Pattern Recognition 47 (2014) 2867–2879

Contents lists available at ScienceDirect

Pattern Recognition

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

Ellipsoidal neighbourhood outlier factor for distributed anomaly detection in resource constrained networks

Sutharshan Rajasegarar [a,*], Alexander Gluhak [c], Muhammad Ali Imran [c], Michele Nati [c], Masud Moshtaghi [b], Christopher Leckie [b], Marimuthu Palaniswami [a]

[a] Department of Electrical and Electronic Engineering, The University of Melbourne, Australia
[b] Department of Computing and Information Systems, The University of Melbourne, Australia
[c] Centre for Communication System Research, University of Surrey, UK

ARTICLE INFO

Article history:
Received 24 July 2013
Received in revised form
6 March 2014
Accepted 1 April 2014
Available online 12 April 2014

Keywords:
Anomaly detection
Outlier factor
Hyperellipsoidal model
Distributed detection
Sensor networks

ABSTRACT

Anomaly detection in resource constrained wireless networks is an important challenge for tasks such as intrusion detection, quality assurance and event monitoring applications. The challenge is to detect these interesting events or anomalies in a timely manner, while minimising energy consumption in the network. We propose a distributed anomaly detection architecture, which uses multiple hyperellipsoidal clusters to model the data at each sensor node, and identify global and local anomalies in the network. In particular, a novel anomaly scoring method is proposed to provide a score for each hyperellipsoidal model, based on how remote the ellipsoid is relative to their neighbours. We demonstrate using several synthetic and real datasets that our proposed scheme achieves a higher detection performance with a significant reduction in communication overhead in the network compared to centralised and existing schemes.

1. Introduction

Resource constrained networks, such as wireless sensor networks (WSNs), composed of compact, cheap, intelligent sensor nodes, provide the ability to collect measurements from sensors that are distributed across a large physical area with a lower cost of building and operating the network compared to a traditional wired network. These benefits come with a trade-off in terms of limited resources, specifically, the battery life of sensor nodes. By detecting interesting or unusual events in sensor networks, we can avoid transmitting uninformative or erroneous measurements. This can reduce the energy consumed in the network, as well as improving the reliability of the data collected. Detecting interesting or unusual events in the network is known as the anomaly detection problem.

An anomaly or outlier in a data set is defined as "an observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data" [1]. In order to identify anomalies in a data set, we need to find a model for the normal data and then the anomalies can be identified as those data vectors that deviate significantly from the normal model. Anomalies in sensor networks can occur due to faulty sensors, gradual drift in the sensor elements, loss of calibration, noisy sensor transducers, movement of sensors, changes in observed phenomena or malicious attacks such as false data attacks. A key challenge in WSNs is to identify any misbehavior or interesting events (anomalies) with minimal communication overhead within the network while achieving high detection accuracy.

In this paper, we propose a novel anomaly scoring mechanism and a distributed anomaly detection architecture for resource constrained networks. Our main contributions in this paper are as follows.

1. We propose a novel anomaly scoring scheme called the hyperEllipsoidal Neighborhood Outlier Factor (ENOF). This assigns an outlier factor (score) to each hyperellipsoidal model of the sensor data at each sensor node, and identifies the globally anomalous set of hyperellipsoids.
2. Using the ENOF, we propose a distributed anomaly detection architecture that enables identification of any anomalies in the network, which has the following characteristics:
   (a) It models the sensor data collected at each node level using *multiple* hyperellipsoids, which enables the modelling of multi-modal distributions in the sensor data. We utilise our

* Corresponding author. Tel.: +61 03 83440112; fax: +61 03 8344 6678.
E-mail addresses: sraja@unimelb.edu.au (S. Rajasegarar), a.gluhak@surrey.ac.uk (A. Gluhak), m.imran@surrey.ac.uk (M. Ali Imran), m.nati@surrey.ac.uk (M. Nati), mosm@unimelb.edu.au (M. Moshtaghi), caleckie@unimelb.edu.au (C. Leckie), palani@unimelb.edu.au (M. Palaniswami).

http://dx.doi.org/10.1016/j.patcog.2014.04.006
0031-3203/

previously proposed HyCARCE algorithm [2] (see Section 4.1) and our ellipsoidal similarity measure [3] (see Section 4.2.1) to automatically infer the set of hyperellipsoidal clusters at each node level.

(b) Only summary information about the hyperellipsoids needs to be communicated among the nodes (along the sensor network routing hierarchy) to infer a set of global hyperellipsoids that correspond to a global model of normal behaviour. In comparison to communicating all the raw measurements to a central gateway for analysis, our approach vastly reduces the communication complexity in the network and helps prolong the lifetime of the network.

(c) It identifies globally anomalous data vectors *at their node level* based on the ENOF score, in addition to the locally anomalous data vectors, using the globally anomalous hyperellipsoids that are communicated to the nodes by their parent nodes. The proposed architecture is complete in this sense of being able to detect all local and global anomalies down to the level of individual nodes.

3. We evaluate our scheme on several real and synthetic data sets and demonstrate that our scheme achieves comparable detection accuracy to centralised and other existing schemes while consuming significantly less communication overhead in the network.

Our proposed outlier scoring methodology and the anomaly detection architecture have many potential applications. These include monitoring the health of coral reefs such as the Great Barrier Reef, Australia [4], monitoring and identifying inefficient energy consumption behaviour in an office or residential (indoor) environment [5], and Internet of Things (IoT) [6] applications such as Smart City monitoring for noise, pollution and environmental parameters [7–11].

The rest of the paper is organised as follows. Section 2 describes different kinds of anomalies in WSNs, followed by a review of existing anomaly detection methods in Section 3. Our proposed architecture and the ENOF are presented in Section 4. Complexity analysis is given in Section 5. Evaluation and comparison of the algorithms using several real and synthetic datasets is presented in Section 6, followed by the conclusion and future work in Section 7.

2. Network topology and local and global anomalies

The network topology of WSNs and the composition of sensor node types used in the network are heavily dependent on their application. One flexible arrangement of sensor nodes is a clustered or multi-tiered hierarchical topology in which a parent–child relationship exists, such as the one shown in Fig. 2(a). We use this hierarchical topology to describe the different kinds of anomalies found in a sensor network and to propose a distributed detection approach. However, our scheme is applicable to any network topology where a set of sensors have communication capabilities with their neighbours in the network.

In WSNs, anomalies can be at the level of individual measurements with respect to the other measurements at the same sensor node, or at the level of the measurements of one node with respect to other sensor nodes in the network. We have identified three categories of anomalies in WSNs [12]. First, anomalies can occur in individual measurements or network traffic attributes at a sensor node (refer to Fig. 1(a)), i.e., some observations at a node are anomalous with respect to the rest of the data. Second, all of the data at a sensor node may be anomalous with respect to neighbouring nodes. In this case, such a sensor node will be identified as an anomalous node as shown in Fig. 1(b). Third, a set of sensor nodes in the network can be anomalous as shown in Fig. 1(c). We refer to these three types of anomalies as *first*, *second* and *third order anomalies*.

The anomalies found in a WSN can be categorised into *local anomalies* and *global anomalies*. Local anomalies are anomalous measurements in a sensor nodes' own (local) data measurements, where only the measurements at the same node are used as a basis for comparison. However, we are also interested in cases where the majority of measurements at a sensor node are anomalous in comparison to other nodes in the network. These global anomalies are anomalous measurements in the union of the measurements collected from multiple sensor nodes in the network.

In order to detect these different kinds of anomalies, a model is required that represents the structure of normal and anomalous data in the network. We use hyperellipsoidal models for this representation, as they can accommodate many different data distributions ranging from hyperspherical to linear. Below we provide a review of the state-of-the-art on the different models used for anomaly detection, and then describe our proposed distributed anomaly detection architecture and the scoring mechanism using the hyperellipsoidal data model in detail.

3. Related work

The task of detecting interesting or unusual events in a general manner is an open problem in the data mining community, and is often referred to as the anomaly detection problem. Several alternative definitions of anomalies have been proposed in the literature, as well as a variety of detection algorithms [13–16]. Anomaly or outlier detection techniques that do not assume any prior knowledge about the distribution of the data are called non-parametric anomaly detection techniques. These techniques are suitable for resource constrained sensor networks where the data distribution may change frequently. Below we look in more detail at some of the recent non-parametric approaches that have been proposed for such networks, namely, nearest neighbour based approaches and clustering approaches.

Nearest neighbour based approaches use distances among data vectors as the similarity metric. Examples of such metrics include the distance to a nearest neighbour data vector (NN), the distance to the *k*th nearest neighbour data vector (kNN), and the distance to the average of the *k* nearest data vectors (Average kNN). Note that *k* is a user defined parameter. These similarity measures are used

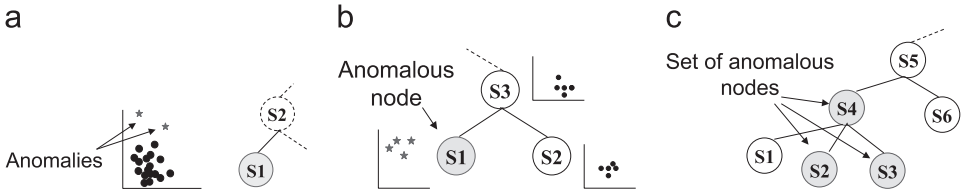

Fig. 1. Types of anomalies in wireless sensor networks. (a) First order anomalies, (b) second order anomalies and (c) third order anomalies.