Journal of Computing and Electronic Information Management

ISSN: 2413-1660

A Data Anomaly Detection Algorithm of Node in Wireless Sensor Networks

Shiping Fan[1] and Chaojie He[2, a]

[1]School of Software Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China

[2]School of communication and information engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China

[a]787931148@qq.com

Abstract: In the application of wireless sensor networks, sensor nodes are typically deployed in harsh environments. They are prone to failure and become damaged nodes. They are disturbed during communication to make the resulting data anomalous and destroy the accuracy of the data. Affecting people's judgments to make the wrong thing. As a result, this paper proposes a node data anomaly detection algorithm based on singular value decomposition. The algorithm take full advantage of the spatial correlation of node data space, and uses the singular value decomposition method to deal with the data compression process. Then it uses the bootstrap method to construct the confidence interval, and detects the node data. The simulation results show that the algorithm has good anomaly detection rate and false alarm rate, and saves the energy of the node.

Keywords: Wireless Sensor Networks (WSNs), Singular value decomposition (SVD), Anomaly detection, Confidence Interval (CI)

1. Introduction

Wireless sensor networks (WSNs) is composed of spatially distributed autonomous sensor nodes for monitoring physical environments (such as temperature, humidity or pressure) [1, 2]. The WSNs deploy these low-cost sensor nodes on a large-scale uncontrolled or harsh environmental conditions, and with its limited energy and low computation ability, the nodes are prone to failure, so that the sensing data becomes unreliable. If the wrong data is transmitted to the monitoring station, it will not only result in the waste of limited resources, but also lead to the wrong response of the

workstation, which can cause serious consequences. Therefore, it is necessary to monitor the nodes' data.

Anomaly detection is for the abnormal behavior of WSNs, which can predict, record, alarm, and so on, that is, to find the data which is not consistent with the expected data. The basic idea is to define the normal mode, and then to detect the transmission of the data is not consistent with the normal model of the data, in order to filter and exclude abnormal data.

Anomaly detection in WSNs is a very important research field, many experts and scholars are doing research in this field. There are many kinds of anomaly detection algorithm, according to the definition of the model, it can be divided into: anomaly detection based on statistics, anomaly detection based on clustering, anomaly detection based on classification and anomaly detection based on near point [3]. In [4], the author puts forward the data aggregation based on spatial correlation algorithm (SCSDA) to eliminate the abnormal data; it combines the spatial correlation weight and the relative energy level to divide the WSN into several network data clusters, and then using the Mahalanobis distance and OGK estimation to detect abnormal behavior. The authors of the document [5] use a spatial model to detect potential errors and use spatial correlation between nodes to verify potential errors. The author of [6] uses the consistency detection method to judge whether the average value of the sensor data is offset, and delete the abnormal data. In [7], the authors propose a fault-tolerant data aggregation algorithm based on trust computing it between sensor nodes through the sensor readings of the time and spatial correlation credibility, eliminating the data of sensor nodes with low credibility. These algorithms have their own advantages, but there are one or more questions: (1) sensors' data should be affected by the Gauss noise, this is usually not true in practice; (2) algorithm consists of complex mathematical formulas, resulting in computational complexity increases; (3) to a large number of original data, this increases the sensor nodes' communication cost and storage space.

A feasible anomaly detection algorithm should meet the requirements of data dimensionality reduction, distributed, adaptive detection and on-line detection. Based on this, this paper improves the anomaly detection algorithm based on statistics, put forward a method based on singular value decomposition (SVD) anomaly detection algorithm SVD-CI (Based on Singular Value Decomposition of Confidence Interval). The algorithm works in layered wireless sensor network, and uses the singular value decomposition to reduction the nodes' data with spatial characteristics, according to the data of time related characteristics, it uses time series to construct a confidence interval, in order to detect abnormal data, which can effectively reduce the

computational complexity of nodes, reduce the energy loss from data transmission, and ensure a good detection performance.

2. Anomaly Detection by SVD-CI

2.1 Data Reduction Based on SVD

The data of sensor nodes have the characteristics of time and space correlation, so the data has strong correlation. The singular value decomposition method has strong data processing ability, can achieve the eigenvalues extraction, dimensionality reduction of data and model simplification, which can compress the data meanwhile keep the original data maximum maintenance.

The singular value decomposition method is an analytical tool for the numerical matrix, the matrix X given as its decomposition:

$$X = U\Lambda V^T \tag{1}$$

In the formula, U is orthogonal matrix of order m, the column vector called left singular vector; V is orthogonal matrix of order n, the column vector called right singular vector; Λ is a singular diagonal matrix m × n, the data on the diagonal matrix is called singular value, diagonal outside data values were 0. That is

$$\Lambda = \begin{bmatrix} S & 0 \\ 0 & 0 \end{bmatrix} \tag{2}$$

is a diagonal matrix, and the S=drag ($\sigma_1, \sigma_2, \cdots, \sigma_r$).In this way, the singular value of matrix X is obtained, that is $\sigma_1, \sigma_2, \cdots, \sigma_r$, The singular value is overall description of the data, if according to the order from big to small words, namely $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r \geq 0$,The larger singular values in the front represent the main distribution of the matrix, and the smaller singular values can be set to zero.

In WSNs, a cluster with m nodes, each node observation n data fixed. Computing capability and the energy is a great test to the cluster head. Using the singular value decomposition method can greatly reduce the amount of data processing nodes, and save nodes' energy.

2.2 The Algorithm of SVD-CI

In order to construct the confidence interval, the cluster head, C, firstly decomposes the data of member nodes, and then calculate the mean and variance of the member

nodes' data. Using the bootstrap method, the variance of member nodes for a few seconds sampling to build a series of data sets. The data sets obtained in this way also has an approximate distribution of real data values. The upper and lower bounds of confidence intervals are constructed by the cluster head, C. Specific steps are as follows:

(1) There are j members of sensor nodes in the cluster head of C. The data value sent by the member sensor node to the cluster header C is$V_c = v_{c1}, v_{c2}, v_{c3}, \cdots, v_{cj}$. Cluster head C calculates the average V_c of the original data setav_c. α Is a predefined limit, so the lower limit of data isd_{lower}, and the upper limit of the data isd_{upper}

$$d_{lower} = av_c \times (1 - \frac{\alpha}{2}) \tag{3}$$

$$d_{upper} = av_c \times (1 + \frac{\alpha}{2}) \tag{4}$$

Let k be an integer between 1 and j, and If v_{ck} belongs to $[d_{lower}, d_{upper}]$, it is the normal value. Otherwise, it is suspicious and will be checked by step 2.

When all data sent to the cluster head C is true, the data will be close to each other, and$v_{c1}, v_{c2}, v_{c3}, \cdots, v_{cj}$ are all belong to $[d_{lower}, d_{upper}]$. The cluster heads treat them as true and do not need to perform step 2, which helps to reduce computation. In this paper, α is a very small number, and $[d_{lower}, d_{upper}]$ is a very narrow range. Almost only when all the data sent to the cluster head of C are close to each other, v_{ck} will be identified as true, which helps to improve the detection rate.

(2) When v_{ck} is not regarded as true in step 1, the cluster head, C, will establish the confidence interval by using the bootstrap method. Steps are as follows:

(a) Calculate the variances: the cluster head, C, calculate the original data setV_c, it is$\Delta_c = \delta_{c1}, \delta_{c2}, \delta_{c3}, \cdots, \delta_{cj}$.

(b) Resampling: the cluster head, C, subsampling from the original variance set Δ_c for j times, to build a resampling variance set $\delta_{c1}^*, \delta_{c2}^*, \delta_{c3}^*, \cdots, \delta_{cj}^*$,and then the cluster head, C, calculates the mean of the set and get $\bar{\delta}_{c1}$.

(c) Sort: the cluster head, C, repeats step b N-1 times to get the$\bar{\delta}_{c2}, \bar{\delta}_{c3}, \bar{\delta}_{c4}, \cdots, \bar{\delta}_{cN}$, and sorts them from small to big, so that to gets a new variance series:

$$\bar{\delta}_{c1} \leq \bar{\delta}_{c2} \leq \bar{\delta}_{c3} \leq \cdots \leq \bar{\delta}_{cN} \tag{5}$$

(d) Confidence interval construction: the cluster head C construct confidence intervals by using variance set in step c. α' is a preset value, then the confidence

probability is 100% × (1-α') that the confidence interval can be expressed as ($\bar{\delta}_{c1}$, $\bar{\delta}_{cu}$). Here u is N×(1-α').Finally, it can get the confidence interval is:

$$[\bar{\delta}_{lower}, \bar{\delta}_{upper}] = [\bar{\delta}_{c1}, \bar{\delta}_{cu}] \quad (6)$$

(e) To eliminate the abnormal data of the original data set: V_c, if the variance of δ_{ck} is in the interval [$\bar{\delta}_{lower}$, $\bar{\delta}_{upper}$], then the node data v_{ck} is regarded as true, otherwise, v_{ck} will be regarded as false and eliminated from the data set.

The method eliminates the data with a large variance, which is usually an exception. The normal data are close to each other, their variance is less than the exception, so it will not be excluded from the data set. In this way, we can ensure that the anomaly detection algorithm proposed in this paper has high detection rate and low false alarm rate.

At the same time, the bootstrap method does not require any complex theoretical calculations or complex data distribution assumptions, which makes it suitable for data anomaly detection in wireless sensor networks.

3. Performance Evaluation

In this paper, we evaluation the performance of SVD-CI based on Matlab platform. The simulation scene is defined as a rectangular area of 100m × 100m, and 100 sensor nodes are randomly distributed. Specific parameters are defined in table 1:

Table 1 Simulation parameters and value

parameters	value
α	0.01
α'	0.05

According to the clustering method, the sensor nodes are divided into several clusters, and the cluster heads are spatially related to the member nodes. The performance of the proposed algorithm is evaluated from the detection accuracy and false alarm rate. The accuracy of the test indicates that the percentage of abnormal behavior is detected successfully, and the false alarm rate is expressed as the percentage of abnormal behavior.

3.1 The performance of SVD-CI compared to CITA

In the simulation, the probability of abnormal percentage is 10%. For example, if there are 100 nodes in the network, the data of about 10 sensor nodes is abnormal.

Fig. 1 and Fig 2 shows the detection rate and false alarm rate of SVD-CI and CITA in the case of abnormal deviation of 20%. As can be seen from the graph, the proposed algorithm has a higher detection rate and lower false alarm rate. Due to the use of bootstrap interval analysis, the confidence interval is constructed reasonably, and the abnormal data with large variance are detected and eliminated. The variance of the normal data is less than the variance of the abnormal data, and it will not be considered as an exception in the confidence interval constructed by the bootstrap method. So SVD-CI has better performance.

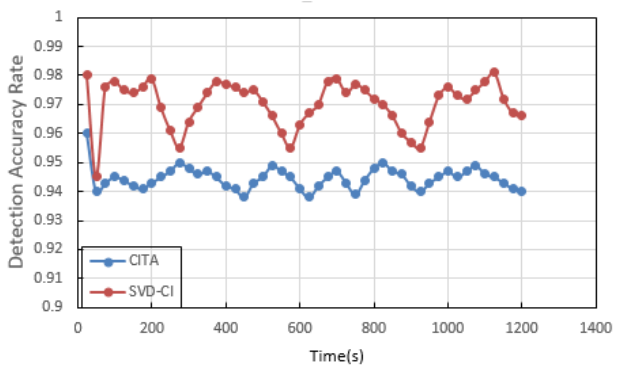

Fig. 1 Comparison of detection rates of SVD-CI and CITA

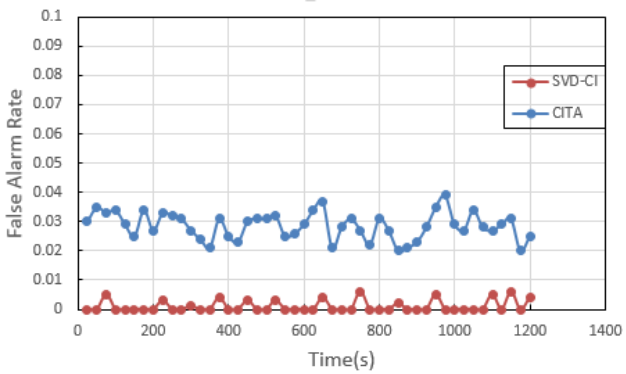

Fig. 2 Comparison of false alarm rate between SVD-CI and CITA

3.2 Compare the average residual energy of the nodes in the cluster

Fig. 3 shows the comparison of the average residual energy of the nodes in the cluster using the SVD-CI algorithm and the CITA algorithm when the anomaly deviation is 20%. It can be seen from the chart using SVD-CI algorithm and cluster node average residual energy is higher, so the algorithm is more efficient. This is because the CITA have abnormal detection for each node, and no data for screening, leading to the great amount of calculation, and more consumption of nodes' energy, and the SVD-CI algorithm of node singular value decomposition and compression the data, and retain the integrity of information and data, so it can reduce the amount of calculation and does not reduce the abnormal detection rate, to achieve energy saving and ensure the detection rate of the objective.

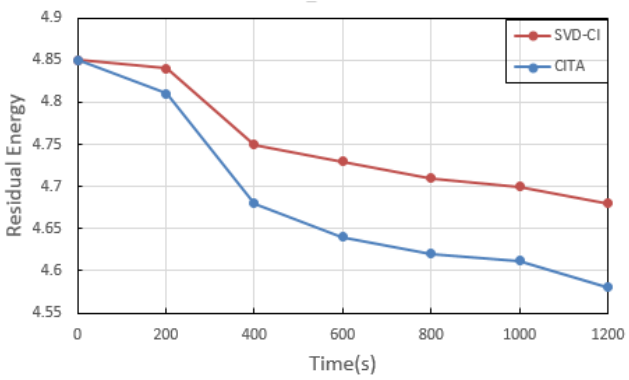

Fig. 3 Residual energy of anomaly detection

4. Conclusion

Wireless sensor network based on the spatial correlation between sensor nodes are divided into several clusters, and cluster nodes' data are very similar, in order to reduce the computational complexity and save node energy, improve the detection rate of abnormal data and remove anomaly nodes in wireless sensor networks, a singular value confidence space anomaly detection algorithm based on decomposition is proposed. In order to make the algorithm adapt to different error distribution, we use the bootstrap method to establish the confidence interval. The experimental results show that the proposed algorithm has higher detection accuracy and lower false alarm rate compared with other abnormal data detection schemes, and can greatly reduce the computational complexity and save the node energy.

Acknowledgements

This paper was financially supported by The Science and Technology Research Project of Chongqing Municipal Education Committee (Grant No. KJ1400422, KJ1500441 and KJ1400431).

References

[1] Akyildiz I F, Su W, Sankarasubramanian Y, et al. Cayirci: "wireless sensor networks: a survey[C]// International Symposium on Computer Networks. 2002.

[2] Puccinelli D, Haenggi M. Wireless sensor networks: applications and challenges of ubiquitous sensing [J]. Circuits & Systems Magazine IEEE, 2005, 5(3):19-31.

[3] Sharma AB, Golubchik L, Goninan R. Sensor faults: Detection methods and prevalence in real-world datasets [J]. Acm Transactions on Sensor Networks, 2010, 6(3):23.

[4] Li G. Spatial Correlation Based Secure Data Aggregation Scheme in Wireless Sensor Networks [J]. Journal of Information & Computational Science, 2013, 10(12):3781-3789.

[5] Ozdemir S, Xiao Y. FTDA: outlier detection-based fault-tolerant data aggregation for wireless sensor networks [J]. Security & Communication Networks, 2013, 6(6):702–710.

[6] Shu J, Hong M, Zheng W, et al. Multi-sensor Data Fusion Based on Consistency Test and Sliding Window Variance Weighted Algorithm in Sensor Networks [J]. Computer Science & Information Systems, 2013, 10(1):197-214.

[7] Sun Y, Luo H, Das S K. A Trust-Based Framework for Fault-Tolerant Data Aggregation in Wireless Multimedia Sensor Networks [J]. Dependable & Secure Computing IEEE Transactions on, 2012, 9(6):785-797.