Dr Ashwini K B et al, International Journal of Computer Technology & Applications,Vol 7(6),768-772

# An Effective Outlier Detection-Based Data Aggregation for Wireless Sensor Networks

Dr Ashwini K B[1]
[1]R V College of Engineering
[1]Master of Computer Applications
[1] Bangalore, India
[1]ashwinikb@rvce.edu.in

Dr Usha J[2]
[2]R V College of Engineering
[2]Master of Computer Applications
[2] Bangalore, India
[2]ushaj@rvce.edu.in

## Abstract

*Data aggregation protocols are essential for wireless sensor networks to reduce energy consumption and prolong network lifetime. However for wireless sensor networks, not only the energy consumption of sensor nodes but also the correctness of the data aggregation results is critical. As wireless sensor networks are usually deployed in harsh and hostile environments, malfunctioning and compromised sensor nodes negatively affect the correctness of the data aggregation results. This paper presents data aggregation scheme that eliminates the outliers, and then, it determines the sensor nodes that have distinct sensed data and collects only one sensor node that has the actual data for each distinct sensed data, and the data aggregator does not accept data from any other sensor nodes. This process ensures that (i) no outlier data is included in the aggregated data and (ii) there is no redundant data with the data aggregator. The simulation results show that the proposed scheme is able to reduce the number of false data transmissions, thereby increasing the data aggregation accuracy.*

## 1. Introduction

WSN is typically composed of large number of sensor nodes which are scattered in the sensor field to measure quantitative data. Sensor nodes in WSN generate a large amount of data that must be communicated to the base station using radio transmission. Sensor nodes rely on small batteries and are usually capable of measuring physical phenomena such as temperature, sound, vibration and pressure. Hence, a WSN must perform the data gathering task in an energy efficient manner so that its lifetime is prolonged. Data aggregation is implemented in WSNs to eliminate data redundancy, reduce data transmission, and improve data accuracy. It is shown that data aggregation results in better bandwidth and battery utilization [1,2] which enhances the network lifetime because communication constitutes 70% of the total energy consumption of the network [3]. In WSNs, data aggregation is performed by sensor nodes, called data aggregators. Data aggregators are responsible not only for collecting and summarizing data but also for in-network analysis of the collected data, and trigger alarms on the basis of this analysis [4]. In particular, since sensor nodes are placed in outdoor for applications such as disaster monitoring and habitat monitoring, a sensor node can malfunction or sensor readings may be incorrect due to external impact or severe external environment. Some sensor node readings may significantly vary due to sudden change in environments. These abnormal sensor readings are called the outliers.

For example, assume that WSN consisting of several sensor nodes dispatched in a forest to monitor forest fire. Assume the forest area is divided into clusters when an aggregated value is send by the cluster head to the forest guard, he can identify the actual forest fire or can initialize the sensor node to check if an outlier is generated. Thus the outlier detection is quite important task to detect an event or maintain sensor networks harmoniously. It is clear from the aforementioned discussion that outlier detection mechanisms must be implemented in WSNs so that data aggregators, that is, decision makers, can correctly trigger alarms. However, outlier detection process is a memory consuming and communication-consuming task by its nature [5]. In distributed and resource-constrained environments, such as WSNs, identifying accurate outliers is a challenging task. Moreover, data aggregation accuracy must be maintained.

The rest of the paper is organized as follows. In section2 the related work in data aggregation and outlier detection in WSN domain is presented. In section3 the existing system model is presented followed by problem definition in section4. We have a

proposed model in section5. Outlier detection and data aggregation is proposed in section6 followed by performance evaluation in section7. Finally concluding remarks are made in section8.

## 2. Related Work

In WSN a lot of outlier detection techniques have been proposed. Paper [6] introduced a framework for cleaning and querying noisy sensors. The authors presented an in-network Bayesian approach to reduce the uncertainty of the data due to random noise. To obtain a better estimation of the sensor node readings, the authors combined the prior knowledge of the real sensor reading, the noise characteristics of the sensor node, and the observed noisy reading. The authors proposed several algorithms based on the introduced uncertainty models and evaluate the proposed algorithms. A comprehensive survey of outlier detection techniques is presented in [7]. To detect outliers in WSNs, the authors of [8] investigated the augmentation of sensor network queries by statistical models. The authors argued that a statistical model may offer a more reliable way to gain insight into the physical phenomena observed. Using statistical models, the authors propose an approach to detect outliers in streaming sensor data. The authors of [9] proposed a histogram-based method to detect outliers in a communication efficient manner. A declarative data cleaning mechanism over sensor node data streams is introduced in [10]. A fuzzy logic-based approach is proposed in [11] to infer the correlation among measurements from different sensors. The proposed technique assigns a confidence value to each measurement and then performs an aggregated weighted average scheme. The authors of [12] proposed a technique based on a weighted moving average that takes into account both recent local samples and corresponding values by neighbouring sensor nodes to estimate actual sensor readings. Localized voting protocols are used in [13] and [14] to identify the faulty sensors. However, the authors of [15] have shown that localized voting schemes are prone to errors if there is no direct communication among sensor nodes that produce the faulty data. In [16], an outlier detection method for real-time events in WSNs is proposed. The proposed method trains and tests the data in real time and has shown to be effective. In [17], an outlier detection protocol that is based on time-series analysis and geostatistics is proposed. The authors presented that the proposed protocol accurately detected outliers in WSN data, taking advantage of their spatial and temporal correlations. In [18], outlier detection techniques for WSN localization problems are investigated, and an outlier detection scheme to cope with noisy sensor data is proposed. Different from the existing work, in this paper a novel method is employed to improve outlier detection scheme and improve the accuracy and efficiency of the data aggregation process.

## 3. System Model

We consider a large sensor network with densely deployed sensor nodes that are assigned unique identification numbers. Sensor nodes have limited computation and communication capabilities. The network is divided into clusters, and each cluster has a dynamically selected data aggregator node. Because of the dense deployment, sensor nodes have overlapping sensing regions and events are detected by multiple sensor nodes, thereby requiring data aggregation to reduce the amount of data transmission. Data are periodically collected and aggregated in data aggregation sessions. The data aggregator sends aggregated data to the base station. We assume that each data aggregator aggregates its cluster data only and hierarchical data aggregation is not allowed.

## 4. Problem Definition

Assume sensor nodes Pi and Pj belong to same cluster P they may be labelled as an outlier because of an event that occurred in the neighbouring cluster. For example, consider the fire monitoring scenario given in Figure 1 where clusters P, Q and R form neighbouring cluster groups. When a fire started inside cluster S, it is expected that the sensor nodes of clusters P and Q that are located close to cluster S detect the fire as well. Temperature readings of such nodes deviate significantly from the temperature readings of the other sensor nodes in clusters P and Q. As a result, data aggregator of clusters P and Q labels these nodes as outliers. This process is not sufficient to label a sensor node as outlier. Hence, the data aggregators should determine the outliers after communicating with its neighbouring data aggregators.

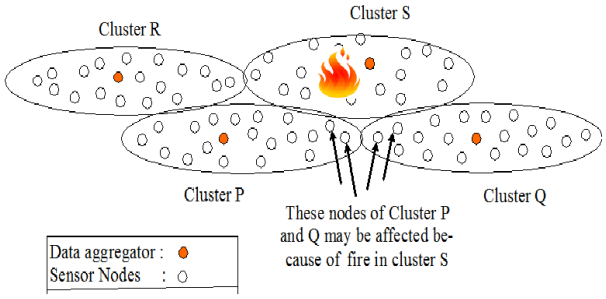


Figure 1

## 5. Proposed Model

The density-based clustering method search for the connected dense spaces that are separated by low dense spaces [19]. The basic idea of the density-based clustering is that for each data point in a cluster the neighbourhood of a given radius (Eps) must contain at least a minimum number of data points (MinPts) [20] , it showed that MinPts in the range of 10–20 can mostly result in good clustering outcomes. In this paper, MinPts is chosen as 16. For a given MinPts, Eps can be determined based on the k-distance graph, as suggested by Ester et al. [21]. The k-distance means the distance from a given point to its kth nearest neighbour point, where k equals to MinPts. k-distance sorts k-distance of all data points of concern. The value of Eps can then be determined at the place where the k-distance starts to change dramatically. If a point p is directly density-reachable from another point q, it must satisfy the below criteria [21].

Eq(1) $p \in NEps(q) \& |N_{Eps}(q)| \geq MinPts$

where, NEps(q) is the set of the data points within the Eps of the point q.

In this proposed algorithm the output is a list of ordered data points with respect to the reachability-distance and core-distance. The reachability-distance of a point p with respect to another point o is defined in Eq. (2), in which the core-distance of the point p is defined in Eq. (3)[20]. Figure 2 graphically demonstrated how the concepts are defined.

equation(2)

Reachability-distance

$$\begin{cases} \text{UNDEFINED}, \text{if} |NEps(O)| < MinPts \\ \max(\text{core} - \text{distance}(o), \text{distance}(o, p)), \text{otherwise} \end{cases}$$

equation(3)

Core-distance

$$\begin{cases} \text{UNDEFINED}, \text{if} |NEps(p)| < MinPts \\ MinPts - \text{distance}(p), \text{otherwise} \end{cases}$$

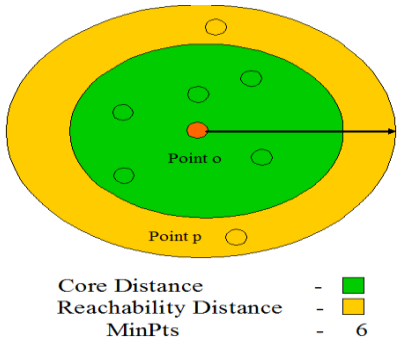


Figure 2

## 6. Outlier Detection and Data Aggregation

Sensor nodes of a cluster send sensed data to the aggregator node at regular intervals. The data aggregator looks for the following:

Phase 1:

If the sensed data of a node is found to be similar with another node, then its support count is increased by 1. The sensor nodes that have a support count, which is less than 1 are labeled as local outliers. These local outliers, however, might be affected by the events that occurred in the neighboring clusters. Therefore, neighboring data aggregators exchange their local outlier lists among them to determine if these outliers can improve their support count. Each data aggregator compares sensor's node data with its neighboring local outliers and updates their support counts. Neighboring data aggregators exchange the updated support counts of local outliers. Data aggregators check the updated support count of their local outliers, and they label the local outliers that have a updated support count less than 1 as outliers.

Phase 2

The data aggregator has the list of outliers and the sensor nodes that have the same sensed data. With this information, the data aggregator decides the sensor nodes that should send their actual data for data aggregation as follows. The data aggregator first eliminates the outliers, and then, it determines the sensor nodes that have distinct sensed data and collects only one sensor node that has the actual data for each distinct sensed data, and the data aggregator does not accept data from any other sensor nodes. This process ensures that (i) no outlier data is included in the aggregated data and (ii) there is no redundant data with the data aggregator. The data aggregator aggregates received data and sends aggregated data to the base station.

## 7. Performance Evaluation

In this section, we evaluate the proposed algorithm in terms of outlier detection performance and data aggregation accuracy. It is simulated using NS2 in a scenario where a cluster-based sensor network is deployed to monitor the temperature for forest fire. Hundred sensor nodes are placed in uniformly distributed random locations within a square area where the base station is located on one corner. There are four clusters in the network, and each cluster has a data aggregator node. Data aggregators reach the base station over a single hop. A synthetic data set in which sensor nodes generate false data with a probability of up to 10%. Each simulation is run 20 times, and the results are averaged.

### 7.1 Data Aggregation Accuracy

The data aggregation accuracy is evaluated. The percentage of false data sent by sensor nodes is also changed in the simulation. The results are presented in Figure 3 where the data aggregation accuracy of the network is defined as [1-Error in the aggregation result]. The error in the aggregation result is the difference between the aggregated data computed by the data aggregator and the aggregated value of the data sent by the sensor nodes without any false data. Hence, the data aggregation accuracy is affected by outlier detection performance. If the network eliminates all the outliers in the network, then the data aggregator does not receive any false data resulting in 100% correct data aggregation results. As seen from Figure 3, the percentage of false data in the network negatively affects the data aggregation accuracy .

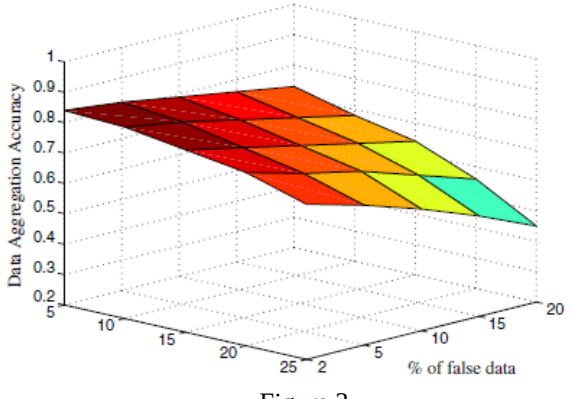


Figure 3

## 8. Conclusion

This paper presents a method that detects outlier data, and data aggregation is performed eliminating the false data. The simulation results show that the proposed scheme, is able to detect outliers in most cases. As a result the number of false data transmissions is reduced thereby increasing the data aggregation accuracy.

## 9. Reference

[1] Intanagonwiwat C, Estrin D, Govindan R, Heidemann J. Impact of network density on data aggregation in wireless sensor networks. Proc. of the 22nd International Conference on Distributed Computing Systems 2002; 575–578.
[2] Ozdemir S, Xiao Y. Secure data aggregation in wireless sensor networks: a comprehensive overview. Computer Networks 2009; 53(12):2022–2037.
[3]. Perrig A, Szewczyk R, Tygar D, Wen V, Culler D. SPINS: security protocols for sensor networks. Wireless Networks Journal (WINE) 2002; 8(5):521–534.
[4]. Akyildiz IF, Su W, Sankarasubramaniam Y, Cayirci E. A survey on sensor networks. IEEE Communications Magazine 2002; 40(8):102–114.
[5] Hodge VJ, Austin J. A survey of outlier detection methodologies. Artificial Intelligence Review 2004; 22(2):85–126.
[6] Han J, Kamber M. Data Mining: Concepts and Techniques. Morgan Kaufmann: San Francisco, 2006.
[7] Hodge VJ, Austin J. A survey of outlier detection methodologies. Artificial Intelligence Review 2004; 22(2):85–126.
[8] Heinz C, Seeger B. Statistical modeling of sensor data and its application to outlier detection. Technical Report 2006/07, University of Stuttgart, 2006.
[9] Sheng B, Li Q, Mao W, Jin W. Outlier detection in sensor networks. Proc. of MobiHoc, 2007; 219–227.
[10]. Jeffery S, Alonso G, Franklin MJ, Hong W, Widom J. Declarative support for sensor data cleaning. Proc. Of Pervasive Computing, 2006.
[11]. Wen Yj, Agogino AM, Goebel K. Fuzzy validation and fusion for wireless sensor networks. Proc. of ASME, 2004.
[12]. Zhuang Y, Chen L, Wang S, Lian J. A weighted moving average-based approach for cleaning sensor data. Proc. of ICDCS, 2007.
[13]. Chen J, Kher S, Somani A. Distributed fault detection of wireless sensor networks. Proc. of DIWANS, 2006.
[14]. Xiao X, Peng W,Hung C, LeeW.Using sensor ranks for in-network detection of faulty readings in wireless sensor networks. Proc. of MobiDE, 2007.
[15]. Deligiannakis A, Kotidis Y, Vassalos V, Stoumpos V, Delis A. Another outlier bites the dust: computing meaningful aggregates in sensor networks. Proc. Of ICDE, 2009.
[16]. Syed Mohamed M, Kavitha T. Real time outlier detection in wireless sensor networks. International Journal of Latest Trends in Computing 2011; 2(1):114–118.
[17]. Zhang Y, Hamm NAS, Meratnia N, Stein A, Voort M, Havinga PJM. Statistics-based outlier detection for wireless sensor networks. International

Journal of Geographical Information Science 2012. doi:10.1080/ 13658816.2012.654493.
[18]. Chen Y, Juang J. Outlier-detection-based indoor localization system for wireless sensor networks. International Journal of Navigation and Observation 2012. doi:10.1155/2012/961785.
[19]. P.N. Tan, M. Steinbach, V. Kumar Introduction to Data Mining Pearson Addison Wesley, USA (2006) (ISBN 9780321321367)
[20]. M. Ankerst, M.M. Breunig, H.-P. Kriegel, J. Sander OPTICS: ordering points to identify the clustering structure ACM SIGMOD Rec., 28 (2) (1999), pp. 49–60 http://dx.doi.org/10.1145/304182.304187
[21] M. Ester, H.-P. Kriegel, J. Sander, X. Xu A density-based algorithm for discovering clusters in large spatial databases with noise 2nd International Conference on Knowledge Discovery and Data Mining (1996), pp. 226–231 (http://citeseer.ist.psu.edu/viewdoc/summary?doi = 10.1.1.121.9220)