Distributed Anomaly Detection in Wireless Sensor Networks

*Ammara Razzaq U00045665 Computer Science Department; University of Sharjah*

# Abstract

Wireless Sensor Networks (WSNs) are a network of sensor nodes which communicate wirelessly with each other to transmit data and perform computations on that data. Sensor nodes consist of very tiny sensors which are used for several purposes like monitoring temperature, humidity and concentration of gases for several purposes like health care, security and protection. Sensor data is not always accurate as several errors can be introduced in it due to faulty sensor devices, strange events and malicious attacks. These misbehaviors in sensor data are termed as anomalies. Anomaly detection in sensor network is very important to detect any faults in the sensor data with minimum energy consumption. In this paper a distributed anomaly detection approach is implemented using fixed width clustering technique. It is demonstrated that distributed scheme achieves comparable performance to centralized scheme with reduced communication overhead.

# Introduction

Wireless sensor networks (WSNs) consist of inexpensive, low power, tiny sensor nodes that can perform multiple functions like monitoring temperature, pressure and humidity. The large numbers of densely deployed sensor nodes communicate with each other through radio signals to form a network of sensors.

Wireless sensor nodes are low in power, bandwidth, memory and computational capabilities. They have limited battery power due to limited size. Maximum energy of a sensor node is expended in data communication between nodes therefore they have small lifespan. Sensor network protocols must be designed with low computational costs in order to consume less power to prolong the lifetime of the network. [1]

The transmitted data between nodes can suffer with anomalous measurements due to multiple reasons. The inaccuracy in data can be due to sensor environment or resource limitations of sensor devices like limited battery power and computational capacities. The performance of sensor nodes is affected due to errors in sensor data.

The observations in the sensor data that are incoherent with the rest of the data are termed as anomalies. They are defined in [2] as “patterns in data that do not confirm a well-defined notion of normal behavior”. The procedure to find these anomalies in the data is called anomaly detection [2]. The use of anomaly detection in sensor networks helps in removing the anomalous data hence making the sensor data accurate and precise. The goal of the anomaly detection techniques is to provide high efficiency in anomaly detection with minimum power consumption. Therefore anomaly detection techniques should have low computational complexity as high computational complexity consumes more power.

For the purpose of anomaly detection different types of model structures have been implemented. Two main types of model structures are centralized and distributed. In centralized approach, the data from every node is transmitted to gateway node where all the data computation is performed. In distributed approach, only a summary of the required data characterized by its local normal reference model is transmitted to gateway node. The gateway node then combines the data from all nodes and computes a global normal reference model and then transmits it back to all nodes. This global normal reference model is then used to detect anomalies at every node. Hence distributed approach of anomaly detection is more energy efficient as compared to centralized approach as it provides less communication overhead [3]. In sensor nodes more energy is consumed in transmitting the data and less energy is consumed in data computation work. Centralized approach is not an appropriate approach, as large amount of raw data is transmitted to the gateway node causing communication overhead and high energy consumption. In distributed approach only a summary of data is transmitted, hence this approach is more energy efficient. However size of the local normal reference model should be small to avoid any communication overhead.

In this project both distributed and centralized approach have been implemented and their accuracy is measured using clustering based anomaly detection technique presented in Rajasegarar et al. [4].

## Clustering based anomaly detection:

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). It is a main task of exploratory data mining [5].

Rajasegarar et al. [4]: in this approach, proposed distributed anomaly detection technique in wireless sensor networks using clustering technique.A comparison between centralized and distributed approach is done in this paper using fixed width clustering technique.

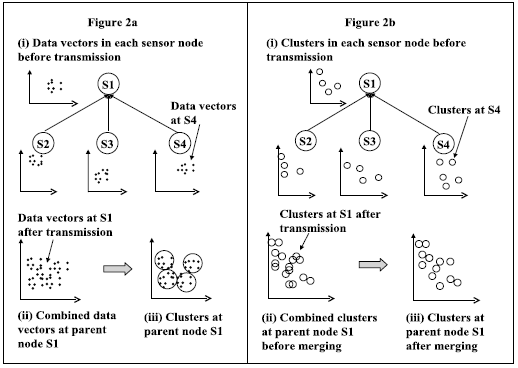
In centralized approach, data vectors at individual nodes are transmitted to the gateway node. At gateway node, all data gathered from children nodes and its own data are combined. Then this data is clustered. After clustering, similar clusters are merged together using cluster merging algorithm. Then the anomaly detection algorithm is applied to clustered data at gateway node to detect anomalous clusters.

In distributed approach data clustering is performed at every individual node.

The main steps of the algorithm are as follows:

1. At every node, the node data is clustered using clustering technique.
2. The clustered data at every node is transmitted to its immediate parent node.
3. Parent node receives clustered data from every node and combines it.
4. Parent node merges the combined set of clusters.
5. Parent node sends the data to its immediate parent until it reaches the gateway node.
6. Anomaly detection algorithm is applied at gate way node to find anomalous clusters from the merged clusters.

Figure 1 shows the difference between centralized and distributed approach.



(b)

(a)

**Figure1.** (**a**) Centralized detection model (**b**) Distributed Model

# Implementation and Experiments

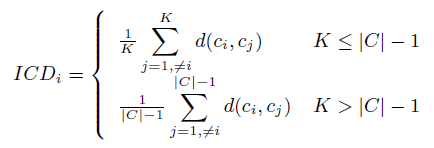
The algorithms are implemented in C++ using test sensor data. Both centralized and distributed algorithms are evaluated on a one level hierarchy topology to determine their accuracy.

Three main algorithms that are implemented are fixed width clustering algorithm, merging of clusters and anomaly detection algorithm. In the anomaly detection algorithm value of K for K nearest neighbors (KNN) is set to 5. In the paper, data conditioning is applied on the sensor data before clustering it. For this project, data conditioning algorithm is implemented but is not applied to sensor data, because sensor data falls in the range [0, 1] after conditioning, and they will almost fall into single cluster for small set of test data.

The clustering algorithm performs fixed width clustering based on the Euclidean distance between data vectors. The width of cluster is specified by user. First data vector in the data set is selected as the centroid of the first cluster. For each data vector, the Euclidean distance between that data vector and the centroid of every cluster is measured. If this distance is less than the width specified by the user then the data vector is added to that cluster and the centroid of that cluster is adjusted to the mean of data vectors that it contains. If the distance is greater than width then a new cluster is created with data vector specified as the centroid of the new cluster. This clustering algorithm is very efficient as only one pass through the data is required to perform clustering.

In merging of clusters algorithm, clusters are merged together based on some similarity criteria. In this algorithm Euclidean distance between the centroids of the clusters is used for detecting similar clusters. If the Euclidean distance between the centroids of two clusters is less than the width then these two clusters are similar and will be merged into one cluster. This new cluster will contain the data points of both clusters and its centroid will be the mean of data points of both clusters.

In anomaly detection algorithm average inter-cluster distance between clusters is used to detect anomalous clusters. First, for each cluster ‘ci’ inter-cluster distance d(ci,cj) with every other cluster is computed. Dci = { d(ci,cj), j = 1…..(|C|-1) ,j!=i}. Dci is the set of inter-cluster distances for every cluster. d(ci,cj) is the Euclidean distance between centroids of clusters ci and cj. |C| is the size of cluster. Second, average inter-cluster distance ICDi for each cluster is computed for K smallest values from Dci using following equation.



Third, a cluster is identified as anomalous if it satisfies following equation.

C:\Users\ammara\Desktop\ddd.png

A cluster ‘ci’ belonging to the cluster set C is anomalous, if its average inter-cluster distance is greater than the sum of mean of inter-cluster distance and standard deviation of inter-cluster distance.

Both centralized approach and distributed approach have comparable accuracy. But in centralized approach more energy is consumed since complete data of all sensors is transmitted to the gateway node. In comparison, distributed approach is more energy efficient since data is clustered at every node and instead of sending the complete raw data, processed data is transmitted which consumes less energy.

To measure the accuracy of the algorithm, false negative ratio of the result of anomaly detection algorithm is measured. A false negative occurs when an anomalous measurement is measured as normal. False negative ratio is the ratio of false negatives to actual anomalous data. False negative ratio is measured for varying cluster widths. It becomes higher for very large and very small cluster widths. At large values of cluster width, small set of clusters are produced. And both anomalous and normal values can be included into single cluster and hence anomaly detection algorithm will not be able to detect these clusters. For smaller values of cluster widths, large numbers of clusters are produced and there is a possibility that every cluster is a singleton. Anomalous data will be distributed in smaller clusters and they will have small inter-cluster distances and anomaly detection algorithm will not detect these clusters. For a range of width between these two extremes anomaly detection algorithm detects anomalous clusters accurately.

A graph is plotted to show the variation of FNR with varying values of cluster widths for distributed and centralized case. They show same accuracy for given range of cluster widths.

## Conclusion and Future Work

In this project, an anomaly detection algorithm is implemented using fixed width clustering technique. Two approaches for anomaly detection i.e. distributed and centralized are implemented and compared with each other. This algorithm is evaluated on test data for one level hierarchy topology. The evaluation shows that distributed approach has comparable accuracy to centralized approach but with less communication overhead.

In future, accuracy of algorithm can be studied for varying K values. Future research can be expanded to evaluate the anomaly detection algorithm for multiple KNN parameters.

## References

1. Akyildiz, I.F.; Su, W.; Sankarasubramaniam, Y.; Cayirci, E. Wireless sensor networks: A survey. *Comput. Netw.* **2002**, *38*, 393–422.
2. Chandola, V.; Banerjee, A.; Kumar, V. Anomaly detection: A survey. *ACM Comput. Surv.* **2009**, *41*, 15.
3. Murad A. Rassam; Anazida Zainal; Mohd Aizaini Maarof. Advancement of Data Anomaly Detection Research in Wireless Sensor Networks: A Survey and Open Issues 2013
4. Rajasegarar, S.; Leckie, C.; Bezdek, J.C.; Palaniswami, M. Distributed Anomaly Detection in Wireless Sensor Networks. In Proceedings of the 10th IEEE Singapore International Conference on Communication Systems (ICCS 2006), Singapore, 30 October–1 November 2006; pp. 1–5.
5. http://en.wikipedia.org/wiki/Cluster\_analysis