Incremental Histogram Based Anomaly Detection Scheme in Wireless Sensor Networks

Ying Wang
Department of Information Engineering
Qinhuangdao Institute of Technology
Qinhuangdao, China
wyqhd@hotmail.com

Guorui Li
School of Computer and Communication Engineering
Northeastern University at Qinhuangdao
Qinhuangdao, China
lgr@mail.neuq.edu.cn

***Abstract*—Many mission critical wireless sensor networks require an efficient and lightweight anomaly detection scheme to identify outliers. In this paper, we propose an incremental histogram based anomaly detection scheme in order to detect the anomaly data values within the network. It first partitions the whole network into several clusters in which the cluster members are physically adjacent and data correlated. Then, the cluster head and cluster members update histogram incrementally and compare histograms in the form of kullback-leibler divergence differentially. We show through experiments with real data that the proposed anomaly detection scheme can provide a high detection accuracy ratio and a low false alarm ratio.**

***Index Terms*—Wireless sensor networks, security, anomaly detection, histogram.**

I. Introduction

A wireless sensor network is composed of many spatially distributed autonomous sensor nodes, to jointly monitor physical or environmental conditions, such as temperature, humidity, pressure, sound or motion [1]. In recent years, it has drawn considerable attention from both the research and industry community. The focuses are ranging from theoretical research to practical applications, such as wildlife monitoring, disaster response, military surveillance and smart building etc.

However, wireless sensor nodes are extremely restricted by their limited resources, including energy, computing, memory, bandwidth and communication [2]. In addition, they are usually deployed in unattended environments without any tamper-resistance owing to cost constraints. Furthermore, the transmitted messages among sensor nodes can be captured by any internal or external devices, caused by the use of publicly accessible communication channels. Therefore, wireless sensor networks are vulnerable to many external and internal security threats, such as bogus routing and sensed data attack, selective forwarding attack, sinkhole attack, wormhole attack and eavesdropping etc [3].

In wireless sensor networks, anomaly detection, also known as outlier detection, is the process of identify measurements that significantly deviate from the normal pattern of sensed data values. It not only can control the quality of measured data and improve the robustness of the data analysis under the presence of noise and faulty sensors, but also can detect malicious sensors and potential network attacks imposed by adversaries [4].

In this paper, we propose an incremental histogram based anomaly detection scheme in wireless sensor networks. It contains two correlated steps, data correlation based clustering step and kullback-leibler divergence based detection step. In the first step, we partition the wireless sensor network into a number of clusters. The nodes within the same cluster have the similar sensed data values and are physically close to each other. In the second step, we execute the anomaly detection using kullback-leibler divergence within each cluster based on the incremental histograms. Only two bins' values in the histogram are recomputed in each round of histogram updating process. Through experiments in which we use data released from the Intel Berkeley Research Lab, we show that our proposed anomaly detection scheme can provide a high detection accuracy ratio and a low false alarm ratio.

This paper is organized as follows. In the next section, we review some related works. In Section III, we present our incremental histogram based anomaly detection scheme. In Section IV, we analyze the scheme and present some experiment results. Finally, we conclude this paper in Section V.

II. RELATED WORKS

In wireless sensor networks, the anomaly detection scheme can be approximately categorized into the following five kinds of schemes, i.e., statistics based scheme, nearest neighbor based scheme, clustering based scheme, classification based scheme and spectral decomposition based scheme. We refer the readers to [5] for a comprehensive survey on anomaly detection schemes of wireless sensor networks.

In statistics based anomaly detection scheme, a probability distribution model which describes the distribution of the sensed data is estimated and evaluated. The data instances are declared as outliers if the probabilities of the data instances to be generated by the statistics model are very low. Based on the adopted probability distribution model, it can be further categorized into parametric and non-parametric scheme. Wu et al. proposed two local anomaly detection schemes which employ the spatial correlation of the sensed data values among neighboring sensor nodes to distinguish between outlying

978-1-4799-1027-4/13/$31.00 ©2013 IEEE

sensors and event boundary in [6]. Jun et al. designed a symmetric α stable distribution based scheme to model outliers being in form of impulsive noise [7]. Sheng et al. presented a histogram based technique to identify global outliers in data collection applications of sensor networks using histogram information to extract data distribution from the network and filter out the non-outliers [8]. Palpanas et al. proposed a kernel based technique for online identification of outliers in stream sensor data [9].

In nearest neighbor based anomaly detection scheme, several kinds of well-defined distances are usually adopted as similarity measurements between two data instances. A data instance is declared as an outlier if it is located far from its neighbors. Zhang et al. proposed a distance based technique to identify a number of global outliers in snapshot and continuous query processing applications of sensor networks by adopting an aggregation tree structure [10]. Zhuang et al. used wavelet analysis method and dynamic time warping distance based similarity comparison to identify outliers in spatiotemporal correlated sensor data [11].

In clustering based anomaly detection scheme, data instances are identified as outliers if they do not belong to any clusters or their clusters are significantly smaller than other clusters. Rajasegarar et al. proposed a clustering based anomaly detection scheme which identifies anomalous clusters when the cluster's average inter-cluster distance is larger than the predefined threshold [12].

In classification based anomaly detection scheme, unsupervised classification technique which requires no knowledge of available labeled training data is used to learn the classification model. Based on the classification model, it can be further categorized into support vector machines (SVM) based and Bayesian network based scheme. Rajasegarar et al. proposed a SVM based scheme using one-class quarter sphere SVM to decrease the computational complexity and locally detect outliers at each node [13]. Elnahrawy et al. presented a Bayesian network based scheme which maps the problem of learning spatio-temporal correlations to the problem of learning the parameters of the Bayesian classifier and then uses the classifier for probabilistic inference [14]. Janakiram et al. described a Bayesian belief network based scheme to capture not only the spatio-temporal correlations that exist among the observations of sensor nodes but also conditional dependence among the observation of sensor attributes [15].

In spectral decomposition based anomaly detection scheme, principal component analysis (PCA) is often used to reduce dimensionality before outlier detection and any data instance which violated the structure for the smallest components is regarded as an outlier. Chatzigiannakis et al. utilize PCA to distributed model the spatio-temporal data correlations in wireless sensor network and identify local outliers spanning through neighboring nodes [16].

III. Incremental Histogram Based Anomaly Detection Scheme

The incremental histogram based anomaly detection scheme includes data correlation based clustering step and kullback-leibler divergence based anomaly detection step. We first introduce the incremental histogram and its updating methods.

A. Incremental histogram

Histogram is a common technique for density estimation and has been widely used as a tool in exploratory data analysis. We build a histogram in an incremental way to reflect the dynamic status of sensed data values collected by a sensor node. The incremental pattern can avoid computing unnecessary data distributions repeatedly. The detailed incremental histogram update process is shown in Fig. 1.

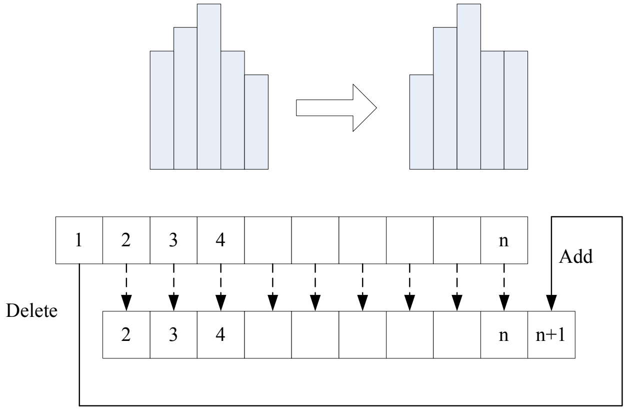

Fig. 1. The incremental histogram

We use a circular buffer of size n to hold current set of historical data values. When a new data value comes, the histogram is updated incrementally. The first element x of the circular buffer is deleted from the buffer and its corresponding bin B_i's value h_{B_i} in the histogram decreases by $1/n$. The rest of the bins' values in the histogram remain unchanged. The delete operation works as follows:

$$h_{B_i} = \begin{cases} h_{B_i} - \dfrac{1}{n} & x \in B_i \\ h_{B_i} & x \notin B_i \end{cases} \qquad (1)$$

Then, the last element y of the circular buffer is added into the buffer and its corresponding bin B_j's value h_{B_j} in the histogram increases by $1/n$. The rest of the bins' values in the histogram remain unchanged. The add operation works as follows:

$$h_{B_j} = \begin{cases} h_{B_j} + \dfrac{1}{n} & y \in B_j \\ h_{B_j} & y \notin B_j \end{cases} \qquad (2)$$

By updating histogram incrementally, only two bins' values in the histogram are recomputed. The whole histogram doesn't need to be rebuilt thoroughly. Therefore, the precious energy of wireless sensor node is reserved and the new histogram is obtained quickly.

B. Data correlation based clustering step

In data correlation based clustering step, we partition the whole sensor network into several data correlated clusters using spatial correlated weight and relative energy level [17]. If the sampled values of node i and j are denoted as *d*$_i$=(x_1,x_2,…,x_n) and *d*$_j$=(y_1,y_2,…,y_n) respectively, the Euclidean distance between them can be calculated as

$$d_{ij}=\sqrt{\left|x_1-y_1\right|^2+\left|x_2-y_2\right|^2+\cdots\left|x_n-y_n\right|^2} \tag{3}$$

Then, the expectation of d_{ij} is

$$E(d_{ij})=\sum\nolimits_{j\in N(i)} d_{ij}\Big/\left|N(i)\right| \tag{4}$$

where $N(i)$ is the h hop neighbor set of node node i. The deviation of d_{ij} is

$$D(d_{ij})=\sum\nolimits_{j\in N(i)}\left(d_{ij}-E\left(d_{ij}\right)\right)^2\Big/\left|N(i)\right| \tag{5}$$

Similar to [18], we can calculate the spatial correlated weight w_i of node i as

$$w_i=\left(\sum\nolimits_{j\in N(i)}\left|d_{ij}-E\left(d_{ij}\right)\right|\right)^2\Big/\left|N(i)\right|^2 D\left(d_{ij}\right) \tag{6}$$

According to the Cauchy-Schwarz inequality, we can get $w_i\in[0,1]$. Moreover, if the residual energy at sensor node i is denoted as e_i, the relative energy level re_i can be calculated by comparing e_i with the average energy available at other nodes within h hop neighbor set of node i. Hence, we can calculate the relative energy level of node i as

$$re_i=\left(e_i+\sum\nolimits_{i\in N(i)} e_i\right)\Big/e_i\times\left(\left|N(i)\right|+1\right) \tag{7}$$

The spatial correlated weight w_i describes the average spatial measurement deviation between node i and its h hop neighbors. Large value of w_i means node i has high spatial correlation with its h hop neighbors, which implies it should be selected as an aggregator preferentially. Small value of re_i means node i has more energy than its h hop neighbors, which implies it should be selected as an aggregator preferentially. The data correlation based clustering step is shown in Table I.

TABLE I. DATA CORRELATION BASED CLUSTERING STEP

```
Broadcast sensor's sampled value and residual energy to $h$ hop neighbors
Calculate $w_i$ and $re_i$ with $h$ hop neighbors
Wait $(re_i/w_i)T$ time
{
  If receive clustering message from $h$ hop neighbor $j$ and has not joined
    any cluster
    {
        Calculate Euclidean distance $d_{ij}$ between $i$ and $j$
        If ( $d_{ij}\le\delta/2$ )
          Join cluster and forward clustering message
    }
}
If $(re_i/w_i)T$ time is up
{
  Elect itself as cluster head
  Broadcast clustering message to its $h$ hop neighbors
}
```

C. Kullback-leibler divergence based anomaly detection step

Kullback-leibler divergence is a convenient and robust method of measuring the difference between two data sets in a statistical sense. Given two data sets P_n and Q_n, and their corresponding probability mass function p_n and q_n from data domain X at time n, the kullback-leibler divergence between them can be calculated as

$$D\left(p_n\|q_n\right)=\sum\nolimits_{x\in X} p_n(x)\log_2\left(p_n(x)/q_n(x)\right) \tag{8}$$

The probability mass function p_n and q_n of wireless sensor nodes' sensed data values can be obtained from incremental histogram. In order to reuse the already calculated part of the previous kullback-leibler divergence value, we can compute the new kullback-leibler divergence value in a differential way [19]. That is,

$$\begin{aligned}
&D\left(p_n\|q_n\right)=\sum\nolimits_{x\in X} p_n(x)\log_2\left(p_n(x)/q_n(x)\right)\\
&=D\left(p_{n-1}\|q_{n-1}\right)\\
&-p_{n-1}(a_n)\log_2\left(p_{n-1}(a_n)/q_{n-1}(a_n)\right)+p_n(a_n)\log_2\left(p_n(a_n)/q_n(a_n)\right)\\
&-p_{n-1}(a_0)\log_2\left(p_{n-1}(a_0)/q_{n-1}(a_0)\right)+p_n(a_0)\log_2\left(p_n(a_0)/q_n(a_0)\right)\\
&-p_{n-1}(b_n)\log_2\left(p_{n-1}(b_n)/q_{n-1}(b_n)\right)+p_n(b_n)\log_2\left(p_n(b_n)/q_n(b_n)\right)\\
&-p_{n-1}(b_0)\log_2\left(p_{n-1}(b_0)/q_{n-1}(b_0)\right)+p_n(b_0)\log_2\left(p_n(b_0)/q_n(b_0)\right)
\end{aligned} \tag{9}$$

where a_n and b_n are the new added bin values and a_0 and b_0 are the dropped bin values of the incremental histogram. Therefore, the major component of the latest kullback-leibler divergence value can be retained and reused. Only a few simple addition/subtraction and logarithm operations are required to update the kullback-leibler divergence value.

In the data correlation based clustering step, we partition the sensor network into clusters where sensors are physically close to each other and have the similar sensed values. Specifically, the normal sensed data values within the same cluster follow the identical trends. However, these values are not necessary in the same data range. Therefore, we can align the incremental histograms of the sensors by finding the significant values of them before computing kullback-leibler divergences. The incremental histogram alignment process is shown in Fig. 2.

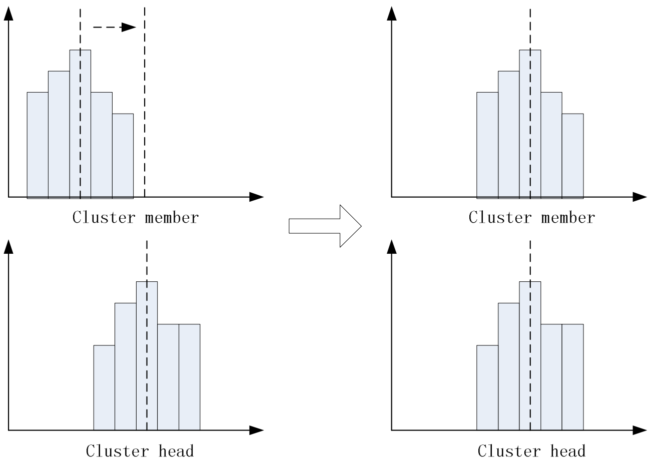

Fig. 2. The incremental histogram alignment process

The cluster head runs the following kullback-leibler divergence based anomaly detection step which is shown in Table II in each detection period T.

IV. EXPERIMENTS

In this section, we verify the performance of our proposed incremental histogram based anomaly detection scheme. The real sensed data collected from 54 Mica2Dot sensors deployed in the Intel Berkeley Research Lab between 28 February and 5 April 2004 were used to test the performance of our scheme.

TABLE II. KULLBACK-LEIBLER DIVERGENCE BASED ANOMALY DETECTION STEP

```
For each cluster member node $i$
{
    Collect $i$'s incremental histogram $h_i = \{p_i, i \in [1,\cdots,n]\}$
    Compute cluster head's incremental histogram $h_{CH} = \{q_i, i \in [1,\cdots,m]\}$
    Align $h_i$ and $h_{CH}$ by finding the significant values of them
    Calculate the kullback-leibler divergence $D(p_n||q_n)$ according to (9)
    If $(|D(p_n||q_n)| > \Delta)$
        Identify node $i$ as anomaly
}
```

The collected data include humidity, temperature, light and voltage values along with timestamp information collected once every 31 seconds. We randomly add some noise following the normal distribution to the tested data in order to simulate the abnormal behaviors.

Fig. 3 shows an instance of the data correlation based clustering step. The whole network was partitioned into 7 clusters and the cluster heads were marked with red circles. Instead of using hops as the measurement, we adopted the space distance to restrict the cluster size in the experiments for the reason of no routing structure is provided within the raw data.

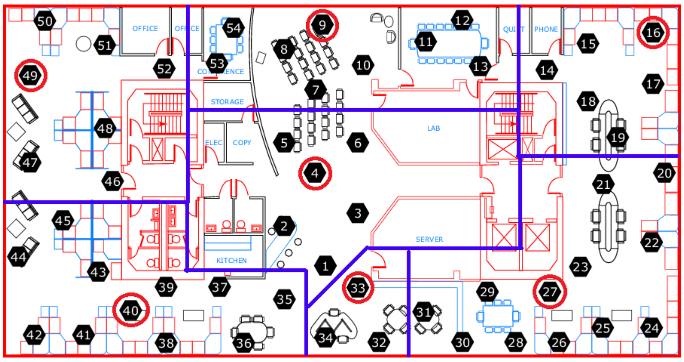

Fig. 3. An instance of the clustering step

Fig. 4 shows the data instances of the first cluster (node 1-4 and 6, node 5 was broken) between 00:00:00 to 23:59:59 at 1 March 2004. We can see that the sensors had the similar sensed data values and followed the identical trends. However, their sensed data values are not exactly the same. Therefore, by aligning the incremental histograms and computing the kullback-leibler divergences among sensors, our proposed scheme can detect the existing outliers.

The effectiveness of the proposed scheme is reflected by detection accuracy ratio and false alarm ratio. Detection accuracy ratio is defined as the percentage of abnormal data values that can be successfully detected. And false alarm ratio is defined as the percentage of the normal data values that are claimed as abnormal values. They are the two key performance indicators for any anomaly detection scheme. Fig.5 shows the detection accuracy ratio of the incremental histogram based anomaly detection scheme where the data are collected from same data set. We can see that the detection accuracy ratio is very high and it increases with the length of data set.

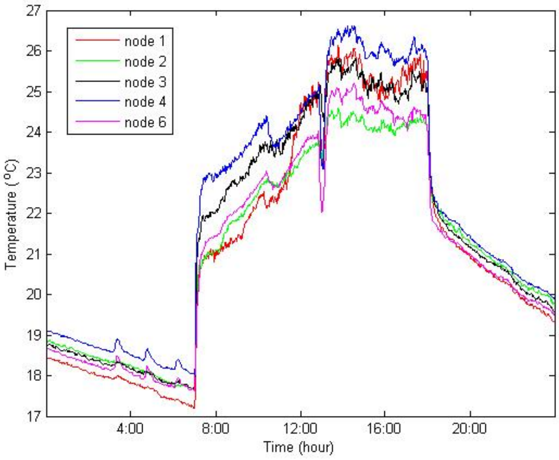

Fig. 4. The data instances of node 1-6 at 1 March 2004

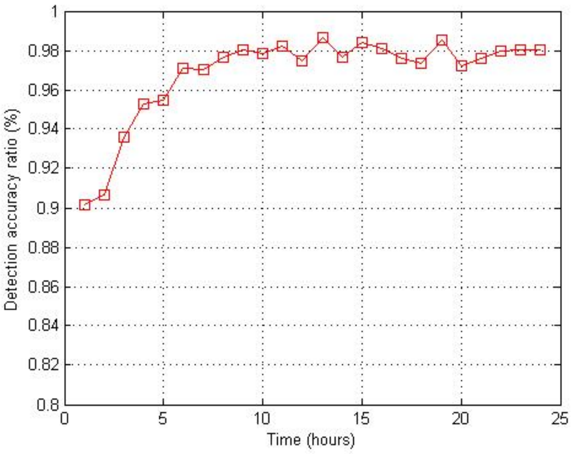

Fig. 5. Detection accuracy ratio

Fig. 6 shows the false alarm ratio of the incremental histogram based anomaly detection scheme when the data are collected from the same data set as before. We can see that the false alarm ratio is very low and it also increases with the length of data set.

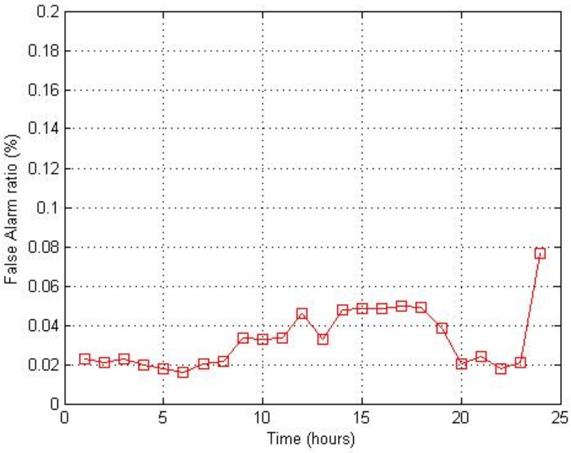

Fig. 6. False alarm ratio

V. CONCLUSION

In this paper, we proposed an incremental histogram based anomaly detection scheme for wireless sensor networks with the goal of detecting anomaly data values. By updating histogram incrementally and computing kullback-leibler divergence between cluster head and cluster members differentially, the proposed scheme can detect the outliers quickly and effectively. Our experiment results show that our detection scheme can achieve a high detection accuracy ratio and a low false alarm ratio. As a future work, we would like to implement the proposed scheme into a wireless sensor network testbed and evaluate its performance in detail.

ACKNOWLEDGMENT

The work is supported by the Fundamental Research Funds for the Central Universities of China (Program No. N120423005), the Natural Science Foundation of Hebei Province, China (Grant No. F2012501014), the Scientific Research Foundation of the Higher Education Institutions of Hebei Province, China (Grant No. Z2010215), the Research Fund for the Doctoral Program of Higher Education of China (Grant No. 20120042120009) and the Science and Technology Research and Development Project of Qinhuangdao (Grant No. 2012021A029).

REFERENCES

[1] J. Yick, B. Mukherjee, and D. Ghosal, “Wireless sensor network survey,” Computer Networks, vol. 52, no. 12, pp. 2292–2330, August 2008.

[2] M. Xie, S. Han, B. Tian, and S. Parvin, “Anomaly detection in wireless sensor networks: A Suvery,” Journal of Network and Computer Applications, vol. 34, no. 4, pp. 1302-1325, July 2011.

[3] C. Karlof, and D. Wagner, “Secure routing in wireless sensor networks: attacks and countermeasures,” Ad Hoc Networks, vol. 1, no. 2, pp. 293-315, September 2003.

[4] S. Rajasegarar, C. Leckie, and M. Palaniswami, “Anomaly detection in wireless sensor networks”, IEEE Wireless Communications, vol. 15, no. 4, pp. 34-40, August 2008.

[5] Y. Zhang, N. Meratnia, and P. Havinga, “Outlier detection techniques for wireless sensor networks: a survey,” IEEE Communications Survey & Tutorials, vol.12, no. 2, pp. 159-170, April 2010.

[6] W. Wu, X. Cheng, M. Ding, K. Xing, F Liu, and P. Deng, “Localized outlying and boundary data detection in sensor networks,” IEEE Transactions on Knowledge and Data Engineering, vol. 19, no. 8, pp. 1145-1157, August 2007.

[7] M. Jun, H. Jeong, and J. Kuo, “Distributed spatio-temporal outlier detection in sensor networks,” In *Proceedings of 1st International Conference on Communication System Software and Middleware*, pp. 5719-5831, January 2006.

[8] B. Sheng, Q. Li, W. Mao, and W. Jin, “Outlier detection in sensor networks,” In *Proceedings of 8th International Symposium on ACM Mobile Ad Hoc Networking and Computing*, pp. 219-228, September 2007.

[9] T. Palpanas, D. Papadopoulos, V. Kalogeraki, and D. Gunopulos, “Distributed deviation detection in sensor networks,” ACM SIGMOD, vol. 32, no. 4, pp. 77-82, December 2003.

[10] K. Zhang, S. Shi, H. Gao, and J. Li, “Unsupervised outlier detection in sensor networks using aggregation tree,” In *Proceedings of 3rd International Conference on Advanced Data Mining and Applications*, pp. 158-169, December 2007.

[11] Y. Zhuang, and Lei Chen, “In-network outlier cleaning for data collection in sensor networks,” In *Proceedings of 32nd International Conference on Very Large Data Bases*, pp. 41-48, September 2006.

[12] S. Rajasegarar, C. Leckie, M. Palaniswami, and J. Bezdek, “Distributed anomaly detection in wireless sensor networks,” In *Proceedings of 10th International Conference on Communication System*s, pp. 1-5, October 2006.

[13] S. Rajasegarar, C. Leckie, M. Palaniswami, and J. Bezdek, “Quarter sphere based distributed anomaly detection in wireless sensor networks,” In *Proceeding of 10th IEEE International Conference on Communication*, pp. 3864-3869, June 2007.

[14] E. Elnahrawy, and B. Nath, “Context-aware sensors,” In *Proceedings of 1st ACM International Workshop on Wireless Sensor Networks*, pp. 77-93, September 2004.

[15] D. Janakiram, A. Mallikarjuna, and P. Kumar, “Outlier detection in wireless sensor networks using bayesian belief networks,” In *Proceedings of 1st International Conference on Communication System Software and Middleware*, pp. 1-6, January 2006.

[16] V. Chatzigiannakis, S. Papavassiliou, M. Grammatikou, and B. Maglariset, “Hierarchical anomaly detection in distributed large-scale sensor networks,” In *Proceedings of 11th IEEE Symposium on Computers and Communications*, pp. 761-767, June 2006.

[17] G. Li, and Y. Wang, “Spatial Correlation based Secure Data Aggregation Scheme in Wireless Sensor Networks,” Journal of Information & Computational Science, vol. 10, no. 12, pp. 1513-1522, July 2013.

[18] Y. Ma, Y. Guo, and X. Tian, “Distributed clustering-based aggregation algorithm for spatial correlated sensor networks,” IEEE Sensors Journal, vol. 11, no. 3, pp. 642-648, March 2011.

[19] G. Li, and Y. Wang, “Differential kullback-leibler divergence based anomaly detection scheme in sensor networks”, in *Proceeding of 12th IEEE International Conference on Computer and Information Technology*, pp. 966-970, October 2012.