Available online at www.sciencedirect.com

ScienceDirect

AASRI Procedia 5 (2013) 9 – 14

AASRI Procedia

www.elsevier.com/locate/procedia

2013 AASRI Conference on Parallel and Distributed Computing and Systems

An Efficient Distributed Anomaly Detection Model for Wireless Sensor Networks

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

[a]Faculty of Computing, Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia
[b]Faculty of Engineering and Information Technology, Taiz University, Taiz, Yemen

Abstract

Wireless sensor networks (WSNs) are important platforms for collecting environmental data and monitoring phenomena. Anomalies caused by hardware and software errors, unusual events, and malicious attacks affect the integrity of data gathered by such networks. Therefore, anomaly detection process is a necessary step in building sensor network systems to assure data quality for right decision making. This paper presents a new distributed online anomaly detection model that measures the dissimilarity of sensor observations in principal component space. The new model distributes the detection process over the network to minimize energy consumption while ensuring high detection effectiveness. The detection effectiveness and efficiency of the new model are proved through experiments on real world dataset from Sensorscope system project. Experimental results reveal that the model achieved high detection rate with relatively few false positive rates compared to an existing model.

*Keywords:*Wireless Sensor Networks; Anomaly Detection; Distributed Processing; Principal Component Analysis; Principal Component Classifier.

* Corresponding author. Tel.: +60-17-3681303; fax: +607 55 38829.
E-mail address: murad.utm@gmail.com.

doi:10.1016/j.aasri.2013.10.052

1. Introduction

Wireless sensor networks (WSNs) are composed of small-sized, cheap, energy-limited, and multifunctional devices called sensors that are deployed to collect data from an environment or monitor a phenomenon[1]. The limited resources such as processing capabilities and energy increase the possibility of these networks to be susceptible to variety of misbehaviors or anomalies. Data anomaly is defined in [2] as, "an observation that appears to be inconsistent with the reminder of a dataset". Hardware and software faults, observations errors, unusual events, or intrusions are possible causes for sensor data anomalies. Therefore, to assure the integrity of sensor observations and identify important interested events in the monitored phenomena, detection of anomalies should be considered for the design of any sensor network system.

Distribution of anomaly detection process over nodes in WSN has been proposed as a solution to reduce the computational load of detection methods in nodes. However, another factor should be considered when designing any distributed detection model which is the communication overhead caused by transmitting large amounts of data between nodes. Some distributed anomaly detection models have been designed to consider the communication overhead such as [3-6]. The concept of distribution in these works relies on that each node implements the anomaly detection method by its own resources and then transmits only a summary of local detection model to the central location such as cluster head or base station in order to construct a global detection model. The global model is then returned to the nodes for using it in anomaly detection of subsequent observations.

A distributed deviation detection model in WSN was described in [4] to avoid the unnecessary communication and computational effort. This model was designed based on a non-parametric statistical technique called kernel density estimator. The hierarchical structure which is always used to implement any distributed solution was not adopted for this model. Instead, the hierarchical structure was emulated by assigning each group of low capacity sensors to one of some limited more powerful sensors based on spatial proximity. An In-network outlier detection algorithm was proposed in [5] and aimed at reducing the communication overhead results from massive data transmission in WSN. This algorithm was based on distance similarity to find the global anomalies in WSN. In this model, each node applies the distance similarity measure to find the anomalies and broadcast its result to its neighbours. Other nodes perform the same task until all nodes agree on a common decision on anomalies. Broadcast communication lead to high energy consumption. The problem is worsen when deal with large scale WSN.

A Quarter Sphere Support Vector Machine (QSSVM) based distributed anomaly detection model was proposed in [6] to identify the anomalous observations in sensors. The procedure of applying this model was as the following:

- Every sensor runs the QSSVM on its own data to find the local anomalous observations and the parameter R_j of the QSSVM that separate normal observations from anomalous.
- Every sensor sends the local parameter R_j to the parent node in the structure.
- The parent node collects the radii of the children nodes and constructs the global radius R_m.
- The parent returns the global radius parameter R_m to its children nodes to detect anomalies in their data using this radius.

Variety of distributed models was designed based on the QSSVM [3, 7-10]; however, the common limitation of these models is the prohibited computational cost of applying the QSSVM on the limited sensor resources. Moreover, the dependency of SVM on some parameters substantially affects the detection effectiveness. These parameters need to be chosen for each WSN application and therefore affect the practicality of the solution. The authors of [15] proposed statistical based outlier detection methods based on temporal and spatial correlations of sensor data. The performance of these methods is highly depending on some statistical parameters that affect the usability of these methods for online detection.

In this paper, a new distributed online anomaly detection model is designed using the one-class principal component classifier (OCPCC) previously proposed in [11] to detect anomalies as they occurred. The difference between the model proposed in this paper and the model in [11] is the selection of detection threshold and the variant of PCA algorithm used. In [11], the threshold was selected by clustering the dissimilarity vector of training data using automated clustering threshold procedure which has high computational cost to be used for online detection. In addition, the called candid-covariance free incremental principal component analysis (CCIPCA) algorithm which was originally proposed in [12] and utilized for data reduction for WSN in [13] was used as a core for the design of the proposed model in this paper. Data samples chosen from sensorscope GSB[14] dataset were used to validate the detection effectiveness of the model in terms of detection rate and false alarm rate. A comparison with an existing distributed model in [15] was conducted to show the advantages of the proposed model.

The rest of paper is structured as the following: section 2 details out the design of the proposed model. Section 3 provides the experimental results and discussion. Section 4 concludes the paper and suggests some future research directions.

2. The proposed model

Two main stages are involved in the design of the proposed model which are training stage and detection stages. Details of each stage are given in the following sections.

2.1. Training stage

In this stage, the normal observations are gathered at every sensor to find out the local normal model (LNM) and send it to the cluster head (CH) for constructing the global normal model (GNM). The procedure of building the LNM is as follows: the training normal observations S_{Train} are standardized by the mean value (μ) and the standard deviation value (σ) of training observations. The eigenvector (V_i), and eigenvalues (λ_i) matrices are then calculated from the standardized observations. The projection of training observations on the principal component space Y_i is then computed by Eq. (1):

$$Y_i = S_{Train}(i) \times V(i) \tag{1}$$

The number of principal component which depends on the application is determined and the D_{train} dissimilarity measure is computed for training observations using Eq. (2).

$$D_{train} = \sum \frac{y_i}{\lambda_i} \tag{2}$$

In this paper, the detection threshold that represents the local normal model (LNM) is chosen to be the maximum *MaxDiss* and minimum *MinDiss* bounds of the dissimilarity vector D_{train}. The LNM parameters are then sent to CH to construct the GNM. Different strategies are used to build the GNM at CH; however, the details of such strategies are not the subject of this paper.

2.2. Detection stage

In this stage, each node tests every observation using the GNM built in training stage. Some other parameters which include the mean (μ) and standard deviation (σ) parameters, eigenvector (V_i) and eigenvalues (λ_i) matrices, are also used in this stage. The GNM model is composed of (*MaxDiss, MinDiss)* values calculated globally at CH. Each observation is identified as either normal or anomalous by comparing its projection on the space with the detection thresholds specified in the GNM model. The score of projecting every new observation on the principal component space is computed using the parameters of the normal profile that were obtained from training stage as in Eq. (3):

$$Y_i = S_{Test} \times V \tag{3}$$

The measure of dissimilarity (D_{test}) for every new observation is computed using Eq. (4).

$$D_{test} = \sum \frac{y^2}{\lambda} \tag{4}$$

At the end, the value of D_{test} parameter is compared against the GNM values to classify each observation as normal or anomalous using the following criterion:

$$(D_{test} > MaxDiss) Or (D_{test} < MinDiss) \rightarrow Anomalous$$
$$else \rightarrow Normal$$

3. Experimental results and evaluation

In this section, the experimental results of the model proposed in this paper on real world data set are presented and evaluated by comparing it to a recent existing anomaly detection model from the literature.

3.1. Datasets

Grand St. Bernard (GSB) [14] is one of sensorscope project deployment dataset was gathered using WSN deployment at the Grand St. Bernard pass that is located between Italy and Switzerland. The network is formed of 23 sensors that record metrological environmental data that include; (ambient and surface) temperature, and humidity. 2 clusters were formed in this deployment in which the small cluster has 5 sensors and the big cluster has 18 sensors. The observations of the small cluster which have 5 sensors namely, Node-25, Node-28, Node-29, Node-31 and Node-32 were used to evaluate the proposed model. In our experiments, the ambient temperature observations in the period 6am-14pm of data recorded on 5th March 2007 were used. The observations are labeled using the Mahalanobis distance measure as used in [15].

3.2. Results and discussion

We have tested our proposed model on dataset samples extracted from nodes 25,28,29,31 and 32. The hierarchical WSN structure was adopted in which one of the above nodes act as CH that receives the LNM from other nodes in the cluster and constructs the GNM that is used by other cluster nodes for detection. The detection rate (DR) and false alarm rates (FPR) which represent detection effectiveness are reported in Table 1.

Table 1. Detection rate and false alarm rate of the proposed model

	Nodes					
	N25	N28	N29	N31	N32	Average
DR (%)	100	91	100	89	100	96
FPR (%)	16	0	2	0	0	7.2

Table 1 reports the results of the proposed model for each node. The results show that an average of 96% detection rate was achieved by the model while 7.2% false positives were produced from misclassification of some normal observations as anomalies. A comparison with the statistical temporal-based outlier detection (TOD) model proposed in [15] using the same data samples and same data labeling approach is shown in Table 2.

Table 2. A comparison with the TOD model proposed in [15]

	SW=15	SW=30	SW=48	SW=60	Our model
DR (%)	82.86	82.86	82.86	82.86	96
FPR (%)	13.3	13.48	11.67	10.13	7.2

Table 2 reports the results of the TOD model with different smoothing windows values. It is clearly shown that our proposed model outperforms the TOD model in terms of detection rate and false positive rate. To evaluate the efficiency of the model, the computational complexity and communication overhead are considered. The computational complexity incurred by our model is *O(N)* where *N* is the number of observed variable in the phenomenon. This complexity is the computational complexity of CCIPCA algorithm. The additional calculations involved in the model are fixed. Since the LNM sent by each node has a fixed size, the communication overhead is a function of the number of nodes *K* in each cluster; therefore, the communication overhead is *O(K).* For the TOD model, there is no communication overhead as the method is implemented locally but the computation cost is *O(c. d .p)* where *d* is the number of variables, *c* is the number of original observations to be modeled, and *p* is the fitting parameter of the AR model used by the TOD.

4. Conclusion

This paper introduces the preliminaries and initial experimental results of new distributed anomaly detection model for WSNs. The model was designed based on the calculation of the dissimilarity between sensor observations in the principal component space. Experimental results and comparison with recent existing work indicate that the new model is promising in terms of achieving high detection effectiveness while efficiently utilizing the limited network resources. Further experiments are needed to investigate the performance of the model with multivariate data and additional real world datasets.

Acknowledgements

This study was supported by Research Management Center (RMC) at Universiti Teknologi Malaysia (UTM) under the RUG grant (VOTE NO. 08J95).

References

[1] Akyildiz I.F, Su W, Sankarasubramaniam Y , Cayirci E. *Wireless sensor networks: a survey.* Computer Networks 2002. 38(4); 393-422.

[2] Hodge V, Austin J. *A Survey of Outlier Detection Methodologies.* Artif. Intell. Rev 2004. 22(2); 85-126.

[3] Zhang Y, Meratnia N, Havinga P. *An online outlier detection technique for wireless sensor networks using unsupervised quarter-sphere support vector machine*. In *International Conference on Intelligent Sensors, Sensor Networks and Information Processing,* 2008.

[4] Palpanas T, Papadopoulos D, Kalogeraki V, Gunopulos D. *Distributed deviation detection in sensor networks.* SIGMOD Rec., 2003. 32(4);77-82.

[5] Branch J, Szymanski B, Giannella C, Wolff R,Kargupta H. *In-Network Outlier Detection in Wireless Sensor Networks*. In *26th IEEE International Conference on Distributed Computing Systems, 2006.*

[6] Rajasegarar S, Leckie C, Palaniswami M, Bezdek J. *Quarter Sphere Based Distributed Anomaly Detection in Wireless Sensor Networks*. in *IEEE International Conference on Communications, ICC '07.* 2007.

[7] Shahid N, Naqvi I.H, Qaisar S.B. *Quarter-Sphere SVM: Attribute and Spatio-Temporal correlations based Outlier & Event Detection in wireless sensor networks*. In *2012 IEEE Wireless Communications and Networking Conference (WCNC),*2012.

[8] Takianngam S, Usaha W. *Discrete Wavelet Transform and One-Class Support Vector Machines for anomaly detection in wireless sensor networks*. In *2011 International Symposium on Intelligent Signal Processing and Communications Systems (ISPACS),* 2011.

[9] Rajasegarar S, Leckie C, Bezdek, J.C, Palaniswami M. *Centered Hyperspherical and Hyperellipsoidal One-Class Support Vector Machines for Anomaly Detection in Sensor Networks.* IEEE Transactions on Information Forensics and Security, 2010. 5(3); 518-533.

[10] Zhang Y, Meratnia N, Havinga P.J. *Ensuring high sensor data quality through use of online outlier detection techniques.* International Journal of Sensor Networks, 2010. 7(3); 141-151.

[11] Rassam M.A, Zainal A, Maarof M.A. *One-Class Principal Component Classifier for Anomaly Detection in Wireless Sensor Network.* In 2012 Fourth International Conference on Computational Aspects of Social Networks (CASoN), 2012; 271-276.

[12] Juyang W, Yilu Z, Wey-Shiuan, H. *Candid covariance-free incremental principal component analysis.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 2003. 25(8); 1034-1040.

[13] Rassam M.A, Zainal A, and Maarof M.A. *An Adaptive and Efficient Dimension Reduction Model for Multivariate Wireless Sensor Networks Applications.* Applied Soft Computing, 2013. 13 (4): 1878-1996.

[14] GSB, *Grand-St-Bernard (GSB) dataset. http://lcav.epfl.ch/cms/lang/en/pid/86035*. 2007.

[15] Zhang Y, Hamm N, Meratnia N, Havinga P. *Statistics-based outlier detection for wireless sensor networks.* International Journal of Geographical Information Science, 2012. 26(8);1373-1392.