Quest Journals
Journal of Electronics and Communication Engineering Research
Volume 4 ~ Issue 1 (2018) pp: 05-10
ISSN(Online) : 2321-5941
www.questjournals.org

Research Paper

Outlier Detection Using Kernel Functions In Wireless Sensor Networks

W.Nancy, G.M.Abisha Grace, D.Ruban Thomas
Jeppiaar Institute of Technology
Jeppiaar Institute of Technology
Vel Tech Multi Tech Dr.Rangarajan Dr.Sakunthala Engineering college
Department of Electronics and Communication
Jeppiaar Institute of Technology
Corresponding Author: W.Nancy

ABSTRACT: *A wireless sensor network consists of spatially distributed autonomous sensor node to monitor physical or environmental conditions, and to cooperatively pass the data through the network to a base station. The quality of data set may be affected by noise and error due to the fact that outliers are one of the sources that greatly influence data quality. Outliers is defined as, "those measurements that significantly deviate from the normal pattern of sensed data". In this paper we provide a comprehensive overview of the existing method in the field of outlier detection in WSNs. One of the existing methods for (S. Subramaniam et al., 2006., S. Papadimitriou et al., 2003) is LOCI (local correlation integral) algorithm. Since this algorithm has computational complexities, a statistical method to detect the outlier is proposed. In this proposed method, Kernel density estimators use kernel functions to estimate the probability distribution function (pdf) for the normal instances. A new instance that lies in the low or high probability area of this pdf is declared as an outlier. In this paper, we propose an outlier detection approach that can be classified both into statistical and density based approaches, since it is based on local density estimation using kernel functions.*
KEYWORDS: *outlier detection, statistical distribution, kernel functions.*

Received 15 Jun, 2018; **A**ccepted 30 Jun, 2018 © The author(s) 2018. **P**ublished with open access at **www.questjournals.org**

Outlier Detection Using Kernel Functions in Wireless Sensor Networks

From consists of a large number of small, low-cost sensor nodes distributed over a large area with one or possibly more powerful sink nodes gathering readings of sensor nodes. The sensor nodes are integrated with sensing, processing and wireless communication capabilities. Each node is usually equipped 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.

The WSN is not only used to provide fine-grained real-time data about the physical world but also to detect time-critical events. A wide variety of applications of WSNs includes those relating to personal, industrial, business, and military domains, such as environmental and habitat monitoring, object and inventory tracking, health and medical monitoring, battlefield observation, industrial safety and control.

A wireless sensor network consists of thousands of low cost nodes which could either have a fixed location or randomly deployed to monitor the environment. Sensors communicate with each other using the technique of multiple hop approach. The sensor networks are being connected out with the help of the base station. The communication is being initiated on with the help of the higher bandwidths.

Data is collected at the wireless sensor nodes (W. Wu, X. Cheng et al., 2007) compressed and transmitted to the gateway, which uses out wireless sensor networks that is present at the system with gateway connection.

I. OUTLIER DETECTION

An outlier is an observation, which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism. These measurements that significantly deviate from the normal pattern of sensed data

Data measured and collected by WSNs is often unreliable. 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 stringent resource constraints such as energy (battery power), memory, computational capacity, and communication bandwidth. The limited resource and capability make the data generated by sensor nodes unreliable and inaccurate.

Especially when battery power is exhausted, the probability of generating erroneous data will grow rapidly. On the other hand, operations of sensor nodes are frequently susceptible to environmental effects.

It is inevitable that in such environments some sensor nodes (M.C. Jun et al., 2006) malfunction, which may result in noisy, faulty, missing and redundant data. Furthermore, sensor nodes are vulnerable to malicious attacks such as denial of service attacks, black hole attacks and eavesdropping, in which data generation and processing will be manipulated by adversaries.

The above internal and external factors lead to unreliability of sensor data, which further influence quality of raw data and aggregated results. Since actual events occurred in the physical world, e.g., forest fire, earthquake or chemical spill, cannot be accurately detected using inaccurate and incomplete data.

It is extremely important to ensure the reliability and accuracy of (B. Sheng et al., 2007) sensor data before the decision-making process due to the fact that outliers are one of the sources to greatly influence data quality.

II. MOTIVATION OF OUTLIER DETECTION

Outlier detection in WSNs has attracted much attention. According to potential sources of outliers as mentioned earlier, the identification of outliers provides data reliability, event reporting, and secure functioning of the network. Specifically, outlier detection controls the quality of measured data, improves robustness of the data analysis under the presence of noise and faulty sensors so that the communication overhead of erroneous data is reduced and the aggregated results are prevented to be affected.

Outlier detection also provides an efficient way to search for values that do not follow the normal pattern of sensor data in the network. The detected values consequently are treated as events indicating change of phenomenon that are of interest. Furthermore, outlier detection identifies malicious sensors that always generate outlier values, detects potential network attacks by adversaries, and further ensures the security of the network. Here, we exemplify the essence of outlier detection in several real-life applications like Environmental monitoring, Habitat monitoring, Health and medical monitoring, Industrial monitoring and Target tracking.

Extracting useful knowledge from raw sensor data is not a trivial task. The context of sensor networks and the nature of sensor data make design of an appropriate outlier detection technique more challenging. According to the following reasons, conventional outlier detection techniques might not be suitable for handing sensor data in WSNs.

Resource constraints: The low cost and low quality sensor nodes have stringent constraints in resources, such as energy, memory, computational capacity and communication bandwidth. Most of traditional outlier detection techniques have paid limited attention to reasonable availability of computational resources. They are usually computationally expensive and require much memory for data analysis and storage. Thus, a challenge for outlier detection in WSNs is how to minimize the energy consumption while using a reasonable amount of memory for storage and computational tasks.

High communication cost: In WSNs, the majority of the energy is consumed for radio communication rather than computation. For a sensor node, the communication cost is often several orders of magnitude higher than the computation cost.

Dynamic network topology: frequent communication failures, mobility and heterogeneity of nodes. A sensor network deployed in unattended environments over extended period of time is susceptible to dynamic network topology and frequent communication failures. Moreover, sensor nodes may move among different locations at any point in time, and may have different sensing and processing capacities. Each sensor node may even be equipped with different number and types of sensors. Such dynamicity and heterogeneity increase the complexity of designing an appropriate outlier detection technique for WSNs.

Large-scale deployment: Deployed sensor networks can have massive size up to hundreds or even thousands of sensor nodes. The key challenge of traditional outlier detection techniques is to maintain a high detection rate while keeping the false alarm rate low. This requires the construction of an accurate normal profile that represents the normal behaviour of sensor data . This is a very difficult task for large-scale sensor network applications.

Identifying outlier sources: The sensor network is expected to provide the raw data sensed from the physical world and also detect events occurred in the network. However, it is difficult to identify what has caused an outlier in sensor data due to the resource constraints and dynamic nature of WSNs.

Thus, a challenge of outlier detection in WSNs is how to identify outlier sources and make distinction between errors, events and malicious attacks. Thus, the main challenge faced by outlier detection techniques for

WSNs is to satisfy the mining accuracy requirements while maintaining the resource consumption of WSNs to a minimum. In other words, the main question is how to process as much data as possible in a outlier and online fashion while keeping the communication overhead, memory and computational cost low.

Types of outliers

Compared to a centralized approach, in which the entire data is processed in a central place, outliers in WSNs can be analyzed and identified at different nodes in the network. This multi-level outlier detection in WSNs makes local models generated from data streams of individual nodes totally different than the global one . Depending on the scope of data used for outlier detection, outlier may be either local or global.

Local outliers are identified at individual sensor nodes, techniques for detecting local outliers save communication overhead and enhance the scalability. Local outlier detection can be used in many event detection applications, e.g., vehicle tracking, surveillance monitoring. Two variations for local outlier identification exist in WSNs. One is that each node identifies the anomalous values only depending on its historical values.

The alternative is that in addition to its own historical readings, each sensor node collects readings of its neighbouring nodes to collaboratively identify the anomalous values. Compared with the first approach, the second approach takes advantage of the spatiotemporal correlations among sensor data and improves the accuracy and robustness of outlier detection.

Global outliers are identified in a more global perspective. They are of particular interest since analysts would like to have a better understanding of overall data characteristics in WSNs. Depending on the network architecture, the identification of global outliers can be performed at different levels in the network. In a centralized architecture, all data is transmitted to the sink node for identifying outliers.

This mechanism consumes much communication overhead and delays the response time. In aggregate/clustering based architecture, the aggregator/cluster head collects the data from nodes within its controlling range and then identifies outliers. While this mechanism optimizes response time and energy consumption, it has the same problem as of centralized approach, it should be mentioned that individual nodes can identify global outliers if they have a copy of global estimator model obtained from the sink node .

III. PROBLEM FORMULATION

Existing System

In this paper, a technique which gives high accuracy in terms of estimating data distribution and high detection rate while consuming low memory usage and message transmission is proposed.

One of the statistical based approaches for the detection of the outlier is being given in LOCI algorithm. In this algorithm a set of the input data is being taken, and the mean and the standard deviation (T. Palpanas et al., 2003) for the input data is being found out.

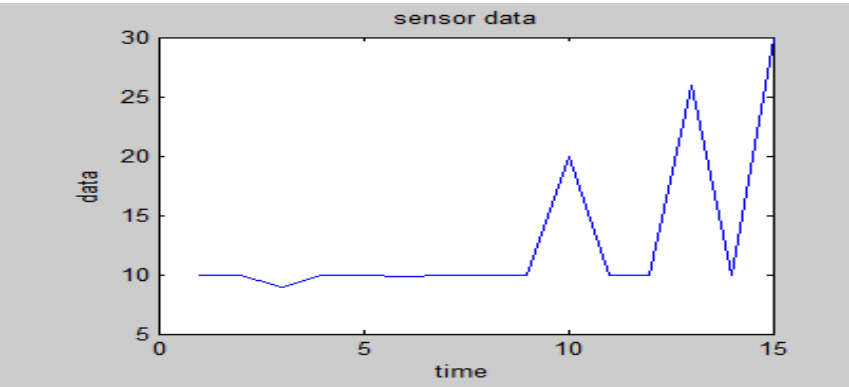

Figure 1. INPUT SENSOR DATA

The maximum and the minimum value of the data set is to be found out and they pave way for the estimation of the maximum and the minimum range of outliers that are being made available in the given region.

$$G = \frac{\max_{i=1,2,\ldots N} \left| Y_i - \overline{Y} \right|}{s}$$

with $\overline{Y}$ and s denoting the sample mean and standard deviation, respectively. The Grubbs test statistic is the largest absolute deviation from the sample mean in units of the sample standard deviation.

This is the two-sided version of the test. The Grubbs test can also be defined as a one-sided test. To test whether the minimum value is an outlier, the test statist

$$G = \frac{\bar{Y} - \bar{Y}_{\min}}{s}$$

with Ymin denoting the minimum value. To test whether the maximum value is an outlier, the test statistic is

$$G = \frac{Y_{\max} - \bar{Y}}{s}$$

Where Y_{max}, denote the maximum value.

The above equation determines the maximum and the minimum threshold value of the outlier that is to be determined.

To flag the point to be a outlier the important factor that is to be found out in the given data set is the multi-granularity deviation factor (MDEF). This factor is being calculated by

$$\text{MDEF} = \frac{(\text{MEAN of the input dataset})}{(\text{MEAN of the dataset})}$$

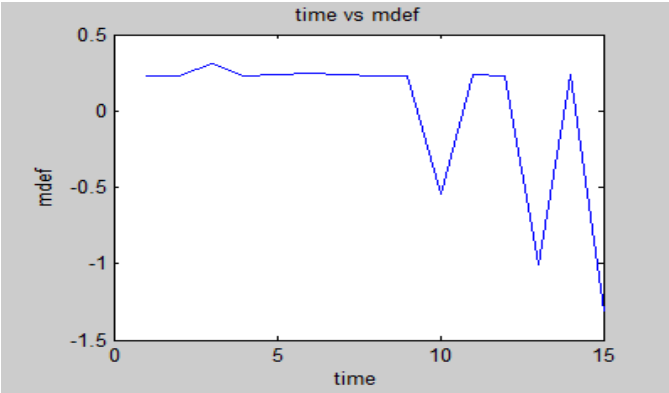

Figure 2. MDEF

The normalized standard deviation pattern of the MDEF is being given out using the formula

$$\text{SMDEF} = \frac{(\text{S.D of the dataset})}{(\text{MEAN of the dataset})}$$

To detect a point to be an outlier, the condition that is being potrayed in the given region is, MDEF > K*SMDEF

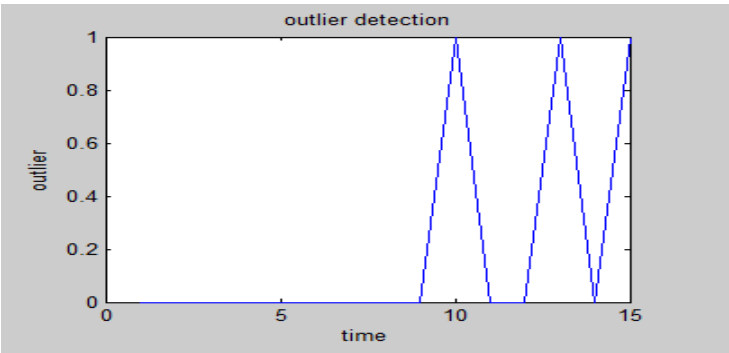

Figure 3. OUTLIERS

These are the necessary conditions that is performed to obtain the outlier using the LOCI algorithm.

Kernel Functions

The kernel density function is a statistical, non parametric based approach. This technique requires no prior known data distribution.It uses kernel density estimator to approximate the underlying distribution of sensor data. Thus, each node can locally identify outliers if the values deviate significantly from the model of approximated data distribution.

Given a data set D = f(x1,x2,....., xn), where n is the total number of data samples in Euclidean space of dimensionality dim. Our first step is to perform density estimate. One of the best-working non-parametric density estimation methods is the variable width kernel density estimator. In this method, given n data samples of dimensionality dim, the distribution density can be estimated as:

$$\tilde{q}(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h(x_i)^{\dim}} K\left(\frac{x - x_i}{h(x_i)}\right)$$

Where, K is a kernel function and h(xi) are the bandwidths implemented at data points xi. In our case, K is a multivariate Gaussian function of dimensionality dim with zero mean and unit standard deviation:

$$K(x) = \frac{1}{(2\Pi)^{\dim}} \exp\left(-\frac{\|x\|^2}{2}\right)$$

Where, ||x|| denotes the norm of the vector. The simplest version of the bandwidth function h(xi) is a constant function h(xi) = h, where h is a fixed bandwidth. The usage of the k-th nearest neighbour in kernel density estimation was first proposed.

The reachability distance for the given data set is being found out to be:

$$rd_k(y,x) = \max(d(y,x), d_k(x))$$

dk(x) is the distance to k-th nearest neighbour of point x.

The name of local density estimate (LDE) is justified by the fact that we sum over a local neighborhood mNN compared to the sum over the whole data set commonly used to compute the kernel density estimate (KDE),

$$LDE(x_j)\alpha \frac{1}{m}\sum_{x_i \in mNN(x_j)} \frac{1}{(2\Pi)^{\frac{\dim}{2}} h(x_i)^{\dim}} \exp(-\frac{rd_k(x_j,x_i)^2}{2h(x_i)^2})$$

$$LDE(x_j) = \frac{1}{m}\sum_{x_i \in mNN(x_j)} \frac{1}{(2\Pi)^{\frac{\dim}{2}} (h..d_k(x_i))^{\dim}} \exp(-\frac{rd_k(x_j,x_i)^2}{2(h.d_k(x_i))^2})$$

In order to be able to use LDE to detect outliers, the local density values LDE(xj) need to be related to the LDE values of neighboring points. We define Local Density Factor (LDF) at a data point as the ratio of average LDE of its m nearest neighbours to the LDE at the point:

$$LDF(x_j)\alpha \frac{\sum_{x_i \in mNN(x_j)} \frac{LDE(x_i)}{m}}{LDE(x_j) + c.\sum_{x_i \in mNN(x_j)} \frac{LDE(x_i)}{m}}$$

IV. CONCLUSION

A comprehensive overview of the existing methods in the field of outlier detection in WSNs is studied. In this phase, one of the existing methods for outlier detection called LOCI (local correlation integral) algorithm is being implemented. Since this algorithm has computational complexities, a statistical method to detect the outlier will be implemented in next phase. In this proposed method, Kernel density estimators use kernel functions to estimate the probability distribution function (pdf) for the normal instances the local density estimate is found and compared with the local density factor. A new instance that lies in the low or high probability area of this pdf is declared as an outlier.

REFERENCES

[1]. S. Subramaniam, T. Palpanas, D. Papadopoulos, V. Kalogerakiand, and D. Gunopulos, 2006. Online Outlier Detection in Sensor Data using Nonparametric Models, *Very Large Data Bases, 187-198.*

[2]. W. Wu, X. Cheng, M. Ding, K. Xing, F. Liu, and P. Deng, 2007. Localized Outlying and Boundary Data Detection in Sensor Networks, *IEEE Trans. Knowl. Data Eng., (19), No. 8, pp. 1145-1157.*

[3]. M.C. Jun, H. Jeong, and C.C.J. Kuo, 2006. Distributed Spatio-Temporal Outlier Detection in Sensor Networks, *Proc. SPIE.*

[4]. B. Sheng, Q. Li, W. Mao, and W. Jin, 2007. Outlier Detection in Sensor Networks, *Proc. MobiHoc.*

[5]. S. Papadimitriou, H. Kitagawa, P.B. Gibbons, and C. Faloutsos, 2003. LOCI: Fast Outlier Detection using the Local Correlation Integral, *International Conference on Data Engineering, pp. 315-326.*

[6]. T. Palpanas, D. Papadopoulos, V. Kalogeraki, and D. Gunopulos, 2003. Distributed Deviation Detection in Sensor Networks, *ACM Special Interest Group on Management of Data, pp. 77-82.*

W.Nancy. "Outlier Detection Using Kernel Functions In Wireless Sensor Networks." Quest Journals Journal of Electronics and Communication Engineering Research, vol. 04, no. 01, 2018, pp. 05–10.