Expert Systems with Applications 40 (2013) 3248–3255

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

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

Data reconciliation in a smart home sensor network

Dorothy N. Monekosso[a,*], Paolo Remagnino[b]

[a] Faculty of Computing and Engineering University of Ulster, Shore Road, Co. Antrim, BT37 0QB, UK
[b] Faculty of Science, Engineering and Computing, Kingston University, Penrhyn Road, KT1 2EE, UK

ARTICLE INFO

Keywords:
Sensor data analysis
Error detection
Principal component analysis
Canonical correlation analysis

ABSTRACT

This paper describes a data-driven approach to sensor data validation. The data originates from a network of sensors embedded in an indoor environment such as an office, home, factory, public mall or airport. Data analysis is performed to automatically detect events and classify activities taking place within the environment. Sensor failure and in particular intermittent failure, caused by electrical interference, undermines the inference processes. PCA and CCA are compared for detecting intermittent faults and masking such failures. The fault detection relies on models built from historical data. As new sensor observations are collected the *model* is updated and compared to that previously estimated, where a difference is indicative of a failure.

1. Introduction

This paper describes a data-driven approach to online data reconciliation and validation for a small network of sensors. The sensors are embedded in an environment such as office facilities to monitor activity and detect unusual activity or behaviors. Using the language of automated diagnosis, the activity or behavior under observation corresponds to the process under observation. Statistical approaches are employed to analyze the data and infer the type of activity taking place. As sensors reading are collected they are checked for validity in real time. These sensors are susceptible to interference or could fail, undermining the performance of the system. This paper presents a method to detect and mask the readings deemed to be in error. Masking can be achieved in a number of ways; in this work statistical models are employed. A difficulty in this application results from the fact that the process (activity) under observation can cause deviations in the sensor outputs that are indistinguishable from noise and/or sensor faults.

The approach adopted is to model the relationships between sensors rather than individual sensors and using these relationships to cross-validate and correct incoming readings. Thus the first step is to determine the relationships between sensor outputs and construct a statistical model. The main contribution of this paper is a data-driven method for detecting permanent and transient faults on a small network of sensors; it is exploits the sensor–sensor relationships to deal with uncertainty in sensor observations caused by noise or sensor failure. Sensor–sensor relationships are discovered using historical data. The search for similarity extends beyond sensors observing a similar process or proximity of sensors. It attempts to characterize any relationship over time. The methodology was tested offline on data from a smart home.

Section 2 introduces related work, while Section 3 describes the methodology. Results are presented in Section 4 and discussed in Section 5. Finally, Section 6 concludes the paper.

2. Related work

The objective is to detect and correct measurement errors. The errors encountered are either systematic (gross errors) or random fluctuations. This work deals with random errors caused by electrical and other types of interference and gross errors caused by sensor faults. In a first step, a sensor reading or readings are identified as anomalous. A second step reconciles all erroneous readings.

The problem is essentially one of diagnosing faults, although, unlike most diagnosis, characterizing the nature of the fault is not necessary. Diagnosis of system faults is a field that has been widely studied in applications such as aerospace (Dearden et al., 2004), process control (Juricek, Seborg, & Larimore, 2004a), and electricity networks (Bauer, Botea, Grastien, Haslum, & Rintanen, 2011). With the proliferation of distributed communication networks, and in particular sensor networks, the diagnosis of such systems has also received much attention (Chen, Kher, & Somani, 2006; Gao, Xu, & Li, 2007; Hai Li, Price, Stott, & Marshall, 2007; Jiang, 2009; Krishnamachari & Iyengar, 2004; Lee & Choi, 2008). Most networks are large networks; however similar techniques have been applied to smaller networks. Kim and Prabhakaran (2011) describe fault diagnosis for a very small network, a body sensor network (BSN). The characteristics of BSN are similar to the sensor network presented here, albeit much smaller in size

* Corresponding author. Tel.: +44 117 3287126.
E-mail addresses: dorothy.monekosso@uwe.ac.uk, d.monekosso@kingston.ac.uk (D.N. Monekosso).

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2012.12.037

than the environment network. There is a relatively small number of sensor nodes; each node comprising a wireless communication unit and a sensing element.

The papers mentioned above are broadly classified as model-based or data driven, although there is some overlap incorporating both model and data-driven. A model may be created from historical data in data-driven methods. The model-based approach may employ physical and other types of models. Irrespective of the type of model, it is used to reason about the behavior of the system, comparing the observed behavior with expected (predicted) behaviors. Approaches to model-based diagnosis differ principally by the type of model employed. These include Bayesian (Abreu, Zoeteweij, & van Gemund, 2009; Krishnamachari & Iyengar, 2004), and hidden Markov models (HMM) (Srivastava, 2005; Ying, Kirubarajan, Pattipati, & Patterson-Hine, 2000). Prediction filters have also been investigated in the form of Kalman filters (Kim, Suk, & Kyung, 2010) and particle filters (Zhou & Liu, 2010). The diagnosis problem may be cast as a feature classification problem and standard classification methods used such neural networks (Maidon, Jervis, Dutton, & Lesage, 1997; Venkatasubramanian, Vaidyanathan, & Yamamoto, 1990), fuzzy classifier (Lo, Fung, & Wong, 2009), and clustering (Iverson, 2004). The work presented a statistical model built from historical data. Statistical multivariate techniques (Ma, Wong, Jang, & Tseng, 2010) have been employed in data-driven diagnosis. A popular statistical approach is principle component analysis (PCA) (Wise, Gallagher, Butler, White, & Barna, 1999; Yue, Qin, Markle, Nauert, & Gatto, 2000; Zhang & Wang, 2004; Zhou, Zhang, & Wang, 2004). More recent is the canonical correlation analysis (CCA) and variants applied to fault detection as proposed in (Chen, Jiang, & Yoshihira, 2006; Juricek, Seborg, & Larimore, 2004b; Kang, Chen, & Jiang, 2010).

The model-based and data-driven methods have been adapted to wireless sensor networks (WSN). Fault detection in WSN exploits the correlation in a network of sensor nodes. For example, two temperature sensors in a network in close proximity are likely to observe a similar temperature. Thresholds that must not be exceeded are defined for the difference between two sensor observations and for the increment in a unit time step (Lee & Choi, 2008). Jiang (2009) improves on this with a weighted average scheme. The author in Krishnamachari and Iyengar (2004) exploits correlation using a Bayesian algorithm. The proposed method exploits correlation; however it does not rely purely on proximity and/or sensor similarity.

In many safety critical systems, a redundant unit replaces a faulty unit. In a static redundancy scheme, a voting system selects one from at least 3 units without performing fault detection. In a dynamic scheme, the faulty unit is replaced by a redundant unit after fault detection has identified the faulty unit. Fault hiding can be used to recover seamlessly from a fault (Guenab, Weber, Theilliol, & Zhang, 2011; Steffen, 2006; Richter, 2011). Once a sensor fault is isolated, the expected sensor output (from the model) is used to validate and if necessary mask the real faulty value in subsequent analyses.

3. Methodology

The objective when analyzing smart environment data is to detect anomalous events or activities. The standard method is to build models of normal events and any new activity detected is compared. If an activity has not been previously detected, it is deemed abnormal. This assumes correct sensor measurements. Any significant deviation of raw measurements from the norm not resulting from the process under observation will undermine the analyses. A method is proposed to disambiguate process deviation from sensor deviation; combining standard data reconciliation techniques with failure detection techniques. The proposed method comprises two concurrent threads. One thread deals with random measurement fluctuations using a standard data reconciliation technique while the second thread detects and locates the source of systematic deviations in sensor readings (gross error) which may be caused either by a faulty sensor or anomalous process. If a failed sensor cannot be found, it is assumed that the process under observation is the source of the systematic deviation. Models are constructed from historical data and refined with incoming sensor readings. Using these models a sensor fault can be detected and the faulty sensor located. There is no need to identify the fault mechanism. Thus expected-behavior is modeled rather than fault-behavior as is normally the case in fault diagnosis. The failed sensor is located by modeling the relationships between the sensors in normal operation. The system dynamically searches for relationships between sensors and models the found relationships. Each new sensor reading is tested against the known relation, to establish its validity. The challenge is to find, for each sensor, the set of sensors that correlate and their relationship which may be dynamic. An intelligent environment contains a heterogeneous array of sensors. Consequently, the sensor readings differ in data type, amplitude, and frequency and may comprise continuous (e.g. temperature) and discrete (e.g. switch state), compound (e.g. an image). A representation that caters for all these types is required. A suitable representation uses a probability distribution allowing all sensors to be treated in the same manner and information to be combined.

3.1. Data reconciliation

The classic data reconciliation technique (Eq. (1)) is not suitable on its own, if reconciling a datum means that an underlying deviation in the process will also be masked. Eq. (1) assumes no systematic errors are present in the measurement and that the measurement noise is random. This is the case for interference seen within the sensor network, which comprises sensors and all measurement related equipment. For an ensemble of n measurements y_i (sensor readings)

$$y_i = x_i + e,$$

where y_i the ith sensor measurement, and x_i is the ith true (unmeasured) value. The value e *is the* measurement error, consisting of random fluctuations only and has a Gaussian distribution. The objective is to minimize the least square correction error

$$\min_{x,y^*} \sum_{i=1}^{n} \left(\frac{y_i^* - y_i}{\sigma_i} \right)^2 \tag{1}$$

Subject to the activities maximum project duration of a day and bounding x and y, the sensor minimum and maximum values:

$$y_{min} \leqslant y^* \leqslant y_{max}$$
$$x_{min} \leqslant x^* \leqslant x_{max}$$

where y_i^* is the corrected value of the ith sensor measurement, and σ_i is the standard deviation, y_i is the sensor measurement that maximizes the correlation with neighboring sensors. Neighborhood does not simply imply physical proximity of two or more sensors but it also implies sensors that exhibit some relationship (i.e. correlate). Canonical Correlation Analysis (CCA) is used to find correlating sensors.

3.2. Systematic error detection

Sensor signals are usually multi-dimensional and might span a large spectrum/space. Signatures of an unfolding scene tend to exist in a subspace. Methods have been proposed to discover the