Expert Systems with Applications 39 (2012) 9087–9096

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

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

Ensemble based sensing anomaly detection in wireless sensor networks

Daniel-Ioan Curiac *, Constantin Volosencu

Automation and Applied Informatics Department, "Politehnica" University of Timisoara, Bd. V. Parvan nr. 2, 300223 Timisoara, Romania

ARTICLE INFO

Keywords:
Wireless sensor networks
Binary classifier
Ensemble based systems
Sensors anomalies
Data accuracy

ABSTRACT

Wireless sensor networks are often used to monitor and measure physical characteristics from remote and sometimes hostile environments. In these circumstances the sensing data accuracy is a crucial attribute for the way these applications complete their objectives, requiring efficient solutions to discover sensor anomalies. Such solutions are hard to be found mainly because the intricate defining of the correct sensor behavior in a complex and dynamic environment. This paper tackles the sensing anomaly detection from a new perspective by modeling the correct operation of sensors not by one, but by five different dynamical models, acting synergically to provide a reliable solution. Our methodology relies on an ensemble based system composed of a set of diverse binary classifiers, adequately selected, to implement a complex decisional system on network base station. Moreover, every time a sensing anomaly is discovered, our ensemble offers a reliable estimation to replace the erroneous measurement provided by sensor.

1. Introduction

A wireless sensor network (WSN) is a collection of tiny, inexpensive and low-power devices that can be deployed throughout a geographical space for fine-grained monitoring and event detection. Besides its computing and communication potential, a WSN is, first of all, an advanced distributed measurement system which is often prone to sensing anomalies that can cause erroneous data, compromising the objectives of the entire network.

There are three major sources of sensor anomalies within WSN: (i) software or/and hardware failures – breakdowns in any subsystem of a sensor node or even battery discharging can produce wrong sensing data; (ii) security attacks – when a malicious entity compels the sensors to report erroneous measurements or to drop measurement data packages (Walters, Liang, Shi, & Chaudhary, 2006); (iii) environment related sources – when sensor nodes cannot measure the physical value correctly due to unfavorable phenomena that can arise in the harsh environments in which they are often deployed.

Generally speaking, sensing anomaly detection refers to the problem of discovering patterns in measurement data that do not match with expected behavior (Chandola, Banerjee, & Kumar, 2009; Rajasegarar, Leckie, & Palansiwami, 2008; An, Heo, & Chang, 2011). This is not a simple task mainly because an estimated model of "correct behavior" is always hard to find. In the case of a WSN there is an inherent feature on which we can rely – sensing redundancy (Curiac, Volosencu, Pescaru, Jurca, & Doboli, 2009), which takes two basic forms: physical redundancy that implies the use of more than one sensor node for measuring the same localized value; and analytical redundancy that implies a mathematical model for evaluating the value provided by one sensor by taking into consideration the past and present values of neighboring sensors (spatial redundancy), the past values of the sensor itself (temporal redundancy) or both (spatiotemporal redundancy).

In the last decade a series of relevant approaches based on assortments employing different types of analytical redundancies and intelligent detection algorithms have been proposed for solving the issue of sensing anomaly discovery.

In (Siripanadorn, Hattagam, & Teaumroong, 2010) a mixture between a competitive learning method called the self-organizing map (SOM) and the discrete wavelet transform (DWT) is used to detect anomalies from synthetic and real-world datasets.

A two-step temporal modeling procedure, developed to analyze and extract semantic symbols from a sequence of observations, is presented in Li, Thomason, and Parker (2010) where an intelligent system detects time-related changes online by using a likelihood-ratio detection scheme. The algorithm is distributed, and supports a hierarchical learning structure that can scale to large number of sensors.

The use of Bayesian networks as means for unsupervised learning and anomaly detection in gas monitoring sensor networks for underground coal mines is described in Wang, Lizier, Obst, Prokopenko, and Wang (2008). The authors showed that the Bayesian network model can learn cyclical baselines for gas concentrations, thus reducing false alarms usually caused by flatline thresholds. Their solution was proved to be efficient in both distributed and centralized approach.

* Corresponding author. Tel.: +40 256 403 227; fax: +40 256 403 214.
E-mail addresses: daniel.curiac@aut.upt.ro (D.-I. Curiac), constantin.volosencu@aut.upt.ro (C. Volosencu).

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2012.02.036

An efficient method applying principal component analysis (PCA) simultaneously on multiple metrics received from various sensors is depicted in Chatzigiannakis and Papavassiliou (2007). One of the key features of this approach is that it provides an integrated methodology of taking into consideration and combining effectively correlated sensor data, in a distributed fashion. Furthermore, it allows the integration of results from neighboring network areas to detect correlated anomalies that involve multiple groups of nodes.

Our paper tackles the sensing anomaly discovery from a new perspective: the modeling of the correct behavior for sensors is done not by one, but by five different models, acting synergically to provide a reliable solution. For this, we developed an ensemble based system (EBS) containing five different binary classifiers, each categorizing every network node as being accurate or erroneous, the final decision being taken by the entire ensemble using a voting procedure. It is broadly accepted that the overall efficiency of such committees of classifiers can occur only if there is diversity among its components (Polikar, 2006). In our proposal, the heterogeneity of classifiers is achieved by using various and carefully selected classifier architectures and different sets of input data. In order to completely solve the problem, our ensemble not only discovers sensing errors, but offers reliable estimations to replace the measurements affected by these anomalies.

The remainder of the paper is organized as follows. In Section 2, we introduce the philosophy of our ensemble based sensing anomaly detection. Section 3 describes the architecture of each individual classifier, pointing out the way in which the correct sensing behavior is modeled and predicted, while Section 4 depicts the decisional block of the ensemble based on weighted voting algorithm. In Section 5, we present the methodology used for training and testing of the ensemble. Section 6 covers a test case illustrating the entire methodology and, finally, conclusions are offered in Section 7.

2. Ensemble based sensing anomaly detection

Discovering sensing anomalies in the context of WSN is a challenging issue due to the complexity of the environment in which sensor nodes are deployed. Often, this subject is tackled using dedicated decisional systems. For acquiring node behavior related decisions, it makes sense to ask more than one decision making entity, because this practice assures indubitably a better, more informed, and trustable final decision. We label these decisional instances as classifiers or experts and their collections as ensemble based systems (Dietterich, 2000; Ho, Hull, & Srihari, 1994; Polikar, 2006; Wang, Hao, Ma, & Jiang, 2011).

In order to periodically detect and investigate each and every sensor anomaly, an ensemble based system has been designed.

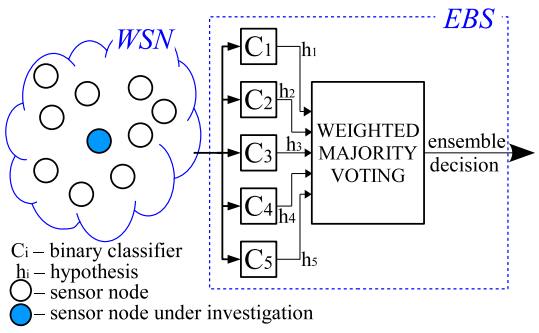

Fig. 1. Ensemble based sensing anomaly detection.

As presented in Fig. 1, this ensemble encloses several binary classifiers that independently categorize the state of each sensor as "reliable" or "unreliable". All the classifier outputs will be aggregated using a weighted majority algorithm to obtain the concluding ensemble decision. This final ensemble decision will be further used by the base station to take all the required actions for the unreliable nodes (e.g. removal of the untrustworthy nodes from the WSN for a specified period of time).

Our ensemble based methodology reveals the following characteristics:

- The EBS input data are represented by past and present measurements gathered by the node under investigation (node A) and respectively, past and present measurements gathered by each of the node A neighboring nodes.
- There are five binary classifiers involving different prediction techniques and different types of inputs to assure the required diversity of classifiers. Each classifier offers its own hypothesis h_i ("reliable" or "unreliable") about the sensing correctness of the node under investigation.
- The overall ensemble decision block is based on a weighted majority algorithm.
- If the investigated node is found as presenting sensor anomalies, the network base station acts in consequence and can exclude that specific sensor from the list of network functioning sensors for a limited period of time. As an example, this can be achieved based on the following rule: if the EBS indicated at least three times that the node A suffers from a sensor anomaly, the base station decides to inactivate the sensor. The base station could later reactivate the sensor after repeating the EBS investigation by testing if new readings became appropriate.
- Using the power of ensemble, a reliable estimation of the measurement affected by anomaly is automatically offered any time when needed.

In the following paragraphs the design of each EBS component, together with details regarding training and testing of the ensemble are offered.

3. Designing the classifiers

The design of individual classifiers to fulfill the EBS requirements is not a simple task. This process has to be governed by one magic word: diversity. As a result, any stratagem for generating the ensemble members must be focused on the ensemble's heterogeneity improvement.

The diversity of classifiers may originate from three basic sources:

a. classifier structure: the use of diverse classifier algorithms (Hsu & Srivastava, 2009; Tsoumakas, Katakis, & Vlahavas, 2004) can assure the required heterogeneity (diversity through structure);
b. classifier internal parameters: by using different training datasets (García-Pedrajas & Ortiz-Boyer, 2009; Kim & Kang, 2010; Li & Sun, 2012; Shirai, Kudo, & Nakamura, 2009) or diverse initializations of the training algorithms, the classifiers having exactly the same structure may cover different regions of the classified workspace, assuring the heterogeneity (diversity through parameters); and
c. classifier inputs: the diversity may be assured by applying different inputs to identical classifiers (diversity through inputs); usually this channel to obtain diversity is irrelevant, but when there is an overabundance of data sources it can be a viable alternative.