Hybrid Approach for Outlier Detection over Wireless Sensor Network Real Time Data

P. R. Chandore[1], Dr. P. N. Chatur[2]

[1]*M.Tech, Department of Computer Science and Engineering,*

[2]*Head of Department Computer Science and Engineering*

[12]*Government College of Engineering, Amravati, Maharashtra, India.*

[1]prakash.chandore@gmail.com, [2]chatur.prashant@gcoea.ac.in

***Abstract* -- Outlier detection has gained considerable interest over real time data stream in data mining community with the realization that outliers can be the key discovery to be made from very large databases or data stream. Wireless sensor networks are one of active research area where massive amount of data measured or recorded. Those measurements that significantly deviate from the normal pattern of sensed data are considered as outlier's .The potential reasons of outliers include noise and errors, events, damage of device, and malicious attacks on the network. Sensor data is real time data recorded continuously with specific requirement and limitations of wireless sensor network in such condition traditional outlier detection techniques are not directly applicable. This paper provides an extensive review of existing outlier detection techniques specifically deployed for the wireless sensor networks and proposed a hybrid approach for detecting outlier in wireless sensor network. We proposed the technique to detect outlier over sensor data using a cluster based approach and distance based approach which contribute a compressive work for finding the outliers in real time data within WNS's.**

***Keywords* -- Outliers, Data Stream, Wireless Sensor Network (WSN), Cluster based approach, distance based approach.**

I. INTRODUCTION

A wireless sensor network (WSN) constitutes a large number of sensor nodes, distributed over a large area with a small number of powerful sink nodes and can be used for a multitude of applications for short-range wireless communication within WSN. Sink nodes gathers readings of sensor nodes which are autonomously integrated with sensing, processing and wireless communication capacities. Where each node is commonly equipped with a radio transceiver, a small microcontroller, power source and multi-type sensors such as temperature, humidity, light, heat, pressure, sound, vibration, etc. [1]. Usually sensor networks are deployed in a multi-hop topology. A typical example of a WSN is shown in fig. 1. The Networks are usually composed of few sinks and large quantity of sensor nodes.

The Wireless Sensor Network is not only used to provide real time data about the physical enviourment but also to detect time critical events. A wide variety of applications of WSNs will be deployed in buildings, cars, and the environment, for monitoring health, traffic, machine status, weather, pollution, surveillance, health and medical monitoring, battlefield observation and so on. In many of these applications, real-time data mining of sensor data to promptly make intelligent decisions is essential [2]. It is likely that they will be in use for a long time, generating a large amount of data. Mining this large data repository for useful information will be crucial.

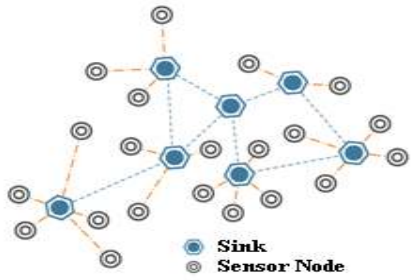


Fig. 1 Wireless Sensor Network

Usually a sensor node follows a hope to hope communication where it forwards data to the next hop node but in certain cases data aggregation is also required at the nodes in order to reduce the use of network resources i.e. used bandwidth and power conservation, consumption and media access delay. Thus two main functions of a sensor node can be defined: data dissemination or data gathering. One is event driven and second is demand driven. In event driven WSN's networks, communication is started by one of the nodes in the network while in the demand driven type the process is initiated by the central gateway or monitoring station [3]. Examples of the event driven systems include monitoring of fire in the forest whereas an inventory control system is demand driven.

WSN's are generating massive amount of data continuously which measured and collected by sink is often unreliable for processing. The quality of data set may be affected by noise and error, missing values, duplicated data, or inconsistent data. The low cost and low quality sensor nodes have rigorous resource constraints such as energy (battery power), memory, computational capacity and complexity, and measured communication bandwidth. The limited resource and capability make the data generated by sensor nodes unreliable

and inaccurate. Especially erroneous data increases when battery power is exhausted [4]. In WSN's operations of sensor nodes are frequently sensitive to environmental effects such as harsh environments where it is inevitable that in such environments some sensor nodes malfunction, which may result in noisy, faulty, missing and anomalous data. Sensor nodes are vulnerable to malicious attacks such as denial of service attacks, black hole attacks and eavesdropping [6], in which data generation and processing will be manipulated by adversaries. The above mentioned factors lead further influence quality of raw data and aggregated results. Since with respective WSN's actual event occurring in physical world such as forest fire, earthquake or chemical spill, cannot be accurately detected using inaccurate and incomplete data [2].

II. CARDINAL OF OUTLIER DETECTION IN WIRELESS SENSOR NETWORKS

This section provides fundamentals of outlier detection in WSN's.

A. Outlier Detection

Outlier detection over large database is currently active research area of data mining which is discovery of data that deviate a lot from other data patterns .Detecting outliers refers to the problem of finding patterns in data that are very different from the rest of the data based on appropriate metrics. Such a pattern often contains useful information regarding abnormal behavior of the system described by the data. Earlier outlier detection work carried in the field of statistics [7]. D .Hawkins [8], gives definition to outlier as: "*an outlier is an observation as to arouse suspicion that it was generated by a different mechanism*". Barnett [9] proposed another classical definition for outlier that "*an outlier is an observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data"*.

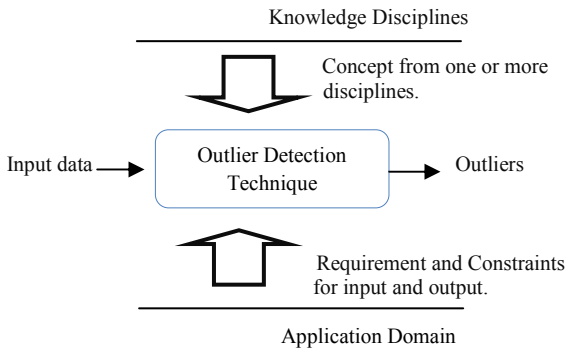


Fig. 2 A general design of an outlier detection technique.

As illustrated in Figure 2, any outlier detection technique has following major ingredients.

1. Nature of data, behavior of outliers, and other restraints and assumptions that collectively constitute the problem formulation.

2. Application domain in which the technique is applied. Some of the techniques are developed in a more generic fashion but are still feasible in one or more domains while others directly target a particular application domain.

3. The concept and ideas used from one or more knowledge disciplines.

Outliers can be defined in WSN's as *"those measurements that significantly deviate from the normal pattern of sensed data"* [10]. This definition is provided with respective WSN's sensor nodes terminology of measured data. Potential sources of outliers in data collected by WSNs include *noise* and *errors*, *actual events*, and *malicious attacks*. Without changing real significance of data that dramatically affect data analysis such noisy data as well as erroneous data should be eliminated or corrected [11].

B. Outlier Detection in Wireless Sensor Network

Outlier detection is a primary step in many data-mining applications. It refers to the problem of finding patterns in data that do not conform to expected normal behavior or anomalous behavior. i.e., *mining useful and interesting information from a large amount of data* [12]. Currently outlier detection is active research area from data mining community; Outlier detection has been widely researched in various disciplines such as statistical analysis, data analysis, machine leaning, information theory, streaming data, and spatial data [10]. Recently, the topic of outlier detection in WSNs has attracted much attention. Due to potential sources of outliers as mentioned earlier, the identification of outliers provides data reliability, event reporting, and secures functioning of the WSN's. The detected values form outliers detection consequently are treated as events indicating change of phenomenon that are of interest. In addition outlier detection identifies malicious sensors that always generate outlier values, detects potential network attacks by adversaries, and further ensures the security of the network.

III. MOTIVATION

The Constraints of WSN's and the nature of sensor data make design of an appropriate outlier detection technique more challenging. Traditional outlier detection techniques might not be suitable for handing sensor data in WSNs.

TABLE I
PARAMETRIC CONTEXT OF WIRELESS SENSOR NETWORKS.

Parameter	Context
Resource constraints.	Dynamic nature of data causese traditional methods to have a high computation cost for evaluation of outlier. Data is passing with a time constraints in wirless sensor network only one pass of scan is possible so we required much memory for data analysis and storage. So it needed that Minimize the energy consumption while using a reasonable amount of memory for storage and computational tasks.
High communication cost	For a sensor node, the communication cost is often several orders of magnitude higher than the computation cost [13]. Traditional methods need much cost for

	communication within sink and node in WSN's
Distributed streaming data.	One of the major problem exisitng approach that handling dynmic change in data where it diffcult to identify prior distribution of data steram. Direct computation of probabilities is difficult [14].Do not meet the requirement of handling distributed stream data.
Dynamic network topology, frequent communication failures, mobility and heterogeneity of nodes.	Environment and physical loaction are one of the adversites in wireless sensor network. outlier detection in WSNs varies in accordance with different sensing device different processing capabilities . Due to heterogenity of data and dynamic in nature it is diffcult to understand prior distribution of processing data.
Identifying outlier sources.	Difficult to identify what has caused an outlier in sensor data due to the resource constraints and dynamic nature of WSNs.
Uncertain data or missing data	If prior distibutuion is not known then it is diffcult to formulate outlier detection model. Due to misssing valuses we may lead to wrong decision.
High Dimensional Data Stream	Due to high dimensional proerty we required a huge storage space and also processing complexity is high.

One of the biggest challenges for outlier detection over wireless sensor network is to keep energy consumption low. In addition with above scenario WSN's should maintain mining accuracy requirement and resource consumption should be maintained as minimum [14]. Another important issue regarding with processing of streaming data; that earlier work carried out over static data. Outlier detection over wireless sensor network should be able to evaluate outlier in online fashion while keeping the communication overhead, memory and computational cost low [1].

IV. RELATED WORK

In this section, we are focusing over earlier work carried out for outlier detection in wireless sensor networks. Due to numerous reasons outlier are generated in data. So it is prior that to identify such causes that will lead us towards important information to research. Following figure shown prior causes for outlier detection in WSN's.

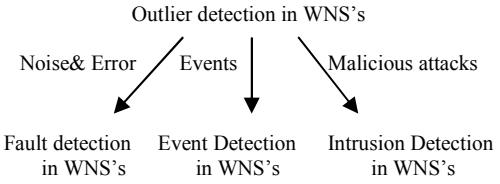


Fig. 3 Three outlier sources in WSNs and their corresponding detection techniques

A. Statistical-Based Approaches

Statistical-based approaches are the earliest approaches deal with the problem of outlier detection these are are essentially *model-based* techniques. A data point is declared as an outlier if the probability of the data point to be generated by this model is very low. The statistical-based approaches are categorized into parametric and non-parametric based on how the probability distribution model is built.

1) Parametric Approach: Parametric techniques assume Availability of the knowledge about underlying data distribution. These approaches further classified into a Gaussian-based models and non Gaussian based models.

Gaussian-based approach: Identification of outlying sensors as well as identification of event boundary in WNS's these two technique presented by Wu et al. [15]. Bettencourt et al. [16] present a local outlier detection technique to identify errors and detect events in ecological applications of WSNs. Hida et al. [17] design a local technique to make simple aggregation operations, such as MAX or AVG, more reliable under presence of faulty sensor readings and failed nodes. Jun et al. [18] present a non Gaussian based approach based on a statistical-based technique, which uses a symmetric α- stable (SαS) distribution to model outliers being in form of impulsive noise.

2) Non-Parametric-Based Approaches: Non-parametric techniques do not assume availability of data distribution. They typically define a distance measure between a new test instance and the statistical model and use some kind of thresholds on this distance to determine whether the observation is an outlier. Histogram based technique to identify global outliers in data collection applications of sensor networks were proposed by Sheng et al. [19]. Use of kernel density function for estimating outliers within WSN's proposed by Palpanas et al. [20].

C. Clustering-Based Approaches

These approaches attempt to detect both either single point outliers or cluster-based outliers, and can assign each outlier a degree of being an outlier. Data instances are identified as outliers if they do not belong to clusters or if their clusters are significantly smaller than other clusters. Clustering technique to identify anomalous measurements in sensor nodes was proposed by Rajasegarar et al. [21]. This technique minimizes the communication overhead by clustering the sensor measurements and merging clusters before communicating with other nodes.

B. Nearest Neighbor-Based Approaches

Commonly followed approach within data mining community to analyze data with respect to nearest neighbor based approach. A data instance is declared as an outlier if it is located far from its neighbors. Branch et al. [22] propose a technique based on distance similarity to identify global outliers in sensor networks.

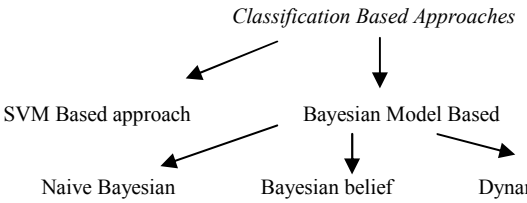
Zhang et al. [23] propose a distance-based technique to identify *n* global outliers in snapshot and continuous query processing applications of sensor networks. Zhuang et al. [24] present two in-network outlier cleaning techniques for data collection applications of sensor networks.

D. Machine Learning Based Approaches

Machine learning based approach is uses a measured estimation or illation of next phase of system on previous measured values. Such techniques are prediction based or learning based for deciding a underlying model of data, to identify new data element as inlier or outlier. These approaches are further classified as follows.

1) *Prediction Based Approaches:* This approach is also termed as model based approach because it uses a prediction model of data using filtering methods. To exploits outliers within spatial and temporal dependencies of sensor data; kalman filter based a prediction technique proposed in [30]. This technique uses only estimation of state of model from previous and current measurement of sensor data. This approach uses two phase for evaluation such as state transition phase which predicates the state of next time based on that of current time where as measuring module is used for measuring neighboring sensor readings as data produced by a virtual sensor device.

2) *Classification Based Approaches:* This approach is also model based approach where model get formulated using the underlying distribution of data which called as classification model. Classification model is built on training data instances to evaluate new data instances. This approach further classified as follows



Network Network Network

Fig 4 Categories of classification based approach

Rajasegarar et al. [25] propose a SVM-based technique for outlier detection in sensor data. This technique uses one-class quarter-sphere SVM to reduce the effort of computational complexity and locally identify outliers at each node.Elnahrawy and Nath [26] present a Bayesian model-based technique to discover local outliers and detect faulty sensors. Janakiram et al. [27] present a technique based on Bayesian belief network to identify local outliers in streaming sensor data. Hill et al. [28] present two techniques based on dynamic Bayesian networks to identity local outliers in environmental Sensor data streams. degree of probabilistic independencies among variables as Naive Bayesian network, Bayesian belief network, and Dynamic Bayesian network Principal Component Analysis (PCA) based approach is categorized under Spectral decomposition based approach. In accordance with this existing approach we have proposed a hybrid approach based on distance based and clustering based as shown in chronology of outlier detection as shown in outlier detection chronology. Distance based approach is categorized under statistical based approach hence proposed approach is from of hybrid approach over WSN.

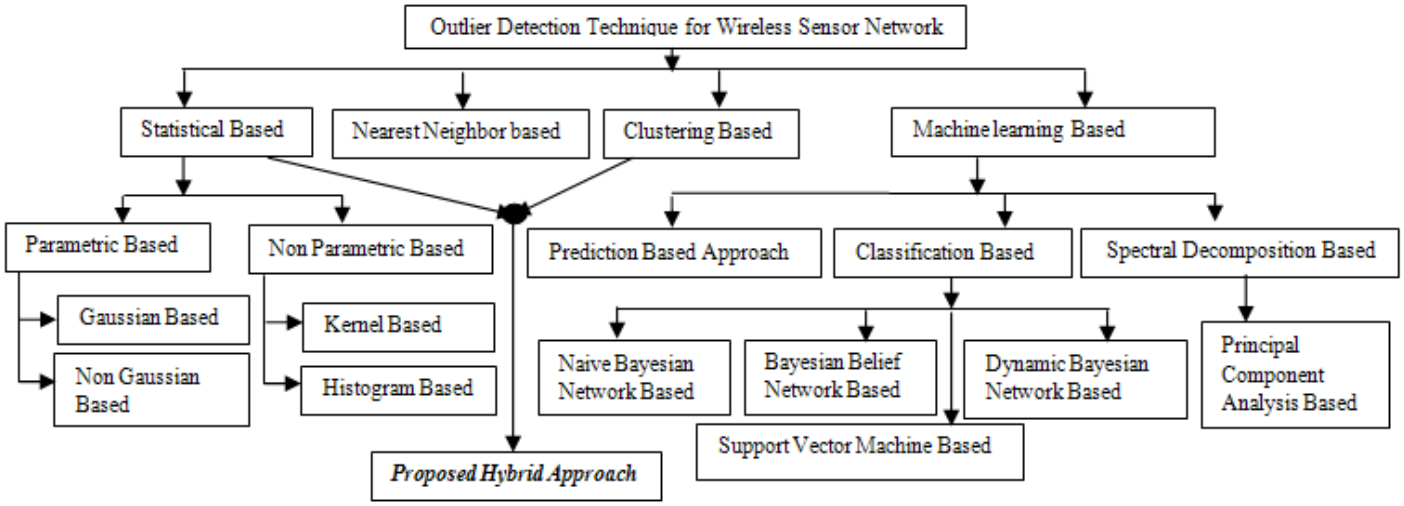


Fig. 5 Outlier Detection Chronology for WSN's and Proposed Hybrid Approach

3) Spectral Decomposition-Based Approaches

This method is used for dimension reduction of processing data. Principal Component Analysis is used for finding normal modes of behavior in the data and prior used to reduce dimensionality of measured data. It also used for finding a new subset of dimension by which we can specifically capture the behavior of the data. The top few principal components capture the build of variability and any data instance that violates this structure for the smallest components is considered as an outlier. Chatzigiannakis et al. [29] propose a PCA-based technique to solve data integrity and accuracy problem caused by compromised or malfunctioning sensor nodes.

V. OUTLIER DETECTION CRONOLOGY FOR WIRELESS SENSOR NETWORKS

Outlier detection techniques for WSNs can be categorized into *statistical-based*, *nearest neighbor-based*, *clustering-based* and *Machine learning based* approaches as shown in fig 5. Statistical-based approaches are further categorized into *parametric based and nonparametric based* approaches where parametric approach categories Gaussian-*based approaches and non-Gaussian based* approaches. In statistical based approach kernel-*based approach and histogram-based* approaches are

belong to non parametric approaches. Machine learning based approach is categorized as Prediction based approach, Classification-based approaches, and spectral decomposition method. Where classification based approach is also termed as a *Bayesian Network based* approaches which is further classified using a specific property of measured sensor real time data

VI. PROPOSED APPROACH FOR OUTLIER DETECTION OVER WIRELESS SENSOR NETWORKS

Proposed work is hybrid approach of existing approaches of distance based approach and cluster based approach as shown in diagram. Work flow of proposed hybrid approach as per mentioned in following points. Partition the data stream into number of chunks and each chunk contain set of data.

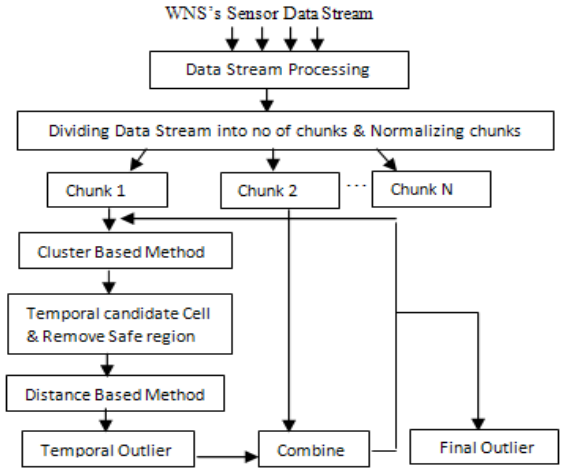


Fig. 6 Proposed Hybrid Approach implementation Architecture

Now over each of these normalized chunks we apply clustering method to find out temporal outlier region and declaring safe region. This safe region will be discarded then we apply distance based outlier detection algorithm over temporal outliers. Temporal outlier found in last chunk again added to respective next chunk of data stream to survive in next stream, and allow it for appropriate number of stream chunks, then declare candidate outliers as real outliers or inliers.

Cluster Based Approach: Clustering is a popular technique used to group similar data points or objects in groups or clusters. Clustering is an important tool for outlier analysis. Cluster based approach is here act as data reduction. First, clustering technique is used to groups the data having similar characteristics. And calculate the centroids for each group.

Distance Based Approach: Distance based technique is used to calculate maximum distance value for each cluster. If this maximum distance is greater than some threshold then it will declare as outlier otherwise as a real object or inliers. Threshold is given by user.

VII. DISCUSSION

Wireless sensor networks generate huge amount of massive data which contain useful information. So it is worthwhile to find out outlier within such data. Outlier detection as a branch of data mining has many important applications and deserves more attention from data mining community. Comparison between Distance based approach and proposed approach are as follows. *Distance-Based Method:* This Operate on whole data. Specifically cannot give number of clusters. Overall calculation of data points within data stream increases that increase computation cost but finally it gives only one value as most expected outlier. *Clustering and Distance-Based:* in comparison with above method proposed approach can group the data in to number of clusters whereas it reduce the size of database that will reduces computation time to each cluster where user can give certain radius to find outliers.

VIII. CONCLUSSION

Outlier detection of is essential for measuring quality of data for predicting information over data analysis. We observed that maintaining a quality and control over data analysis is prior fundamental task of outlier detection. From Above extensive review we reviewed that outlier detection method should be able to work with high dimensional data, online manner over multivariate streaming data, communication and computation overhead should low. We observed that individual method are not able to handle above scenarios hence we proposed a hybrid approach in which it first groups the data having similar characteristics in to number of clusters. Due to reduction in size of data stream, the computation time reduced considerably. Using a distance based approach over cluster it will remarkably reduce a huge amount extra calculation cos. Hybrid approach will take less computation time.

REFERENCES

[1] C. F. Garca-Hernndez, P. H. Ibargengoytia-Gonzlez, J. Garca Hernandez, And J. A. Prez-Daz, "Wireless sensor networks and applications: A survey," IJCSNS International Journal of Computer Science and Network Security, VOL.7No.3,pp.264-273,2007.

[2] X. Ma, D. Yang, S. Tang, Q. Luo, D. Zhang, and S. Li, Online Mining in Sensor Networks, *IFIP international conference on Network and parallel computing*, Vol. 3222, pp. 544-550 2004.

[3] I. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, A survey on sensor networks," Communications Magazine IEEE, vol. 40, pp. 102 -114, aug 2002.

[4] S.Subramaniam, T. Palpanas, D. Papadopoulos, V. Kalogerakiand, and D. Gunopulos, Online Outlier Detection in Sensor Data using Nonparametric Models, *J. Very Large Data Bases*, VLDB 2006.

[5] F. Martincic and L. Schwiebert, Distributed Event Detection in Sensor Networks, *Proc. International Conference on System and Networks Communication*, pp. 43-48, 2006.

[6] A. Perrig, J. Stankovic, and D. Wagner, Security in Wireless Sensor Networks, *CACM*, Vol. 47, No. 6, pp. 53-57, 2004.

[7] V. Hodge and J. Austin, A Survey of Outlier Detection Methodologies, *Artificial Intelligence Review*, Vol. 22, pp. 85-126, 2003.

[8] D.M. Hawkins, *Identification of Outliers*, London: Chapman and Hall, 1980.

[9] V. Barnett and T. Lewis, *Outliers in Statistical Data*, New York: JohnWiley Sons, 1994.

[10] V. Chandola, A. Banerjee, and V. Kumar, Anomaly Detection: A survey Survey,*Technical Report*, University of Minnesota, 2007.
[11] P.N. Tan, M. Steinback, and V. Kumar, *Introduction to Data Mining*, Addison Wesley, 2006.
[12] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, MorganKaufmann, San Francisco, 2006.
[13] I.F. Akyildiz, W. Su, Y. Sankara subramaniam, and E. Cayirci, Wireless Sensor Networks: A Survey, *J. Computer Networks*, Vol. 38, No. 4, Pp.393-422, March, 2002.
[14] M. M. Gaber, Data Stream Processing in Sensor Networks. In J. Gamaand M. M. Gaber, *Learning from Data Streams Processing Techniquesin sensor Network*, pp. 41-48. Springer Berlin Heidelberg, 2007.
[15] W. Wu, X. Cheng, M. Ding, K. Xing, F. Liu, and P. Deng, Localized Outlying and Boundary Data Detection in Sensor Networks, *IEEE Trans.Knowl. Data Eng.*, Vol. 19, No. 8, pp. 1145-1157, 2007.
[16] L.A. Bettencourt, A. Hagberg, and L. Larkey, Separating the Wheatfrom the Chaff: Practical Anomaly Detection Schemes in EcologicalApplications of Distributed Sensor Networks, *Proc. IEEE InternationalConference on Distributed Computing in Sensor Systems*, 2007..
[17] Y. Hida, P. Huang, and R. Nishtala, Aggregation Query under Uncertainity In Sensor Networks, 2003.
[18] M.C. Jun, H. Jeong, and C.C.J. Kuo, Distributed Spatio-TemporalOutlier Detection in Sensor Networks, *Proc. SPIE*, 2006.
[19] B. Sheng, Q. Li, W. Mao, and W. Jin, Outlier Detection in Sensor Networks *Proc. MobiHoc*, 2007.
[20] T. Palpanas, D. Papadopoulos, V. Kalogeraki, and D. Gunopulos, Distributed Deviation Detection in Sensor Networks, ACM *Special Interest Group on Management of Data*, pp. 77-82, 2003.
[21] S. Rajasegarar, C. Leckie, M. Palaniswami, and J.C. Bezdek, Distributed Anomaly Detection in Wireless Sensor Networks, *Proc. IEEE ICCS, 2006.*
[22] J. Branch, B. Szymanski, C. Giannella, and R. Wolff, In-Network Outlier Detection in Wireless Sensor Networks, *Proc. IEEE ICDCS*, 2006.
[23] K. Zhang, S. Shi, H. Gao, and J. Li, Unsupervised Outlier Detection in Sensor Networks using Aggregation Tree, *Proc. ADMA*, 2007.
[24] Y. Zhuang and L. Chen, In-Network Outlier Cleaning for Data Collection in Sensor Networks, *Proc. VLDB*, 2006.
[25] S. Rajasegarar, C. Leckie, M. Palaniswami, and J. C. Bezdek, Quarter Sphere Based Distributed Anomaly Detection in Wireless Sensor Net S. Rajasegarar, C. Leckie, M. Palaniswami, and J. C. Bezdek, Quarter Sphere Based Distributed Anomaly Detection in Wireless Sensor Networks, *Proc. IEEE International Conference on Communications*, pp.3864-3869, 2007.
[26] E. Elnahrawy and B. Nath, Context-Aware Sensors, *Proc. EWSN*, 2004.
[27] D. Janakiram, A. Mallikarjuna, V. Reddy, and P. Kumar, Outlier Detection in Wireless Sensor Networks using Bayesian Belief Networks, *Proc. IEEE Comsware*, 2006.
[28] D.J. Hill, B.S. Minsker, and E. Amir, Real-Time Bayesian Anomaly Detection for Environmental Sensor Data, *Proc. 32nd Congress of the International Association of Hydraulic Engineering and Research*, 2007.
[29] V. Chatzigiannakis, S. Papavassiliou, M. Grammatikou, and B. Maglariset, Hierarchical Anomaly Detection in Distributed Large-Scale Sensor Network *Proc. ISCC, 2006*
[30] M. Shuai, K. Xie, G. Chen, X. Ma, and G. Song, \A kalman _lter based approach for outlier detection in sensor networks," in Computer Science and Software Engineering, 2008 International Conference on, vol. 4, pp. 154 -157, dec. 2008.