CPU, memory) footprint.

1. Introduction

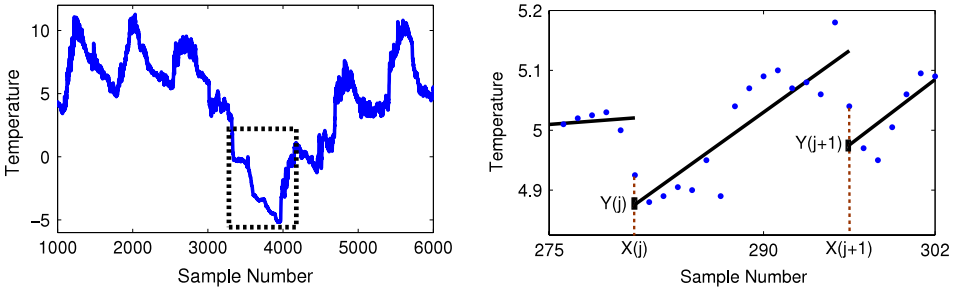
Wireless sensor systems have significant potential for aiding scientific studies by instrumenting the real world and collecting measurements, with the aim of observing, detecting, and tracking scientific phenomena that were previous only partially observable or understood. However, one obstacle to achieving the full potential of such systems is the ability to process, in a timely and meaningful manner, the huge amounts of measurements they collect. Given such large volumes of collected measurements, one natural question might be: *Can we devise an efficient automated approach for identifying the "interesting" parts of these datasets*? For instance, consider a marine biology application collecting fine-grained measurements in near real time (e.g., temperature, light, micro-organism concentrations)—one might want to rapidly identify "abnormal" measurements that might lead to algal blooms which can have devastating consequences. We can view the identification of such "interesting" or "unexpected" measurements (or events) in collected data as anomaly detection. In the remainder of the paper, we use the generic term "anomaly" for all interesting (typically, other-than-normal) events occurring either on the measured phenomena or the measuring equipment. Automated *online (or near real-time) anomaly detection in measurements collected by sensor systems* is the focus of this paper.

Anomalies can have a variety of lengths, magnitudes, and patterns. For instance, Fig. 1(a) depicts a long-duration, relatively gradual change in sensor reading, whereas Fig. 2(b) includes several short-duration, quite abrupt changes in sensor readings. Both scenarios correspond to anomalous events and should be accurately detected by an anomaly detection methodology.

Thus, a good anomaly detection methodology should have the following properties. First, it should be able to accurately identify all types of anomaly as well as normal behavior (i.e., it should have low false negative and false positive rates).

* Corresponding author.
E-mail addresses: yuanyao@usc.edu (Y. Yao), absharma@usc.edu (A. Sharma), leana@usc.edu (L. Golubchik), ramesh@usc.edu (R. Govindan).

0166-5316/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.peva.2010.08.018



(a) Long-duration anomaly. (b) Piecewise linear model for time series data.

Fig. 1. (a) Dataset with long-duration anomaly and (b) example of a piecewise linear model.

Second, it should be robust, i.e., the methodology should be relatively insensitive to parameter settings as well as pattern changes in the datasets. Third, it should require relatively small amounts of resources, as these are typically limited in sensor systems. That is, to run on sensor systems, it should ideally have low computational complexity, occupy little memory space, and require little transmission power. Last, it is also desirable for a detection algorithm to be able to detect anomalies in real time or near real time. This is particularly important for sensor systems corresponding to temporary deployments (as it might not be as useful to detect anomalies once the deployment is over) and those monitoring hazardous natural phenomena (e.g., the spread of contaminants in aquatic ecosystems), where prompt detection (and reaction) can be essential to reducing loss of life and money.

Anomaly detection, in general, has been studied in a number of system contexts, most notably in networking, where several techniques have been proposed for detecting network traffic anomalies [1–4]. While one might take the approach of adapting one (or more) of these techniques to sensor systems, we believe that they do not satisfy all the desirable properties described above, at least in their current form. In Section 6, we provide (a) quantitative results from applying network anomaly detection techniques to data collected by real sensor system deployments, and (b) intuition for why these techniques did not yield good results on such data. Consequently, the properties required of an effective anomaly detection method for sensor data and our experience with applying network traffic anomaly detection techniques to sensor measurements motivated us to explore methods different from those in prior work in network anomaly detection.

We also note that little exists in the literature on the topic of anomaly detection in sensor system data. Most efforts are focused on the detection of faulty sensor readings, such as those depicted in Fig. 3(a) and (b)—these are typically short-duration events, with values significantly deviating from the "normal" sensor readings [5]. Often, such sensor data faults are modeled as outliers and can be detected using simple rule-based approaches or by using statistical models to capture the pattern of normal sensor readings and flagging any significantly different samples as faulty [6]. In this work, we view faulty sensor readings as a special case of anomalies. As illustrated in Section 4, our approach is able to capture such faulty readings, as well as other long-duration, "gradual" anomalies such as the one depicted in Fig. 1(a).

To the best of our knowledge, the only efforts focused on anomaly detection in sensor system data are those reported in [7–9]. Briefly, [7,8] view measurements collected by a sensor system as coming from the same (unknown) distribution and they "pre-define" anomalies as outliers. The main focus of that effort, which is an *offline* approach, is on minimizing the communication overhead (in transmitting data needed for anomaly detection) and corresponding energy consumption. In contrast, we focus on an *online* approach that, on the fly, builds an adaptive model of "normal" data and does *not* a priori define what is an anomaly. For instance, the approach in [7,8] might only flag the most extreme measurement in Fig. 1(a) as an anomaly, whereas our approach would flag the entire event (outlined by the dashed rectangle) as an anomaly. We give a more detailed description of [7,8] and a quantitative comparison in Section 6. In [9], a change point detection-based approach is used for detecting distribution changes (e.g., mean, variance, covariances) in sensor measurements. However, (a) this approach assumes knowledge of the (time-varying) probability distribution from which sensor measurements are sampled (information often not available in real-world deployments), and (b) such techniques do not meet (at least in their current form) our efficiency criteria (see Section 6).

In this work, we formulate the problem of anomaly detection in sensor systems as an instance of the problem of identifying unusual patterns in time series data. Of course, one possible direction would then be to construct a time series-based approach, e.g., based on [6]. However, we also did not find this direction to be effective as such techniques are (typically) not well suited for detecting long-duration anomalies. So, we do not pursue this direction further here, but in Section 6, we do illustrate quantitative results corresponding to applying a representative time series-based approach to data collected by real sensor system deployments and provide intuition for why such a technique did not yield good results.

In contrast, the basic idea behind our approach is to compare the collected measurements against a reference time series. But, to do this efficiently and robustly, the following challenging problems need to be solved: (1) How do we define a reference time series? (2) How do we compare two time series efficiently? (3) What metric do we use in deciding whether two sensor data time series are similar or different? and (4) How do we update the reference time series, to adapt to (normal) changes in sensor data patterns?