IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 14, Issue 5 (Sep. - Oct. 2013), PP 99-103
www.iosrjournals.org

An Efficient Approach for Outlier Detection in Wireless Sensor Network

Prof. Suneet Shukla[1] , Jyotika Saxena[2]
[1]*Computer science & Engineering Deptt. ,CET,IFTM, Moradabad ,India)*
[2]*Computer Science & Engineering Deptt. ,CET,IFTM , Moradabad ,India)*

Abstract: *Wireless Sensor Networks are those networks which include many sensors, sensors have many sensor nodes that are spread all over the world. A wireless sensor network (WSN) normally has many sensor nodes which are very modest, having minimum cost dispersed over the world having high-powered sink nodes which collect the readings of the sensor nodes. These nodes are comprised with high sensing power, processing and wireless communication abilities. A sensor network involves large number of sensors, collecting and communicating the sensor measurements of observing physical world scenarios. The main problem in WSN is outlier. Basically outlier is an element of a data set that falls in an abnormal range. For the capabilities of WSNs, it is clear that we need an accurate and robust technique which can be used for multivariate data too and can give the optimum results in output threshold, and that technique should help to increase the special characteristics of WSNs such as node mobility, network topology change and making distinction between errors and events. Among all the techniques KERNEL FUNCTIONS is the best technique because it can be used for multivariate data also. Except it FTDA and LSH are also produce optimum results.*

Keywords: *fault-tolerant data aggregation scheme, Kernel Density , LSH, Outlier Detection , Wireless Sensor Network*

I. Introduction:

A wireless sensor network (WSN) usually comprises of a huge number of little, short cost sensor nodes spread over a huge region with one or probably more great sink nodes association readings of sensor nodes. The sensor nodes are mixed with logic, methods and wireless communication capacities. Each node is usually prepared with a wireless radio transceiver, power source, small microcontroller and multi-type sensors such as temperature, humidity, light, heat, pressure, vibration sound, etc. The sensor nodes are included with sensing, dispensation and wireless communication abilities. Each node is frequently able to with a wireless radio transceiver, a small microcontroller, a power source and multi-type sensors such as temperature, humidity, light, heat, pressure, sound, vibration, etc. Wireless Sensor Networks mainly consists of **sensors. Sensors** are - "A **sensor** (also called **detector**) is a device that processes a physical quantity and converts it into a signal which can be read by spectator or by an instrument." Wireless networks can also be installed in **extreme environmental** conditions and may be disposed to enemy attacks.

An outlier is an element of a data set that falls in a typical range. It means a number that is much LARGER or much SMALLER than the rest of the numbers.

Types of outlier:

- Local outliers are outliers that can be seen only in the local contiguity of a sensor node by gathering data from nearby particles.
- Global outliers are outliers that can be seen from all of the data collected by every node**.**

1.1 Problem In Wsn: The main problem in WSN is outlier. Basically outlier is an element of a data set that falls in an unusual range.

Recently, the topic of outlier detection in WSNs has attracted much attention. The identification of outliers provides data dependability, event writing, and secure working of the network. Specially, outlier detection controls the quality of calculated data, recovers toughness of the data analysis under the presence of noise and faulty sensors so that the communication overhead of flawed data is reduced and the gathered results are prevented to be affected.

II. Background Work

John[1] in [1995] says that an outlier might also be "surprising veridical data". He states that a point which belongs to class say A might actually be situating inside another class say B resulting in, true (veridical) classification of the point, a surprising one to the observer. According to Aggarwal and Yu [2001] [2] an outlier can be thought of as noise points lying outside a set of predefined clusters or the outliers might be considered as the points that lie outside of the set of clusters and might also be separated from the noise itself. These outliers behave differently from the norm. Hence outlier detection is the crucial task in an application processing, since outliers points to the abnormal behavior of various factors while processing of an application. Moreover, these outliers would lead to the significant performance degradation of selected application. A lot of work has been done related to outlier detection. In 2003, Eiman Elnahrawy and Badri Nath[3] , gave a structure for cleaning and querying noisy sensors. Specifically, they presented a Probability approach also known as Bayesian approach for reducing the uncertainty related to data that arise due to random noise. He presented this in an online fashion. This technique combined the prior knowledge of the true sensor reading, the characteristics of the noise of this sensor, and the observed noisy readings. This pre-processing step is very important and can be performed either at the sensor level or at the base-station. The authors have introduced several algorithms based on these proposed uncertainty models and using a statistical approach.

In 2004, Victoria J. Hodge & Jim Austin, [4] presented a survey of techniques for outlier detection. He gave many techniques for outlier detection. A user should select an algorithm that is suitable for their data set. It should be suitable in terms of the correct distribution model, the correct attribute types, the scalability, the speed, any incremental capabilities to allow new exemplars to be stored and accuracy. The user should also reflect on which of the approaches is suitable for their problem. There are three basic approaches a clustering approach, a classification approach or a novelty approach. The selection of suitable approach is depending on: the data type, whether the data is pre-labeled, how the authors want to detect outliers and how the authors have wished to handle them. Method of handling outliers is very important for a user. There may be many methods as some users have wished to expunge them from future processing in a diagnostic clustering and some users want a recognition system or retain them with an appropriate label in a accommodating clustering or they may want a classification system.

In 2006, Christoph Heinz and Bernhard Seeger [5] suggested that the first step to handle an important problem in sensor processing. This step is known as detection of outliers, with a statistical model and investigated the augmentation of sensor network querying by meaningful statistical models. This method is used instead of exploring the raw sensor readings. A statistical model is more reliable to gain insight into the physical phenomena observed. A key ingredient of statistical models is the probability density function (PDF) as it provides a comprehensive summary. Based on probability density function, the authors have presented an initial approach to detect outliers in streaming sensor data. In 2006, S. Subramaniam, UC Riverside, T. Palpanas and D. Papado-poulos, V. Kalogeraki, D. Gunopulos [6] proposed a framework that works in a distributed fashion. An approximation of multi-dimensional data distributions was given in order to enable complex applications in resource-constrained sensor networks. The authors gave these techniques in the context of the problem of outlier detection. The authors have discussed how the frameworks can be used to identify either distance or density-based outliers in a single pass over the data with limited memory requirements. The authors have studied the problem of outlier detection in sensor networks. Outlier detection is very important in this context, since it enables the user to focus on the interesting events in the network. In 2008,Yang Zhang, Nirvana Meratnia, Paul Havinga[7] , gave a frame-work of the techniques for the detection of outlier. The authors did a survey and also gave a decision tree for the selection appropriate technique for outlier detection. The authors gave many techniques and his techniques were based on pattern recognition and data mining.

III. Proposed Approach

3.1 Problem formulation

A sensor network constructs huge amount of data quickly in the form of a range of streams hence these values became doubtful, isolated and disloyal for our claim. And also in these streams there force be some relative among other streams or might be within the exacting stream itself. Processing of these data now becomes much more rough and boring since we have to take into reflection all these description of sensors and all these operating situation in our work. So we have to build up a capable algorithm or technique for this scenario. For this our method should have to be based on allocation of data on a data freedom at an exacting time. It means we have to guesstimate the data allocation for sensor analysis and with this on our hand we can know the compactness of the data space roughly each value, which promote simplicity our crisis of detecting outliers.

In this thesis, we propose a fault-tolerant data aggregation scheme using an in-network outlier detection mechanism, called FTDA and Kernel Density Estimation.

Kernel Density Estimation In Kernel Density Estimation the whole procedure go just around the Kernel Functions and it is the sum over Kernel Functions which are centered at sample points of disturbed data stream. Generally a kernel function depicts the way of distributing the weights in the area near to the values or points which is being processed in kernel density estimation. To find out the density estimation for whole data set we must have to combine all the kernel functions.

Let us assume that (X_1... X_d) are the elements of a block in a data stream then the kernel density estimator (KDE) may be defined as in [8]

$$f(x)=\frac{1}{d\times B}\sum_{i=1}^{d} K\left(\frac{x-X_i}{B}\right)$$

Here K denotes Kernel Function and B is the bandwidth or window width. Bandwidth is also called smoothing parameter by some authors. The bandwidth B is very essential and crucial parameter in our work of density estimation since its more or less controls the influential region of a point in a data stream as discussed in [5]. In the Paper [5] various methods are also discussed for making the bandwidth choice.

The bandwidth B is set by using Scott's rule [9] where the standard deviation σ of the values is determine within the random sample S taken from sensor data stream under processing.

$$B=\sqrt{5}\sigma|S|^{-\frac{1}{d+4}}$$

And the Kernel Functions have some basic properties like smoothness, continuous, positive and symmetric. And the Kernel Density Estimation exhibits all properties of Kernel Functions. Moreover, the Epanechnikov Kernel, Laplacian Kernel, Gaussian Kernels, Quadratic kernel, uniform Kernel etc are the standard kernel functions [website 1]. We have more focus towards Epanechnikov Kernel and Gaussian Kernels in our work. The Epanechnikov Kernel and Gaussian Kernels are described by standard formula [website 1] that are

$$\textbf{Epanechnikov Kernel } K(x)=\frac{3}{4}(1-x^2)1_{(|x|\le 1)}$$

And

$$\textbf{Gaussian Kernel } K(x)=\frac{1}{2\pi}e^{-\left(\frac{1}{2}\right)x^2}$$

Performance analysis and simulation results show that FTDA protocol increases the accuracy of aggregated data and reduces the amount of data transmission in the network.

For our proposed method, we need a dataset. And, we got the real data set from Intel Berkeley Research lab [website 2] and downloaded it. This dataset is freely available on their website [website 2] and have information about data collected from 54 sensors deployed in the lab between February 28th and April 5th, 2004. There is a log of about 2.3 million readings collected from these sensors in this file. The size of file is 34MB zipped, and 150MB unzipped. The schema is as follows:

date:	Time:	Epoch	moteid:	temperature:	humidity:	light:	voltage:
(yy-mm-dd)	(hh:mm:ss.xxx)	(int)	(int)	(real)	(real)	(real)	(real)

Table 3.1: Schema of dataset(Intel Lab Data)

For sensor1

Dataset	**Min**	**Max**	**Mean**	**Median**	**StdDev**
Temperature	17.1954	122.1530	35.8824	22.1444	33.6511
Humidity	-4	50.7387	34.3193	38.6334	13.8804
Voltage	2.0065	2.7624	2.5196	2.5823	0.1690

Table 3.2: Statistical characteristics for sensor1

For sensor2

Dataset	**Min**	**Max**	**Mean**	**Median**	**StdDev**
Temperature	3.4068	122.1530	40.2018	22.4286	37.8656
Humidity	-3	50.5784	34.2987	40.1284	18.0268
Voltage	0.0180	2.7244	2.4584	2.4850	0.1697

Table 3.3: Statistical characteristics for sensor2

FTDA[10] The outlier detection mechanism is based on the locality sensitive hashing (LSH) technique. The LSH algorithm used in FTDA allows compact representation of sensor data, which reduces the communication overhead of outlier detection.

FTDA takes advantage of LSH technique by estimating the similarity of sensor data from their compact sketches (LSH codes These LSH codes are sent to the data aggregator. To find out the local outlier nodes, the data aggregator compares the similarity between each LSH code pair. Then, the data aggregator communicates with the neighboring data aggregators to discover if its local outliers are affected from the phenomena occurred in the neighboring regions. The data aggregator does not include the faulty data of outliers to data aggregation process and computes the aggregated data. In addition, while detecting outliers, the data aggregator also discovers the sensor nodes that have the exact same LSH codes (i.e., sensor nodes that have the same data) and prevents redundant data transmission from these sensors. Elimination of redundant data transmission improves the bandwidth and energy efficiency of FTDA.

Our contribution in this paper is twofold. First, we propose a novel FTDA scheme using an in-network outlier detection mechanism based on LSH technique. With the help of LSH technique, FTDA protocol is able to detect outliers in a distributed and energy-efficient manner. using LSH codes, FTDA protocol eliminates the redundant data transmission from sensor nodes to data aggregators thereby incrementing the efficiency of data aggregation process. Second, We have applied the FTDA and Kernel Density technique for readings of sensor-1 and sensor-2. It estimated the density at various kernel points and gives the points where we have to plot the density. For temperature, humidity and voltage we have plotted the density estimate for sensor-1 respectively. In our work, we have derived column vector of single attributes for individual sensors with the help of 30-40 lines of C code. Each one is having approx. 50000 data samples. These vectors are used as input dataset for density estimation and also for outlier detection components. The main features have been calculated in advance as these will be used further. We have described the table of features for sensor1 and sensor2 only and on the basis of table we got following plots:

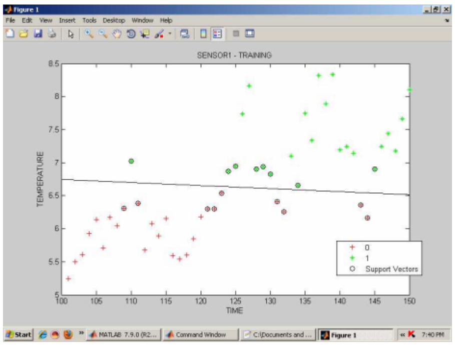

Figure 3.1: Training Set of Sensor Node

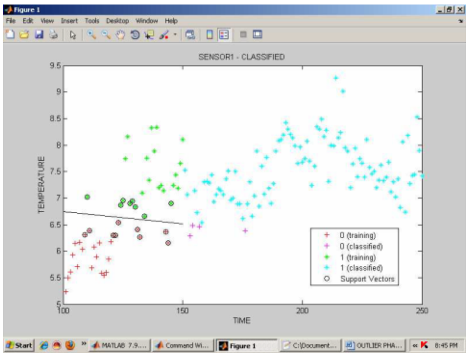

Figure 3.2: Testing of Sensor Node Data

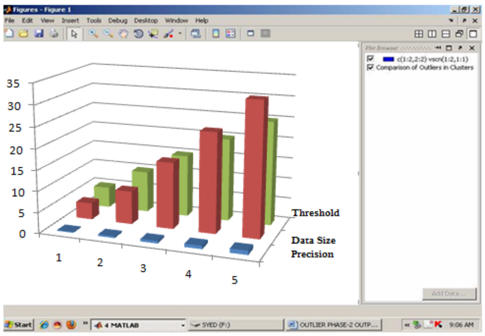

Figure 3.3: Average precision values of FTDA for different data sizes and similarity thresholds

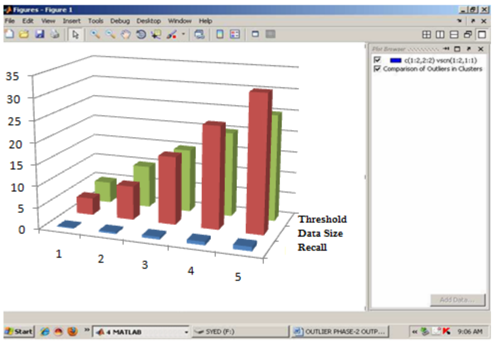

Figure 3.4: Average recall values of FTDA for different data sizes and similarity thresholds

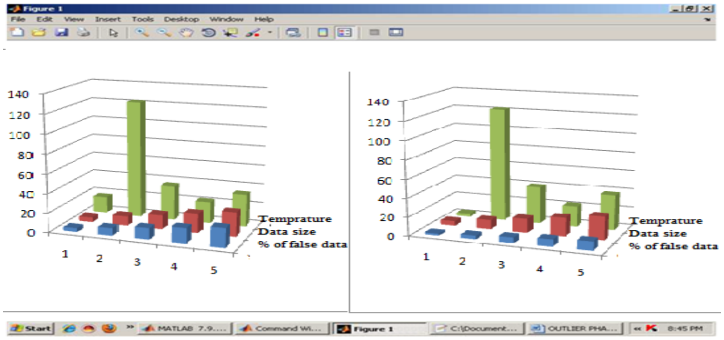

Figure 3.5: Total data transmission of (a) Sensor 1and (b) Sensor 2

IV. Conclusion And Future Work

In this work, we firmed on the focal inconvenience of outlier detection in wireless sensor networks (WSNs). Usually Outlier detection techniques centre the developer's or user's perception into the attractive events, or startling results in the network which has very short probability of happening. Instead of working on the rare sensor evaluations first, a statistical modelling technique changes it into essential information which will lean to efficient output, hence under inspection, offering a more appropriate way to achieve insight into the substantial phenomena.

Similarly, we have vacant a model that is based on the computation of the sensor data allocation. This method focuses a variety of characteristics and features of streaming sensor data. With the help of a rest of experiments with datasets taken from Intel Berkeley research lab we have processed and evaluated our planned scheme. The experiments verify that our algorithm can accomplish very high precision and recall rates for finding outliers, and display the success of the proposed approach.

We will centre on other density estimation methods like orthogonal series expansion (wavelet density estimation) as our future work. The major idea of this method is to analyze the division of dimension by determining the coefficients of its Fourier transform. Latest studies and works have confirmed that due to its local nature, wavelet based density estimation methods are superior to others. At present we are working for only attribute sensors but we will try to extend our plan for multi-attribute sensors in upcoming. And for outlier detection for multi-attribute sensors we will centre on some other idea if mandatory.

References

Reference to a book:

[1] Mohammad ilyas and Imad mahgoub , *handbook of sensor networks*: *compact wireless and wired sensing systems* , CRC Press

[2] S. Theodoridis, K. Koutroumbas, *"pattern Recognition",* 4th edition, Academic Press

Reference for an article:

[1] John, G. H. (1995), Robust linear discriminant trees, in *"Fifth International Workshop on Artificial Intelligence and Statistics", Ft. Lauderdale, FL, pp. 285-291.*

[2] Charu C. Aggarwal , Philip S, Yu, Outlier detection for high dimensional data, *ACM SIGMOD Record Volume 30 Issue 2, June 2001 Pages37-46*

[3] E. Elnahrawy and B. Nath. 2003. Cleaning and Querying Noisy Sensors. *In Proc. of the 2nd ACM International conference on WSNA'03, pages 78-87.*

[4] Victoria J. Hodge & Jim Austin 2004, a survey of techniques for outlier detection , *Artificial Intelligence Review Volume 22, Issue 2 , pp 85-126*

[5] Christoph Heinz and Bernhard Seeger. 2006. Statistical Modeling of Sensor Data and its application to Outlier Detection. Technical Report 2006/07, University of Stuttgart; *5. GI/ITG KuVS Fachgespräch "Drahtlose Sensornetze", Stuttgart.*

[6] S. Subramaniam, T. Palpanas, D. Papadopoulos, V.Kalogeraki and D. Gunopulos. 2006. Online Outlier Detection in Sensor Data Using Non•Parametric Models. *VLDB'06, September 12- 15, Seoul, Korea, pages 187-198.*

[7] Yang Zhang, Nirvana Meratnia, Paul Havinga(2008) Outlier detection Techniques for wireless sensor networks: *A survey, pp .11-20.*

[8] Blohsfeld, B., Heinz, C.,2005, Seeger, B.: Maintaining Nonparametric Estimators over Data Streams. *In: Proc. of BTW.*

[9] D. Scott. Multivariate Density Estimation: Theory, Practice and Visualization. Wiley & Sons, 1992.

[10]Suat Ozdemir : FTDA: outlier detection-based fault-tolerant data aggregation for wireless sensor networks *, Application of Information and Communication Technologies (AICT), 2011 5th International Conference on Baku , ISBN 978-1-61284-831-0 , pp 1-5*

Reference to web page:

[1] http://en.wikipedia.org/wiki/Kernel_(statistics)

[2] http://db.csail.mit.edu/labdata/labdata.html