PROCEEDINGS OF SCIENCE

An Abnormal Data Detection Method Based on the Temporal-spatial Correlation in Wireless Sensor Networks

Yang Liu[1]
Department of Computer Science & Technology, Harbin Institute of Technology at Weihai,Weihai, 264209, China
E-mail: Liuyang322@hit.edu.cn

Ning Wang[2]
Department of Computer Science & Technology, Harbin Institute of Technology at Weihai,Weihai, 264209, China
*E-mail:*1530857750@qq.com

GuoDong Xin[3]
Department of Computer Science & Technology, Harbin Institute of Technology at Weihai,Weihai, 264209, China

Yu Peng
Automatic Test and Control Institute, Harbin Institute of Technology, Harbin150008, China

Jia Song
Department of Computer Science & Technology, Harbin Institute of Technology at Weihai,Weihai, 264209, China

In recent years, the abnormal data detection has played a vital role in the environmental monitoring of wireless sensor networks (WSNs). However, several key issues in the abnormal data detection still remain to be solved, such as the real time capability of the detection and the classification of the abnormal data. To meet the classification requirements, this paper proposes an abnormal data detection method based on the temporal-spatial correlation in WSNs. Spatial and temporal correlations of the data in WSNs are analyzed in this paper. In addition, abnormal data detection methods of two dimensions are proposed respectively. By Combining the detection results of these two dimensions, the abnormal data can be found and classified accurately. Simulations show that the method proposed in this paper can guarantee both rationality and accuracy for WSNs.

ISCC 2015
18-19, December, 2015
Guangzhou, China

[1]Corresponding Author
[2]Speaker
[3]This study is supported by the National Natural Science of China (Grant No. 61170262，61371177）

PoS(ISCC2015)070

1. Introduction

The abnormal data in WSNs is very important for the environmental monitoring. In actual situations, the abnormal data include the malicious and event data [1]. The malicious data affect the monitoring results of the base station, and reduce reliability of the network. The event data perform an important role in environmental changes. It is common and meaningful to accurately capture changes in the monitoring area and guarantee the security of network at the same time. How to achieve efficient and accurate abnormal data detection method in WSNs is the key point of this paper.

At present, abnormal data detection methods include the statistical method and the data mining method. The data mining method involves the abnormal detection based on the clustering and the abnormal detection based on the proximity [2, 3]. Statistical methods assume that the data meet a distributed model or a probabilistic model [4]. However in practice, the data distribution is usually unknown. Zhuang Y proposed a data cleaning method based on the distance to detect the abnormal data [5], which was used to mine the abnormal data after collecting an extreme abundance of data. The obvious lag of this method is that it fails to meet the real time capability requirement [6]. HU Shi proposed a method based on the neural network. Through training the neural network, he used the historical data to forecast the next moment and found the abnormal data [7]. Bernacki J proposed an abnormal detection method based on the exponential smoothing method, which could be successfully used to model the data, predict the data and detect some unusual events in the network traffic [8]. In WSNs, the spatial correlation refers to that the monitoring data of nodes is similar to its neighbors, while the temporal correlation is that each node's data is a time series. However, the temporal correlation or spatial correlation of the data has been neglected in the above methods. The abnormal data comprise the malicious data and event data. The abnormal data could be found by the above methods, but the malicious data and event data could not be distinguished.

In this paper, an abnormal data detection method in WSNs is proposed for the purpose of solving issues mentioned above. We analyze the temporal correlation and spatial correlation of the monitoring data in WSNs. By Combining the temporal correlation with the spatial correlation, a detection method based on the temporal-spatial correlation is proposed. In the spatial dimension, K-Means is used to divide nodes into several clusters. Nodes in the same cluster have a spatial correlation. Then the data mining algorithm based on distance is used to detect the abnormal data. In the temporal dimension, data are detected by using the time series analysis based on the sliding window [9]. By conducting extensive simulations, results show that the method proposed in this paper can not only identify the abnormal data, but also effectively distinguish between the malicious data and event data, and it is suitable for WSNs.

2. Abnormal Data Detection Method

In WSNs, we make the following assumptions:

1) The distribution of nodes is relatively concentrated, and there are public coverage areas.

2) The network is deployed in a trusted environment initially.

3) If no special event occurs, the data difference of a node is small in a short period of time.

4) The node comes with a clock, and the data packets are sent with the time stamp.

5) The node sleep time and monitoring period are shorter.

2.1 Spatial Abnormal Data Detection Based on the Spatial Correlation

Sensor nodes at a short range have spatial correlation, meaning that the monitoring data of these sensor nodes at a short range are similar. XUE An rong held that nodes at the communication range have a spatial correlation [10]. It is pointed out that the time complexity of determining the spatial correlation nodes is $O(n^2)$, and a large number of data packets between nodes are required. In this paper, K-Means is used to divide nodes into clusters. Clustering can make the cluster structure even and reasonable. According to the assumption 1, sensor nodes in the same cluster have a spatial correlation. In other words, if the monitoring data of a certain node have a large deviation from the other nodes in the same cluster, the data are abnormal. This method not only reduces the computation and communication overhead, but ensures rationality of the space division.

We choose the data mining method based on distance to detect the abnormal data from the spatial dimension [11], where Euclidean distance is a measure standard. The Euclidean distance of the vector $X_a(x_{a1}, x_{a2}, \ldots, x_{an})$ and $X_b(x_{b1}, x_{b2}, \ldots, x_{bn})$ is:

$$d(a,b)=\sqrt{\sum_{k=1}^{n}(x_{ak}-x_{bk})^2} \tag{2.1}$$

In this paper, only one dimensional vector is used. The Euclidean distance of X_a and X_b is :

$$d(a,b)=|X_a-X_b| \tag{2.2}$$

Test procedures are as follows:

a) For each datum X_i , calculate $d(i,j)$ from X_i to the other datum in the cluster.

b) Select an empirical value δ.

c) Count the number N of $d(i,j)<\delta$.

d) Calculate . P is the ratio of adjacent nodes of X_i , and N_0 is the number of data in this cluster.

e) Given the empirical critical value, if $P\leqslant\beta$, assess the data X_i as the abnormal data.

2.2 Temporal Abnormal Data Detection Based on the Temporal Correlation

In WSNs, sensor nodes acquire the data at a certain time circle. Each sensor node's data are time series with a temporal correlation. The monitoring data contain many types, such as temperature, humidity, light and so on. The range and form of these types are different. If different time series models are established according to different types, the time complexity will be too high, and it will be difficult to apply to WSNs. Without considering the long time dormancy, it can be found that the common concern of monitoring indexes is that the data will not have obvious fluctuation in a short period of time. Similar to the idea of integration, the monitoring change is accumulated over a series of stationary states.

Therefore, the exponential smoothing forecasting method based on the sliding window is proposed to measure the temporal dimension of the data. Compared with the time series

forecasting method, the exponential smoothing method has low computational complexity, which is suitable for WSNs. The sliding window size is set as L , meaning only L historical data of the node are analyzed. At the same time, for the measured data have no obvious trend in a relatively short period of time, an-exponential smoothing method is chosen [12].

This paper uses the exponential smoothing forecasting method with a sliding window to measure the temporal dimension of the data. Detection steps are as follows:

a) Slide the window to the previous L *moments of data* X_i .

b) Use the historical data in the window to make an exponential smoothing prediction.

c) Compare the predictive results with the actual data.

d) Set that the empirical value =， if the deviation between actual and predictive results is more than α *, assess the data* X_i *as the abnormal data.*

2.3 Abnormal Data Detection Method Based on the Temporal-spatial Correlation

In this paper, a detection method based on the temporal-spatial correlation is proposed to detect and classify the abnormal data. According to the location of nodes, these nodes are divided into several clusters in accrodance with the K-means clustering method, and nodes in the same cluster have a spatial correlation. In the spatial dimension, the data are detected by using the data mining algorithm based on distance. In the temporal dimension, the data are detected by using the time series analysis with a sliding window. The system flow chart is shown in Fig. 1. Finally, considering the two dimensional results, the data type can be classified as shown in Table 1. The event data are abnormal in a single dimension, and the malicious data are abnormal in two dimensions, while data under other conditions are normal data.

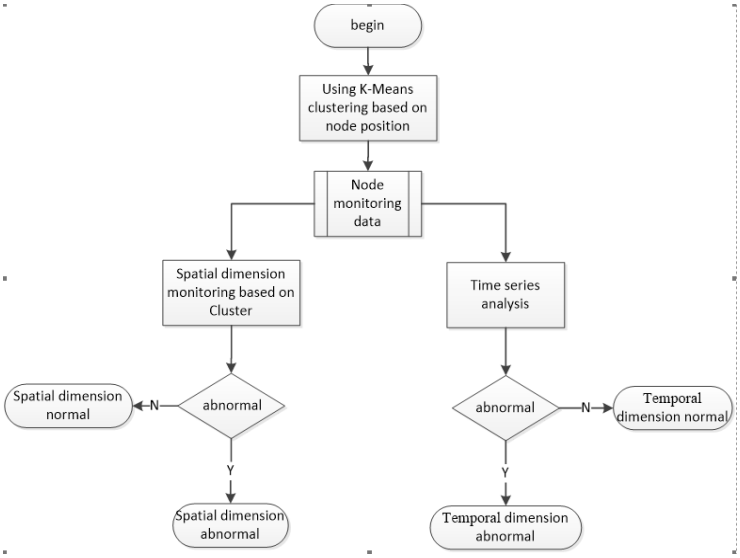

Figure 1: Abnormal Data Detection Flow Chart

Spatial dimension	Temporal dimension	Detection result
Normal	Normal	Normal data
Normal	Abnormal	Event data
Abnormal	Normal	Event data
Abnormal	Abnormal	Malicious data

Table1: Result Determination

Results show that the reputation value of nodes which send malicious data can be reduced, and those malicious nodes can be added into the network blacklist. By analyzing the event data, environmental changes in the monitoring area can be found.

3. Experiment

In this part, experiments are conducted to prove rationality and validity of the method proposed. First, a simulation is carried out to assess rationality of the space division. Then, sensor nodes are deployed to get the monitoring data. Finally, accuracy of this method is tested.

3.1 K-Means Clustering Simulation

In order to verify rationality of the K-Means Clustering method for the space division, a simulation is carried out. The simulation platform is MATLAB. The simulation models a network consisting of 100 sensor nodes placed randomly within a 100m × 100m area. The cluster number is 5. Simulation results are shown in Fig. 2. The five colors represent five clusters. From the result, we can see that the K-Means clustering can make the cluster structure even and reasonable. So using K-Means to divide nodes is feasible, and the sensor nodes in the same cluster have a spatial correlation.

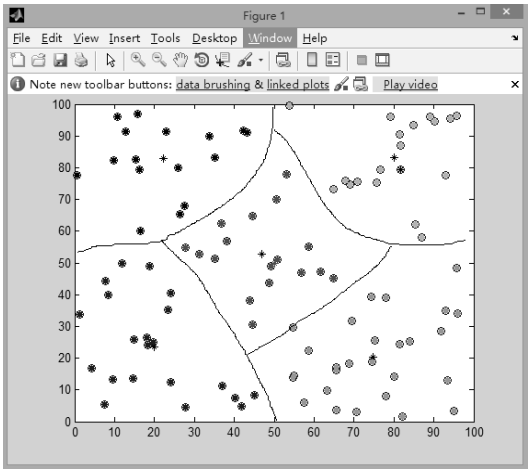

Figure 2: Clustering Cluster Simulation Results

3.2 Data and Parameters

The actual monitoring data are obtained by sampling at a small scale in the laboratory. The sensor node type is CC2530. Locations of these nodes are shown in Fig. 3.

Node distribution satisfies the assumption 1 that the distance between each other is small, so 10 nodes can be analyzed as a cluster. The time interval of the node monitoring is 30s. From the time 0, data within each 30s are processed at the same time. Temperature is chosen as the test index. We take 1,200 consecutive data from the monitoring data in about an hour (including a total of 10 nodes, each node with 120 data) to determine the validity of this method.

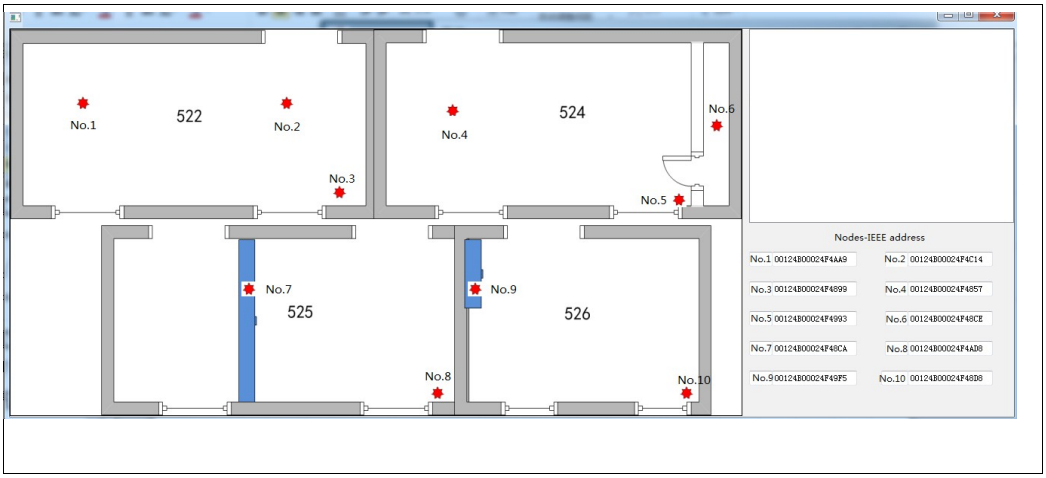

Figure 3: Nodes Location Map

The data are modified by the manual marking in the following ways:

1) All the temperatures of node 4 increase by 10℃. According to the assumptions 2, we hold that the node 4 is in a special environment. So the abnormal data belong to the event data.

2) Randomly select 4 moments, and temperatures of all nodes in these moments increase by 10℃. According to the assumption 1, it is believed that special circumstances occur at these moments, such as fire. So the abnormal data belong to the event data.

3) Randomly modify 40 data as the malicious data.

Therefore, in 1200 data, we obtain 160 event data and 40 malicious data, a total of 200 abnormal data.

Temperature is chosen as the measurement index, so parameters of this experiment can be set as shown in Table 2. δ and β can be set according to the accuracy requirement in the practical application. In this paper, an experiment is carried out with δ = 5 and= 0.75. Smoothness index, smoothing initial value and other parameters are set following the empirical value[11].

Parameter	Value
Distance empirical value δ (℃)	5
Empirical critical value β	0.75
Smoothness index α	0.8
Smoothing initial value	Average values of the first three in the sliding window
Sliding window size	10

Table 2: Parameter Values

3.3Test of Abnormal Detection Method

To evaluate the performance of algorithms, two metrics, DR (outlier detection rate) and FR (false alarm rate), are examined [13]. DR refers to the ratio of the number of correctly detected outliers to the total number of actual outliers. FR is the ratio of the number of normal objects that are misinterpreted as outliers to the total number of alarms. These two metrics are defined in (3.1) and (3.2):

$$DR = \frac{TP}{TP + FN} \tag{3.1}$$

$$FR = \frac{FP}{FP + TP} \tag{3.2}$$

The definition of TP、FP、FN、TP are shown in Table 3:

Test result	Actual state	
	Abnormal	Normal
Abnormal	TP	FP
Normal	FN	TN

Table 3: Possible Detection Results

Using the data mining algorithmbased on distance[5], the time series analysis methodbased on exponential smoothing[8] (Method 3) and the method based on the temporal-spatial correlation in this paper (Method 4) to detect the abnormal data. The data mining method includes two types, the batch data mining (Method 1) and distributed real-time mining (Method 2). DRs of above methods are shown in Fig. 4.

From experimental results, we can see that the DR of the data mining method based on distance is relatively high in the abnormal data detection. Among them, the method of batch data analysis has the highest DR, but it lacks of real-time capability. DR of the real-time mining analysis is 91%, which is slightly lower than that of the batch analysis. DR of the time series analysis method is 39.5%, and results are poor when a special circumstance occurs. Howeer, it must be pointed out that these methods can't effectively distinguish between the malicious data and event data.

Results of the detection method based on the temporal-spatial correlation proposed in this paper indicate that DR of the abnormal data detection is 91%; DR of the malicious data detection is 100%; FR of the malicious data detection is 4%; DR of the event data detection is 86.87%, as shown in Fig. 5.

From the above experimental results, we can see that, in the abnormal data detection, the temporal-spatial correlation detection method proposed in this paper is slightly poorer than the data mining method, but it has reached the acceptable range. More importantly, this method can effectively distinguish between the malicious data and event data, which is convenient for WSNs to make corresponding responses. For nodes sending malicious data, we can take measures to reduce the reputation value, and add these nodes to the blacklist. We can also take other security measures to prevent further damages.

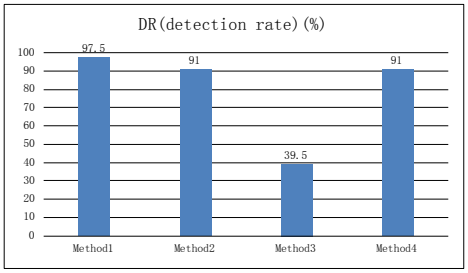

Figure 4: Abnormal Data Detection Contrast

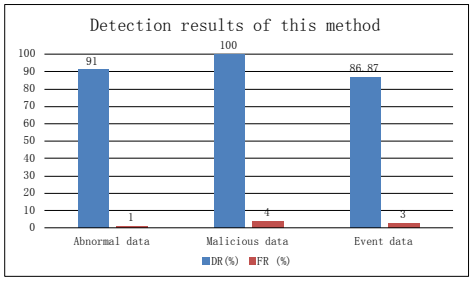

Figure 5: Results Based on the proposed method

4. Conclusion

In this paper, we firstly analyze remaining issues with regard to traditional abnormal data detection methods. Then, a detection method based on the temporal-spatial correlation is proposed to solve those issues. In this method, the data are detected through two dimensions: the temporal dimension and spatial dimension. Finally, the data type can be determined according to two dimensions results. Compared with the traditional method, this method based on the temporal-spatial correlation can effectively find and classify the abnormal data. Malicious data and event data can be distinguished accurately. In the future research, we will focus on efficiency improvements so as to make it more suitable for large-scale WSNs.

References

[1] M. M. Zeng, H. Jiang, X. Wang. *A Novel Traceback Scheme of Malicious Nodes in Wireless Sensor Networks*[J]. Chinese journal of sensors and actuators, 2013,26(1)：122-127 (In Chinese)

[2] Y. Zhang, N. Meratnia, P. Havinga. *Outlier detection techniques for wireless sensor networks: A survey*[J]. Communications Surveys & Tutorials, IEEE, 2010, 12(2): 159-170

[3] H. Li, Q. X. Wu. *"Research of Clustering Algorithm Based on Information Entropy and Frequency Sensitive Discrepancy Metric in Anomaly Detection.*[C]″ In Information Science and Cloud Computing Companion (ISCC-C), 2013 International Conference on, pp. 799-805. IEEE, 2013

[4] X. B. Jiang, G. Y. Li, S. LIAN. *Outlier detection algorithm based on variable width histogram wireless sensor network*[J]. Journal of Computer Applications, 2011,31(3): 694-697 (In Chinese)

[5] Y. Z. Zhuang, C. Lei. *"In-network Outlier Cleaning for Data Collection in Sensor Networks*[J].″ Cleandb Workshop in Vldb, 2006.

[6] H. Kim, J. K. Min. *An Energy-Efficient Outlier Detection Based on Data Clustering in WSNs*[J]. International Journal of Distributed Sensor Networks, 2014, 2014(2):365-370.

[7] S. Hu, G. H. Li, W. W. Lu. *Outlier Detection Methods Based on Neural Network in Wireless Sensor Networks*[J].Computer Science, 2014, 41(B11): 208-211 (In Chinese)

[8] J. Bernacki, G. Kołaczek. *Anomaly Detection in Network Traffic Using Selected Methods of Time Series Analysis*[J]. International Journal of Computer Network and Information Security (IJCNIS), 2015, 7(9): 10-18

[9] F. Lin, H. Zhang. *Spatiotemporal Correlation-based Outlier Detection Algorithm in Wireless Sensor Networks*[J].Computer Applications and Software, 2013, 30(6): 114-115 (In Chinese)

[10] A. R. Xue, M. Li. *Anomaly Reading Detection Algorithm in WSN*[J]. 2010,27(9): 3452-3455(In Chinese)

[11] S. Li, R. Lee, S. D. Lang. *Mining distance-based outliers from categorical data*[C]. Data Mining Workshops, 2007. ICDM Workshops 2007. Seventh IEEE International Conference on. IEEE, 2007: 225-230.

[12] C. J. Wang. *Selection of Smoothing Coefficient via Exponential Smoothing Algorithm*[J].Journal of North University of China(Natural Science Edition),2007,27(6): 558-561 (In Chinese)

[13] Y. Thakran, D. Toshniwal. *Unsupervised Outlier Detection in Streaming Data Using Weighted Clustering*[J]. Intelligent Systems Design and Applications, IEEE, 2012, 24(3): 947–952.