In-network outlier detection in wireless sensor networks

Joel W. Branch · Chris Giannella · Boleslaw Szymanski · Ran Wolff · Hillol Kargupta

Received: 3 September 2009 / Revised: 4 August 2011 / Accepted: 29 November 2011 / Published online: 18 January 2012 © Springer-Verlag London Limited 2012

Abstract To address the problem of unsupervised outlier detection in wireless sensor networks, we develop an approach that (1) is flexible with respect to the outlier definition, (2) computes the result in-network to reduce both bandwidth and energy consumption, (3) uses only single-hop communication, thus permitting very simple node failure detection and message reliability assurance mechanisms (e.g., carrier-sense), and (4) seamlessly accommodates dynamic updates to data. We examine performance by simulation, using real sensor data streams. Our results demonstrate that our approach is accurate and imposes reasonable communication and power consumption demands.

J. W. Branch
Middleware and Application Transformation Department, IBM T. J. Watson Research Center, Hawthorne, NY 10532, USA
e-mail: branchj@us.ibm.com

C. Giannella (✉)
The MITRE Corporation, 300 Sentinel Dr Suite 600, Annapolis Junction, MD 20701, USA
e-mail: cgiannel@acm.org

B. Szymanski
Network Science and Technology Center and Department of Computer Science, Rensselaer Polytechnic Institute, Troy, NY 12180, USA
e-mail: szymansk@cs.rpi.edu

R. Wolff
Department of Information Systems, University of Haifa, Haifa, Israel
e-mail: rwolff@is.haifa.ac.il

H. Kargupta
Department of Computer Science and Electrical Engineering, University of Maryland, Baltimore County, Baltimore, MD 21250, USA
e-mail: hillol@cs.umbc.edu

H. Kargupta
AGNIK, LLC, Columbia, MD, USA
Keywords  Outlier detection · Wireless sensor networks · In-network computation

1 Introduction

Outlier detection, an essential step preceding most any data analysis routine, is used either to suppress or amplify outliers. Suppressing outliers (also known as data cleansing) improves the robustness of data analysis, for instance, in clustering [31], time series analysis [9], or text categorization [60]. Amplifying outliers helps find rare patterns in domains such as fraud analysis [47], intrusion detection, and Web purchase analysis.

Several factors make wireless sensor networks (WSNs) especially prone to outliers. First, they collect their data from the real world using imperfect sensing devices. Second, they are battery powered, and thus, their performance tends to deteriorate as power dwindles. Third, since these networks may include a large number of sensors, the chance of error accumulates. Finally, when used for security and military purposes, sensors are especially prone to manipulation by adversaries. Hence, it is clear that outlier detection should be an inseparable part of any data processing routine in WSNs.

Simply put, outliers are observations whose probability of occurrence is extremely small. Since the actual distribution of the data is usually unknown, direct computation of probabilities is difficult. Because the problem is fundamental, a huge variety of outlier detection methods have been developed. In this paper, we focus on non-parametric, unsupervised methods. A simplistic implementation of these methods would require collecting all data at one central node and executing the outlier detection algorithm there. Such a centralized solution has several disadvantages in WSNs [34]. The two most important are (i) the energy overhead incurred by sending the data across the WSN and (ii) the time delays incurred between data collection and processing. Both disadvantages increase proportionally to the average distance a data item needs to travel. They are also exacerbated by the dense flows arising in the areas close to the central node into which data from the entire WSN, by necessity of a centralized algorithm, converge. The delays increase in such areas because of likely packet collisions, so either the colliding packets need to be retransmitted or bandwidth sharing needs to be imposed. Additionally, the nodes in such areas transmit far more than the average number of messages and therefore expend their energy faster than other nodes.

We developed a technique for the computation of outliers in WSNs. This technique (1) is flexible with respect to the outlier definition, (2) computes the result in-network to reduce both bandwidth and energy consumption (see, e.g., [30]), (3) uses only single-hop communication, thus permitting very simple node failure detection and message reliability assurance mechanisms (e.g., carrier-sense), and (4) seamlessly accommodates dynamic updates to data. In addition to these essential features, the algorithm presented here has two highly desirable properties: it is generic—suitable for many outlier detection heuristics, and its communication load is proportional to the outcome (i.e., the number of outliers reported).

We exemplify the benefits of our algorithm by means of two different outlier detection heuristics and 53 sensors, which we simulate using the SENSE network simulator [20] on real data streams. Our results show that the algorithms both converge to an accurate result with reasonable communication load and power consumption. In most tested cases, our algorithms outperform the equivalent centralized algorithm.

The rest of the paper is organized as follows. In the next section, we provide a motivating example, in which a network of acoustic sensors attempts to locate the possible source of a sound. In Sect. 3, we discuss related work, including prior publications on outlier detection, wireless sensor networks, and distributed data mining. We introduce preliminaries in