Information Fusion 33 (2017) 41–56

Contents lists available at ScienceDirect

Information Fusion

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

Spatial anomaly detection in sensor networks using neighborhood information

Hedde HWJ Bosman[a,b,*], Giovanni Iacca[a], Arturo Tejada[c], Heinrich J. Wörtche[a], Antonio Liotta[b]

[a] INCAS[3], Dr. Nassaulaan 9, 9401HJ, Assen, The Netherlands
[b] Department of Electrical Engineering, Eindhoven University of Technology, P.O. Box 513, 5600MB, Eindhoven, The Netherlands
[c] TNO, Integrated Vehicle Safety Department, 5700 AT Helmond, The Netherlands

ARTICLE INFO

Article history:
Received 29 February 2016
Revised 24 April 2016
Accepted 25 April 2016
Available online 26 April 2016

Keywords:
Anomaly detection
Sensor fusion
Sensor networks
Collaborative WSN

ABSTRACT

The field of wireless sensor networks (WSNs), embedded systems with sensing and networking capability, has now matured after a decade-long research effort and technological advances in electronics and networked systems. An important remaining challenge now is to extract meaningful information from the ever-increasing amount of sensor data collected by WSNs. In particular, there is strong interest in algorithms capable of automatic detection of patterns, events or other out-of-the order, anomalous system behavior. Data anomalies may indicate states of the system that require further analysis or prompt actions. Traditionally, anomaly detection techniques are executed in a central processing facility, which requires the collection of all measurement data at a central location, an obvious limitation for WSNs due to the high data communication costs involved. In this paper we explore the extent by which one may depart from this classical centralized paradigm, looking at decentralized anomaly detection based on unsupervised machine learning. Our aim is to detect anomalies at the sensor nodes, as opposed to centrally, to reduce energy and spectrum consumption. We study the information gain coming from aggregate neighborhood data, in comparison to performing simple, in-node anomaly detection. We evaluate the effects of neighborhood size and spatio-temporal correlation on the performance of our new neighborhood-based approach using a range of real-world network deployments and datasets. We find the conditions that make neighborhood data fusion advantageous, identifying also the cases in which this approach does not lead to detectable improvements. Improvements are linked to the diffusive properties of data (spatio-temporal correlations) but also to the type of sensors, anomalies and network topological features. Overall, when a dataset stems from a similar mixture of diffusive processes precision tends to benefit, particularly in terms of recall. Our work paves the way towards understanding how distributed data fusion methods may help managing the complexity of wireless sensor networks, for instance in massive Internet of Things scenarios.

1. Introduction

In the last decade, the vision of an internet of things (IoT) has rapidly become reality. Recent advances in technology, together with ever-decaying prices of electronic components, have made networked embedded systems ubiquitous in our life. These devices are in most cases endowed with sensing, actuating and networking capabilities and are often connected to the Internet. Noteworthy applications of these systems can be found, for instance, in home automation, automated transportation, or large scale environmental data collection [1].

While at present white goods, smart cities and buildings are being equipped with IoT technology [2], one of the earliest IoT related systems were (and are) wireless sensor networks (WSNs), with typical applications in environmental monitoring [3] and tracking of mobile agents [4]. Such applications usually require numerous sensor nodes to be deployed in remote locations. To make such systems affordable, costs are saved by reducing the quality of the sensors and the hardware resources available on each node (such

* Corresponding author.
E-mail address: sgorpi@gmail.com (H.H. Bosman).

http://dx.doi.org/10.1016/j.inffus.2016.04.007

as battery and computing elements), while the overall measurement quality of the networked system is often ensured by a high level of redundancy in measurements. For this reason, the past decade of WSN research focused mostly on optimizing resource usage [5–7].

With this body of research maturing, and the sensor technology advancing, the attention of the field is now shifting towards applications [8–11]. However, these harbor some hard theoretical problems related to the envisioned scale of the network deployments, such as the analysis of large amounts of data, stemming from, e.g., sensor networks deployed in large outdoor areas or from the many networked appliances in a smart home. The collected data is often analyzed in order to find specific information at a given point in time that is meaningful for the application to act upon. For example, seismic data could be analyzed for patterns that denote seismic activity [12], body sensor data can be analyzed to provide early health warnings [13], or vibration data could be mined for events that potentially point to a failing machine [14]. Often, such patterns or events are out of the ordinary or anomalous.

Anomaly detection can be defined as the detection of events, behaviors or patterns that are unexpected relative to a concept of what is normal [15]. A typical example is the detection of fraud in, e.g., credit card transactions or the detection of identity falsification [16]. One can also think of climate events, such as heat waves and droughts. What defines climate events as anomalous depends on multiple variables, such as location, and the proper context (drought in the Sahara desert, for instance, is not anomalous) [17]. Anomaly detection approaches are also used to detect intrusions in information systems, ever more relevant in present-day cloud computing [18].

Anomaly detection approaches is popular in applications with large central storage and processing facilities, such as those employed to process big data [19]. However, their application to lightweight systems, such as WSNs, is still limited due to the severe resource limitations posed by these systems. Limited memory and the high communication costs, for example, preclude the scenario where all WSN nodes send all information to a central facility for storage and processing [20]. To address these problems, one must either adapt to the aforementioned limitations the approaches available in the literature (which however are devised, in general, for general-purpose computers), or develop new solutions. Moreover, due to the lack of contextual information that is often not present at design time, such methods need self-adaptive mechanisms or dynamic model fitting approaches, such as machine learning techniques, to allow them to operate on data of different, unpredictable environmental conditions. Such learned models can be bootstrapped with the little information available during design time, or be learned completely unsupervised during deployment.

The decentralized nature of WSN results in measurements taken in different points in space, over time. Due to the decreasing cost of the hardware, more nodes can be deployed which results in higher quality data through redundancy. However, the measurements can contain anomalies that occur with respect to local sensors, to neighborhood information or to global information. Using anomaly detection techniques a node can, for instance, generate an initial estimate of the reliability of measurements through aggregation of local spatial neighborhood information, thus reducing the amount of data sent to a central processing facility and allowing the generation of a local and timely response to anomalies. The central processing facility could then use all the aggregated data to provide a second detection or estimation stage to improve anomaly detection accuracy, using its abundant storage and computing power resources.

In this paper, we address the following question: Can the local detection of anomalies be improved (in terms of precision or recall) by combining data from groups of spatially co-located sensor nodes? To answer this question, we devise a novel anomaly detection system based on a decentralized unsupervised online learning scheme, which incorporates local neighborhood information. We extensively evaluate this approach over a broad range of real-world network deployments and datasets from different domains. Then, in order to show the effect of the neighborhood information on the anomaly detection, we compare the performance of the framework with and without the use of neighborhood information.

The remainder of this paper is structured as follows: The next section provides a short summary of the literature related to our work. Section 3 presents our new anomaly detection approach and describes our experimental setup, while Section 4 shows and discusses our experimental results. Finally, Section 5 provides our conclusions.

2. Related work

Anomaly detection is often used in applications such as fraud detection [16], network intrusion detection [21], data centers [22], or airline safety [23]. Historical (or, *a priori*) data is used to construct a model of the normal behavior of the process (or system) under consideration, and newly arriving data is tested for fitting with the model. Patterns or behaviors that do not fit are then classified as anomalous, as fraudulent, as faulty, or simply as events that require further human analysis.

Within the research related to networked embedded devices (such as WSNs), one can often see a similar approach: Data is collected at a central point, where it is analyzed to find the anomalies. This allows, for instance, the use of multiple classifiers in an ensemble, each of which can excel in different aspects of the complex dynamics of the system under monitoring [24]. Furthermore, it allows complex transforms of multivariate time-series [25] or human reinforcement as additional detection method in, e.g., a large oceanic dataset [26].

However, central techniques have several drawbacks. The notable ones in the context of WSN systems have mainly to do with their resource usage. The wireless communication scheme also has inherent drawbacks, such as packet loss, while many detection techniques often assume reliable periodic data and, thus, have to deal with delayed packets due to retransmissions [27]. Furthermore, models learned from previously acquired data may not be suitable at any given time, and thus may require frequent model updates. Depending on the detection method used, these updates may be intrinsic and lightweight, or may require the reprocessing of all the acquired data [28].

To overcome some of these drawbacks, hybrid approaches create and update models offline that are suitable for online use in limited-resource environments. Such approaches offload the learning to a more powerful node and, thus, allow more complicated models to be learned. For example, time series are often modeled using an autoregressive moving average (ARMA) model [29]. Although, the model parameters could be estimated online, offline parameter estimation ensures that the model represents normal data, and leaves valuable computing cycles to run additional detection and classification techniques on the nodes. More complex models can only be trained offline due to resource limitations. For instance, echo state networks, a form of recurrent neural networks, can model complex time series with historical data offline. The resulting neural network can be used in WSN nodes to classify anomalies [30]. One can also think of another type of hybrid approach, where resource-limited nodes only provide basic anomaly detection methods to provide early warnings, while more complex detection methods are executed at a base station. This approach is applied, for example, in electronic health care, where WSN nodes provide early warnings based on sliding window features (such as thresholds of the mean), while a base station performs complex