***An Improved Distributed PCA-Based Outlier Detection in Wireless Sensor Network***

ZHENG Wentian

*(College of Compute and Software Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210046, China)*

***Abstract***—Outlier detection in wireless sensor network is essential to ensure data quality, secure monitoring and reliable detection of interesting and critical events. Principal Components Analysis(PCA) have received a great interest in the machine learning community since their introduction, especially in Outlier Detection in Wireless Sensor Networks (WSN). An efficient and effective methods called Improved Distributed PCA-Based Outlier Detection (IDPCA) has been proposed in this paper. The proposed method operates in every sensor nodes respectively, thus reduces the communication overheads and prolong the network’s life. We take advantage of spatial correlations that exist in sensor data of adjacent nodes to reduce the false alarm rate and make accurate distinction between events and errors in real-time. Experiments with both synthetic and real data collected from the Intel Berkeley Research Laboratory indicates that IDPCA achieves a higher detection rate with a lower false alarm rate, while minimizes the communication overhead than previous methods.

***Keywords*:** wireless sensor network, outlier detection, principal components analysis, spatial correlation

(I)INTRODUCTION

Wireless sensor networks (WSNs) have become a growing area of research and development over the past few years. Generally, a wireless sensor network consists of a hierarchical or nonhierarchical structure of low-cost and low-power sensor nodes which are capable of sensing various attributes of an environment under observation [1]. Most WSN’s applications etc. precision agriculture require precise and reliable information to provide for the end user [2]. However, raw sensor observations collected from these nodes suffer from low data quality and reliability due to the harsh and unattended environmental effects, malicious attacks and resource constrains such as energy, memory, computation ability. To ensure the quality of sensor measurements, outlier detection methods allow cleaning and refinement of collected data and let providing the most useful information to end users, while maintaining low energy consumption and preserve high computational efforts due to the limited energy resources of sensor nodes.

In WSNs, outliers also known as anomalies are those data measurements deviate from the normal behavioral pattern of the sensed data [3]. Therefore, a straightforward method for outlier detection in WSN is to define a normal behavior model of sensors data and consider those observations have significant difference from the defined normal behavior as outliers. An elaborate introduction of the outlier detection techniques in WSNs has been discussed in [4] and [5], mainly include: Statistical-Based Approaches, Clustering-Based Approaches, Nearest Neighbor-Based Approaches, Classification-Based Approaches and Spectral Decomposition-Based Approaches. Unfortunately, these methods are constrained by two main drawbacks. On the one hand, existing methods mainly belongs to centralized approaches which accounts for a great communication and computation overload to the network. On the other hand, those techniques often do not distinguish errors and events and regard outliers as errors, which results in loss of important hidden information about events [4].

Principal component analysis (PCA) is a powerful technique for analyzing and identifying patterns in data [6]. It finds the most important axis to express the scattering of data. By using PCA, the first principal component (PC) is calculated, which reflects the approximate distribution of data. In this paper, a novel outlier detection approaches named Improved Distributed PCA-Based Outlier Detection (IDPCA) has been proposed. We partition a WSN into several groups. In each group, IDPCA operates automatically and find the conformed sensor nodes in the group. We take advantage of spatial correlations that exist in sensor data of adjacent nodes to reduce the false alarm rate and make accurate distinction between events and errors in real-time. Experiments with both synthetic and real data collected from the Intel Berkeley Research Laboratory show that IDPCA achieves a higher detection rate with a lower false alarm rate, while minimizes the communication overhead than previous methods.

The rest of this paper is organized as follows: Section II review some related work, Section III formally states the problem of anomaly detection in WSN. Section IV describe proposed IDPCA method, Section V presents the experimental results and Section VI finally draw some conclusions.

(II) RELATED WORK

*A Existing Anomaly Detection Methods*

As described in the first part, anomaly detection techniques can be classified into statistical-based, nearest neighbor-based, clustering-based, and classification-based approaches.

Statistical-based methods are the earliest method to deal with the outlier detection problems [7,8]. The essential principle behind statistical models is to estimate a statistical normal model in the form of probability distribution which represents the distribution of data in a reference model and evaluates each pattern with respect to that model. Any deviation from the reference model is considered as outlier. However, they rely heavily on the correctness of probability distribution model, so they are not useful since in many real-life scenarios, no a priori knowledge of the sensor stream distribution is available.

Nearest neighbor-based approaches are the most commonly used approaches to analyze a data instance with respect to its nearest neighbors in the data mining and machine learning community. They assume that the normal patterns of data are always found in a dense neighborhood while the anomalous ones are far from their neighbors [9]. Unfortunately, the computation of the distance between data patterns in multivariate datasets is very expensive which could not meet the requirement of low computation resources for WSNS.

Classification-based models are important models of machine learning and data mining community mainly include Support Vector Machine-Based(SVM) approaches and Bayesian Network-Based Approaches. They learn a classification model by a couple of training data instances, afterwards, distinguish an unseen observation into learned class. There are two main drawbacks of these approaches (1) expensive computation resources (2) difficult to acquire the labeled data. Therefore，a novel unsupervised method called one-class classifier which can learn a boundary between normal data and outliers with unlabeled training data has been proposed [10]. Although, One-class classifier is more suitable for WSNS，they are also constrained by some limitations; Such as, the one-class SVM classifier highly affected by the parameters and choice of the proper kernel function. Once an improper kernel function or parameters is selected, there will be a high false alarm rate and low detection rate in the network.

*B PCA-Based Models*

Principal Component Analysis (PCA) is a multivariate data analysis technique used for reducing the dimensionality of the set of correlated data observations by transforming them into a set of uncorrelated variables called Principal Components (PCs) [11]. Researchers have proposed both centralized(CPCA) and distributed PCA-Based anomaly detection methods(DPCA) in [12]. However, they cost high communication overheads and suffer from a comparative high false alarm rate. In [13], authors introduced how to detect outliers and identify faulty nodes using PCA. This model showed two types of analysis: offline analysis and real-time analysis. More recently, Kernel Principal component analysis (KPCA) has used for nonlinear case which can extract higher order statistics. Due to the attractive capability, KPCA-based methods have been extensively investigated, and have showed excellent performance [14,15]. Regrettably, they are also suffering from huge computation pressure and highly depend on the selected kernel function.

Compared to previous PCA-based detection approaches, our IDPCA methods possess a better performance on detection rate and false alarm rate with comparative low communication overloads, moreover, by taking fully advantage of the spatial correlation among the neighbor nodes, we distinguish the outliers from events and errors accurately.

(III)Problems Statement

We consider a WSN composed of a set of sensor nodes deployed in a homogenous environment. The sensor nodes are synchronized and their sensed data belong to the same unknown distribution. Then we partition the network into several groups. Each group consists of a cluster header and a couple of members. It’s generally assumed that nodes in the same group monitor similar readings and within the transmission range of the cluster header. In this paper, we will not concern so much about the cluster approaches and consider the network has already been correctly partitioned.

Let be a group of sensor nodes where  selected as the group head. At every time intervaleach sensor node in the set  measures a data vector. Letdenote the data vector measured atrespectively. Each data vector comprises of features:,A subset  represents a closed neighborhood of a cluster header nodes shown as Figure1.



**Figure.1** A closed neighborhood of 

According to the requirements of applications, outliers can be classified into local outlier and global outlier. Local outliers represent those outliers that are detected at individual sensor node only using its local data. Global outliers represent those outliers that are detected in a more global perspective by considering a cluster of sensor nodes. Our goal is to identify the newly coming data measurement atas normal or anomalous in real-time by local detection and global detection. In addition, we conducted further research about the source of outliers and thus make real-time distinction between events and errors.

(IV) IDPCA Outlier Detection Techniques

In this section, we describe the improved distributed PCA-based outlier detection techniques in detail. The approach consists of four phases: training, outlier detection, outlier source detection, updating.

1. ***Training phase***

The training phase aims to establish a normal profile model at each sensor node respectively. Let be the data vectors collected from the normal network traffic by sensor node  in  time windows, wherecan be written as:

|  |  |
| --- | --- |
|  | (1) |

eachis comprised of features. first normalizeto a range [0,1] and then computes the column-centered matrix of it:

|  |  |
| --- | --- |
|  | (2) |

where is the column means of  and is a vector with the length . The principal components (PCs) of are given by a singular value decomposition (SVD) of :

|  |  |
| --- | --- |
|  | (3) |

Where is the matrix of PCs of and is the diagonal matrix of eigenvalues arranged from maximum to minimum. The first PC of is denoted as. Afterward,calculates the projection distance of each feature vector  from  (see Fig2):

|  |  |
| --- | --- |
|  | (4) |



**Figure.2** Projection distance of a feature vector  from the first PC

The maximum projection distance of all feature vectorsfromdefined as: ,Finally, each sensor node uses the triple to establish the normal profile prepare for the detection phase.

1. ***Outlier detection phase***

By exploiting the high degree spatial correlations among neighbor nodes in a densely deployed WSN, each node possesses sufficient information to detect outliers. The detection phase not only depends on the individual criteria of its own, but also supported by the criteria provided by the spatially neighboring nodes. Therefore, it can reduce false alarm rate in some degree.

The pseudocode of our proposed outlier detection technique is shown in Table (1). Initially, each node acquires the first primary component and the maximum projection distance  using its sequential data

Table 1. pseudocode of IDPCA approach

|  |
| --- |
| 1. letbe the max projection distance from the first pc of node; |
| 2. letbe the median projection distance of ’s neighboring nodes’; |
| 3. letbe the median projection distance of the set ; |
| 4. letbe the amount of data measurements for learning theand first pc; |
| 5. letbe a new data measurement arrive at ; |
| 6. letbe the data vectors arriving at’s neighboring nodes at the same time interval |
| 7. let be the projection distance from the first pc at node in time windows t; |
| 8. let  be the median projection distance offrom their own first pc in time windows t; |
| 9. procedure learning and first pc  ①each node collects  data measurements for learning its own first pc and compute, then broadcasts the distance information to its group node ;  ②each group node computesand;  ③initiate IsOutlier(,) for each node;  return; |
| 10. procedure IsOutlier(,) for  when arrives at ,computes ;  if (>&&>)  indicates an outlier;  SourceOfOutlier(,,,) for;  else  indicates a normal one;  endif  return; |
| 11. procedure SourceOfOutlier(,,,);  ①collects the projection distance from its neighbor respectfully and compute the median projection distance;  ②if (>&&>)  if (>&&>)  may indicates an event;  else  may indicates an erroneous data measurement;  endif;  else  may indicates an erroneous data measurement;  endif;  return; |
| 12. procedure Updating global maximum projection distance using the newly m time windows |

vectors. Afterwards, every node locally broadcast its maximum projection distance to his spatially neighboring group node  which then computes the median projection distance as the global maximum distance . Finally, a median distanceof its closed neighborhood is calculated for source of outlier detection. One should note that to estimate the ‘center’ of the data set, the median is more robust than the mean.

When a new data measurementarrives at the node, it computes the projection distance firstly, then a comparison betweenandis executed. If,then the data can be considered as a normal data. Otherwise, it may be thought as a potential outlier. In this case, further compares thewith the  computed by the group head node. If,will finally be classified as an outlier in the set. Thus, the decision function can be formulated as:

|  |  |
| --- | --- |
|  | （5） |

Where the data measurement equals -1 will be considered as an outlier.

1. ***Outlier source detection phase***

Identifying what has caused the outlier in sensor data is an important task. Potential sources of outliers in WSNs include noise & errors, actual events, and malicious attacks. In this paper, we mainly talk about how to distinguish outlier sources between errors and actual events. Our proposed technique provides a preliminary method to make real-time distinction between events and errors by exploiting the spatial correlation of sensor data among neighboring nodes. The main idea is that if a data observation is considered as an outlier by, then the group head node collects the currently data vector’s projection distance from its neighboring nodes and compute the median distance. If an event occurred in ,andwill be temporally different, this means thatandwill both exceed their maximum projection distance respectively.

1. ***Updating Phase***

There might be changes over time in the conditions of the environment in which a WSN is deployed. Therefore, it is necessary to update the global normal pattern. Let  be the current time window, to update the global  every member node needs to calculate the local  on the normal dataset collected at the  previous time windows. As shown in Figure3, the purpose of this phase is to reduce the importance of normal dataset in old windows and improve the accuracy of training model by real-time data vectors.



Figure.3 Updating the global normal pattern

(V) **Experiment result**

To validate the proposed detection model, we do some experiments with Matlab in this section. Our aim is to compare the performance of the proposed IDCPA with DPCA and CPCA. In our experiments, we have used synthetic data as well as real data gathered from a deployment of WSN in the Intel Berkeley Research Laboratory. A brief description of each dataset in the following subsections.

1. synthetic Dataset

For the sake of obtaining a general idea of the performance of IDPCA, the 3-D synthetic dataset which composed of a mixture of three Gaussian distribution with uniform outliers is used for each node, the mean is randomly selected from (0.3, 0.35, 0.4) and the standard deviation is selected as 0.03. We choose 1000 normal data vectors for the training phase in 4 time windows and 200 normal data vectors with 50 artificial outliers in every time window for the testing phase, the size of each time window is set to 130 minutes and the third component of the anomalous data vector is uniformly distributed in the interval [0.5, 0.7]. A distribution of the synthetic training and testing data is plotted as shown in Figure4.

1. IBRL Dataset

The IBRL dataset collected from a closed neighborhood from a WSN deployed in the Intel Berkeley Research Laboratory is commonly used



Fig.4 plot for synthetic training and testing data

to evaluate the performance of most existing anomaly detection models in WSNs. Four kinds of data records are measured by the network; temperature, humidity, light and voltage. The measurements were collected every 31 seconds intervals. In our simulation, we consider a group of nodes as shown in Figure5. The closed neighborhood contains the node 35 and its 6 spatially neighboring nodes, namely nodes 1, 2, 33, 34, 36, 37. We use a 9am-17pm period of data recorded on 28th February 2004 with two attributes: temperature and humidity for each data measurement. However, the IBRL dataset is a collection of normal data measurements. To evaluate the anomaly detection models using this dataset, some artificial anomalies are injected. This procedure is common in many of proposed anomaly detection models for WSNs in the literature. In the meantime, for the purpose to have a more intuitive comparison of false alarm rate among three approaches, we have introduced some Gaussian noise to the normal data, and the intensity of the noise is measured by signal-to-noise(SNR).

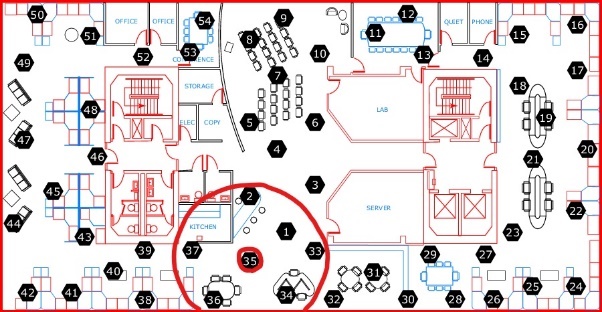


Fig.5 sensor nodes deployment in IBRL dataset

1. Performance Analysis

We used two performance measures: detection rate(DR) and false alarm rate(FAR). The detection rate is defined as the percentage of anomalous data vectors that are successfully detected. The false alarm rate is defined as the percentage of normal data vectors that are incorrectly detected as anomalous.

Figure6 describes the performance of our proposes IDPCA model by using synthetic dataset. As the figure shows that IDPCA can achieve a satisfactory detection rate with average value about 96% and a comparative lower false alarm rate with 1.5% in different time windows which indicates that the proposed approach could detect most of the anomaly data in the network while maintain a low probability to identify the normal data as outliers.



Fig.6 Performance of IDPCA on synthetic data

Figure7 presents the detection rate of applying the proposed IDPCA model and DPCA, CPACA models presented in [12] on the IBRL dataset. The results show that during the given time windows the proposed IDPCA model outperforms the DPCA and CPCA models, as it achieves an average detection rate about 98.5% higher than that of DPCA and CPCA about 95%. In the experiment, the size of time windows is set to 130 minutes and the SNR is 20dB.

Figure8 compares the false alarm rate of applying IDPCA，DPCA and CPCA models on the IBRL dataset with SNR varies from 5dB to 45dB. As the figure shows, with the increase of SNR, all of the three approaches achieve a lower false alarm rate and the proposed IDPCA possess the lowest false alarm rate in any SNR. It reveals that under the same environment conditions (intensity of noise), the false alarm rate of IDPCA is superior to another two models.



Figure.7 DR of different approaches



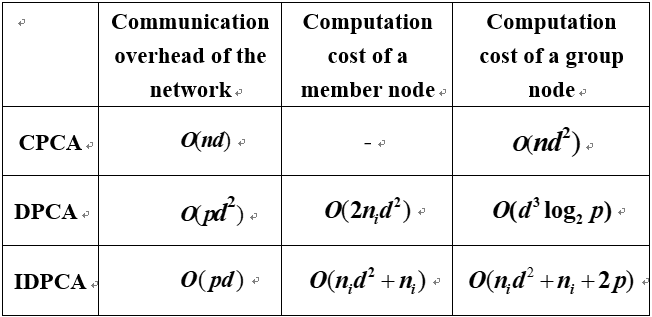
Figure.8 FAR of different approaches

From the experiment results above, we may find that our proposed IDPCA model is more effective and efficient than previous method which availably increase the percentage of outlier detection rate and decrease the false alarm rate.

1. Complexity Analysis

In this section we analyze the computational cost and communication overhead of the proposed Improved distributed approach in detail. The cost of DPCA and CPCA have been described clearly in [12]. To normalize the data vectors, every member node in the group sends a pair of minimum and maximum vectors to the group head firstly. Afterwards, the group head return the global minimum and maximum vector to the member nodes. Every member node normalizes the data vectors and then compute the column-centered matrix of it, finally the PCs of  is given by a singular value decomposition. After getting the first primary component, every member node computes its respectively and sends it to the group head which obtains the global followed by. Hence, during the period every node has 5d communication cost, there are p nodes, so the final communication cost is. Every member node needs to compute covariance matric and the first PC by SVD which require a cost of [13]. In addition, to acquire the an extra cost of  is needed, so the total cost of computation is for the member nodes. As for the group node, it needs extra computation cost to find the global  and identify whether an event happened, so the final computation cost of the group head is . A comparison of the main costs of IDPCA, CPCA and DPCA is shown in table 2.

Table.2 Comparison of different approaches on Computation costs and Communication overhead





(VI) **Conclusions**

In this paper, an efficient and effective methods called Improved Distributed PCA-Based Outlier Detection (IDPCA) has been proposed in the Wireless Sensor Network. In the approach, we partition the sensor nodes into several groups, each group measure similar attributes and possess a group head. Rather than sending all data vectors to the group to acquire the global maximum distance, the proposed method operates PCA in every sensor nodes respectively and send the local maximum distance to the group head. Moreover, by taking advantage of spatial correlations that exist in sensor data of adjacent nodes the proposed method can make accurate distinction between events and errors in real-time. Experiments with both synthetic and real data collected from the Intel Berkeley Research Laboratory indicates that our proposed IDPCA approach achieves a higher detection rate with a lower false alarm rate, while minimizes the communication overheads than previous methods. Our future work will mainly focus on the reduction of computation complexity of proposed method.

**References**

[1] Shahid, N. and I. H. Naqvi (2011). Energy efficient outlier detection in WSNs based on temporal and attribute correlations. 2011 7th International Conference on Emerging Technologies.

[2] Ghorbel, O., et al. (2015). A Novel Outlier Detection Model Based on One Class Principal Component Classifier in Wireless Sensor Networks. 2015 IEEE 29th International Conference on Advanced Information Networking and Applications.

[3] Zhang, Y., et al. (2009). Adaptive and Online One-Class Support Vector Machine-Based Outlier Detection Techniques for Wireless Sensor Networks. 2009 International Conference on Advanced Information Networking and Applications Workshops.

[4] Zhang, Y., et al. (2010). "Outlier Detection Techniques for Wireless Sensor Networks: A Survey." IEEE Communications Surveys & Tutorials 12(2): 159-170.

[5] McDonald, D., et al. (2015). "A Survey of Methods for Finding Outliers in Wireless Sensor Networks." Journal of Network and Systems Management 23(1): 163-182.

[6] M. Ahmadi Livani and M. Abadi, “An energy-efficient anomaly detection approach for wireless sensor networks,” in Proceedings of the 5th International Symposium on Telec-ommunications, Tehran, Iran, December 2010.

[7] S. M. A. Bettencourt, A. A. Hagberg, and L. B. Larkey, "Separating the wheat from the chaff: practical anomaly detection schemes in ecological applications of distributed sensor networks," presented at the Proceedings of the3rd IEEE international conference on Distributed computing in sensor systems, Santa Fe, NM, USA, 2007.

[8] B. Sheng, Q. Li, W. Mao, and W. Jin, "Outlier detection in sensor networks," presented at the Proceedings of the 8th ACM international symposium on Mobile ad hoc networking and computing, Montreal, Quebec, Canada, 2007

[9] V. Chandola, A. Banerjee, and V. Kumar,"Anomaly Detection: A Survey," ACM Computing Surveys, 2009

[10] Zhang, Y., et al. (2013). "Distributed online outlier detection in wireless sensor networks using ellipsoidal support vector machine." Ad Hoc Networks 11(3): 1062-1074.

[11] Rassam, M. A., et al. (2012). One-Class Principal Component Classifier for anomaly detection in wireless sensor network. 2012 Fourth International Conference on Computational Aspects of Social Networks (CASoN).

[12] Livani, M. A. and M. Abadi (2010). Distributed PCA-based anomaly detection in wireless sensor networks. 2010 International Conference for Internet Technology and Secured Transactions.

[13] Fernandes Jr, G., et al. (2015). "Autonomous profile-based anomaly detection system using principal component analysis and flow analysis." Applied Soft Computing **34**: 513-525.

[14] Ghorbel, O., et al. (2015). A Novel Outlier Detection Model Based on One Class Principal Component Classifier in Wireless Sensor Networks. 2015 IEEE 29th International Conference on Advanced Information Networking and Applications.

[15] Ghorbel, O., et al. (2016). One class outlier detection method in wireless sensor networks: Comparative study. 2016 24th International Conference on Software, Telecommunications and Computer Networks (SoftCOM).