Knowledge-Based Systems 60 (2014) 44–57

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Adaptive and online data anomaly detection for wireless sensor systems

Murad A. Rassam [a,b,*], Mohd Aizaini Maarof [a], Anazida Zainal [a]

[a] Information Assurance and Security Research Group (IASRG), Faculty of Computing, Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia
[b] Faculty of Engineering and Information Technology, Taiz University, 6803 Taiz, Yemen

ARTICLE INFO

Article history:
Received 3 June 2013
Received in revised form 18 November 2013
Accepted 3 January 2014
Available online 18 January 2014

Keywords:
Wireless sensor networks
One-Class Principal Component Classifier
Data anomaly detection
Adaptive anomaly detection
Sensor data analysis
Sensor data quality

ABSTRACT

Wireless sensor networks (WSNs) are increasingly used as platforms for collecting data from unattended environments and monitoring important events in phenomena. However, sensor data is affected by anomalies that occur due to various reasons, such as, node software or hardware failures, reading errors, unusual events, and malicious attacks. Therefore, effective, efficient, and real time detection of anomalous measurement is required to guarantee the quality of data collected by these networks. In this paper, two efficient and effective anomaly detection models PCCAD and APCCAD are proposed for static and dynamic environments, respectively. Both models utilize the One-Class Principal Component Classifier (OCPCC) to measure the dissimilarity between sensor measurements in the feature space. The proposed APCCAD model incorporates an incremental learning method that is able to track the dynamic normal changes of data streams in the monitored environment. The efficiency and effectiveness of the proposed models are demonstrated using real life datasets collected by real sensor network projects. Experimental results show that the proposed models have advantages over existing models in terms of efficient utilization of sensor limited resources. The results further reveal that the proposed models achieve better detection effectiveness in terms of high detection accuracy with low false alarms especially for dynamic environmental data streams compared to some existing models.

1. Introduction

Wireless sensor networks (WSNs) are formed by a number of small, cheap, battery-powered, and multi-functional devices called sensors which are densely or sparsely deployed to collect information from environments or to monitor phenomena [1]. However, the constrained sensor resources in terms of storage, processing, bandwidth, and energy make WSNs more vulnerable to different types of misbehaviors or anomalies. Anomaly is defined in [2] as, “an observation that appears to be inconsistent with the reminder of a dataset”. These anomalies always correspond to sensor software or hardware faults, reading errors, and malicious attacks. They may also correspond to some events of interest such as sudden changes in the monitored parameters that may indicate unusual phenomenon. Therefore, it is desirable to efficiently and accurately identifying anomalies in the sensor measurements to ensure the quality of these measurements for making right decisions.

The default approach to detect anomalies is to separate them from normal data using various types of classifiers. However, the absence of ground truth labeled data hinders the direct classification of sensor measurements using traditional classifiers. Instead, the available normal data are modeled using one class classifiers as in [9–11] and then identify any deviation from this model as anomalies.

A variety of anomaly detection models for WSNs have been proposed in the literature and they can be classified based on the place in which the detection is performed into *centralized* and *de-centralized*. In the *centralized* models, it is assumed that the data is available in a central location (such as base station or cluster head) for further analysis [3]. Therefore, the normal model is built using the whole data sent by all nodes in a specific time period and used to detect any significant deviations. However, these models are not suitable for energy constrained WSNs because they assumed the availability of whole data at central location for further analysis and therefore cause prohibitive communication overhead. On the other hand, these models may be useful as baseline models for comparison with different detection algorithms [4].

Meanwhile, in the *de-centralized* models, each node uses its own data to build local normal model. The *de-centralized* models can be further classified based on the decision making mechanism and they can be called *node-level* and *network-level* models. In the *network-level* models [5–9], some parts of processing are locally

[*] Corresponding author at: Information Assurance and Security Research Group (IASRG), Faculty of Computing, Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia. Tel.: +60 173681303.

http://dx.doi.org/10.1016/j.knosys.2014.01.003
0950-7051/© 2014 Elsevier B.V. All rights reserved.

performed in each node and a cooperation mechanism between a group of nodes (i.e. in cluster) is conducted to decide about the potential anomalies. The distributed models aimed to solve the problem of high communication overhead of the centralized models by sending a summary that represents the normal reference model at each node to the central location. In the central location, the global normal reference model is computed by merging the received local models and sent back to each node for local anomaly detection. The problem of some models of this category is the data heterogeneity which makes the global model unsuitable representative of the network normal behavior. Moreover, the size of the normal model that is exchanged between nodes and the central location may increase the communication overhead which is the main factor of quick sensor energy consumption. In the *node-level* models such as [4], the processing, analysis, and decision about anomalies are totally performed locally at each node. The decision is then sent to the central location for further remedial action. Besides, most of the existing distributed detection models such as [9–12] are not suitable for online detection as their detection methods incur high computational complexity that quickly consume the limited sensor energy.

The effectiveness of anomaly detection models is affected by dynamic changes of deployed environments. Therefore, adaptive detection of anomalies in such environments is an important challenge for assuring the quality of sensor measurements. These changes increase the false alarm rates and therefore affect effectiveness of detection models. Few works have tackled the adaptive detection issues such as [13,14]. However, these models have drawbacks such as the additional communication overhead incurred in distributed models and the high computational cost produced by the incremental learning procedures.

The contribution of this paper is three folds:

(i) We show how the Unsupervised Principal Component Classifier which was originally proposed in [29] can be adopted as a One-Class Principal Component Classifier suitable for anomaly detection in sensor measurements in the absence of ground truth labeled data.
(ii) A new efficient and online anomaly detection model for WSN, namely Principal Component Classifier-based Anomaly Detection (PCCAD) model is proposed. The new model is totally local and does not incur any additional communication overhead.
(iii) An adaptive APCCAD model that incorporates an incremental learning method in the design of the PCCAD model to track the data changes in dynamic environments. As a result, the misclassification error caused by this dynamic change is reduced.

The efficiency and effectiveness of the proposed models are validated and compared with some existing detection models using real life datasets collected from real sensor network deployments.

The remainder of this paper is structured as follows: related anomaly detection models in WSNs are presented in Section 2. The proposed models are described in Section 3. The experimental results, analysis, and evaluation of the proposed models are reported in Section 4. Section 6 concludes the paper and suggests some directions for future research.

2. Related works

Anomaly detection in WSNs has many applications including intrusion detection, event detection, fault detection, and outlier detection [3]. The term fault refers to a deviation from an expected value regardless of the cause of that fault. Hence, data faults can be considered as anomalies in the absence of ground truth values [15]. Three types of faults which are *short*, *noise*, and *constant reading* have been studied in [15]. The authors evaluated the performance of three methods to data fault detection in WSN that fall into three approaches: rule-based, estimation-based, and learning-based. It is reported that the methods worked well with high and medium intensity of *short* injected faults and high intensity *noise* injected faults. However, in most cases, these methods failed to detect long or *constant* injected faults and low intensity *short* and *noise* injected faults. For the real world datasets, it is reported that these methods performed generally well as these datasets have experienced high intensity faults.

A combination of discrete wavelet transform (DWT) and self organizing map neural networks (SOM) to detect data anomalies in WSNs was developed in [16]. In this study, data faults were considered as anomalies. The data measurements were first encoded at each node using DWT and then sent to the base station where SOM was applied on a batch of wavelet coefficients. A similar combination of DWT and one-class support vector machines (OCSVM) was proposed in [17] to develop an anomaly detection model for WSNs. In this model, DWT was used for encoding the data measurements at each node like in [16]. The encoded measurements were then examined for anomalies by the OCSVM at the base station. In both models [16,17], the anomaly detection process was taken place in the base station on batches of encoded measurements. These models can be considered as distributed model as they apply DWT on data of each sensor before sending the encoded coefficients to the base station. They are also considered as centralized models as they perform the anomaly detection on batches of coefficients at the base station. The high computational complexity of OCSVM and SOM methods make these models unsuitable for online detection in sensor nodes.

A segmented sequence analysis (SSA) algorithm was used to design an online anomaly detection model in [18]. Data anomalies were detected by comparing the constructed piecewise linear scheme of data collected in a fixed time period with a reference model using similarity metrics and flagged anomalies when there is a significant difference. A data from real world sensor network deployments was used to evaluate and demonstrate the effectiveness and efficiency of the proposed model. It is claimed that this model is efficient and effective in detecting data anomalies compared to some existing models [15]. This model can be considered as distributed model that has two layers of detection. The first layer is locally in the scope of each sensor node and the second layer is in the cluster head. However, there was no feedback from the cluster head to the local nodes and this issue was left as future work.

In [5,6], two distributed anomaly detection models were proposed based on clustering ellipsoids of sensor measurements. The normal model was calculated locally in sensors and a summary of the model is sent to the cluster head to calculate the global normal model which is sent back to nodes. The proposed models in these works were aimed to detect anomalies in heterogeneous sensor networks where the distribution of data is evolving. A distance based anomaly detection model for WSNs was proposed in [19]. In this model, PCA was first applied to reduce the dimension of data and then a distance-based method was applied to detect anomalies. However, this model is static since it detects anomalies in batches of sensor measurements in a fixed period of time.

Recently, the authors of [20] proposed an online histogram-based anomaly detection model to detect anomalies in a distributed manner. It is claimed that the proposed model overcome the existing histogram based models in such that the verification procedure was exempted by the use of probability estimation. Two important issues were not properly addressed in the works [19,20] which are: *first*, they assume a high correlation between