Spatial–Temporal Anomaly Detection Algorithm for Wireless Sensor Networks

Liu Xin[1,2], Zhang Shaoliang*[1]
[1]School of Environment Science and Spatial Informatics, China University of Mining and Technology,
Xuzhou City, Jiangsu Province, 221008, China
[2]School of Medicine Information of Xuzhou Medical College,
Xuzhou City, Jiangsu Province, 221004, China
*Corresponding author, e-mail: flyinsky6@189.cn

Abstract

Traditional anomaly detection algorithms cannot effectively identify spatial–temporal anomalies in wireless sensor networks (WSNs), so we take the CO_2 concentration obtained by WSNs as an example and propose a spatial–temporal anomaly detection algorithm for WSNs. First, we detected outliers through the adaptive threshold. Then, we extracted the eigenvalue (average) of the sliding window to be detected, constructed the spatial–temporal matrix for the relationship between neighboring nodes in the specified interval, used the fuzzy clustering method to analyze the eigenvalue of adjacent nodes in spatial–temporal correlation and classify them, and identified the abnormal leakage probability according to the results of the classification. Finally, we used real datasets to verify this algorithm and analyze the parameters selected. The results show that the algorithm has a high detection rate and a low false positive rate.

Keywords*: wireless sensor networks, spatial–temporal anomaly, data stream*

1. Introduction

In recent years, wireless sensor networks (WSNs) have been applied in many fields, such as environmental and habitat monitoring, object and inventory tracking, health and medical monitoring, battlefield observation, and industrial safety and control [1]. However, the data measured and collected by WSNs is sometimes unreliable because of the resource containments of the nodes or status changes in the monitoring object. Abnormal data in sensor networks can be divided into abnormal points called “outliers” and abnormal events called “events” [2]. Abnormal points are a result of resource limitations of WSNs and sensor nodes in poor environments, which often lead to node failure and therefore result in abnormal data [3]. Abnormal events [4] are often described as a series of abnormal values in a data stream.

At present, the methods widely used in sensor anomaly detection are mainly based on several categories, including statistical model technology [5], adjacent degree technology, wavelet analysis technology [6], and cluster technology [7]. The method based on the statistical model is unsuitable for abnormal distribution data. The method based on clustering depends on the number of clusters. The methods based on adjacent degree technology and wavelet analysis are complex.

The traffic forecast model [8] uses the correlation coefficient of predicted traffic sequences and the actual flow sequence for anomaly detection. The spatial–temporal correlation characteristics of the sensor data were considered in [9], which used time and spatial correlations to generate outliers. The local outliers are converged to sink for the global outliers. This method is applied only to detect abnormal points, but sometimes, abnormal sequence detection helps to reveal abnormal events that occur. Thus, the time series of anomaly detection was more valuable in [10], which proposed to rapidly compare the similarities of two time series based on the Chebyshev coefficient and found an abnormal time sequence. The literature focused on outlier detection and a single time series of anomaly detection in the sensor data. The spatial–temporal characteristics of the sensor when the abnormal event occurred was overlooked [11].

The objective of this study is to identify the phenomenon of CO_2 leakage by analyzing the abnormal readings of sensors. Considering the analysis of CO_2 data streams, we analyze

Received March 9, 2015; Revised June 2, 2015; Accepted June 18, 2015

the spatial–temporal characteristics of each sensor when CO_2 leaks [12] and then identify the abnormal leakage effectively. In this paper, we propose the spatial–temporal anomaly detection (STAD) algorithm for WSNs. First, 3σ rules for the anomaly detection of adaptive threshold value is used. Second, Euclidean distance is employed to determine the neighbor node and extract the mean of time sequence within the sliding window, which is the eigenvalue. Then, a fuzzy similar spatial–temporal matrix of the neighbor node is constructed. Afterward, the fuzzy clustering algorithm is used to identify the abnormal probability model. Finally, the algorithm is verified using a real dataset, and the detection rate (DR) and false positive rate (FPR) of the different parameter settings are analyzed. Several references for parameter selection in the future are provided.

2. Problem Description and Definition

2.1. Background

The greenhouse gas CO_2, which is emitted by industrial production and human activities, is gradually causing global warming. The earth's environment on which people rely is increasingly deteriorating. Carbon capture and storage (CCS) is a technology that can reduce the greenhouse effect of CO_2 by storing the gas underground. The main risk of CCS is leakage. Thus, to monitor the safety of the CCS system, various monitoring technologies were used to establish a three-dimensional monitoring system. Surface CO_2 concentration monitoring is one of them. Identifying the leakage caused by monitoring data collected by sensors is the research target of the paper.

2.2. CO_2 Sensor

According to the gas diffusion in the guiding principles for environmental evaluation, while leakage occurs, the concentration of CO_2 is mainly affected by the wind speed, wind direction, and other weather conditions. Therefore, after a comprehensive consideration of system monitoring requirements, we selected sensors for temperature, humidity, wind speed, wind direction, and CO_2 concentration. The CO_2 concentration monitoring device diagram is shown in Figure 1.

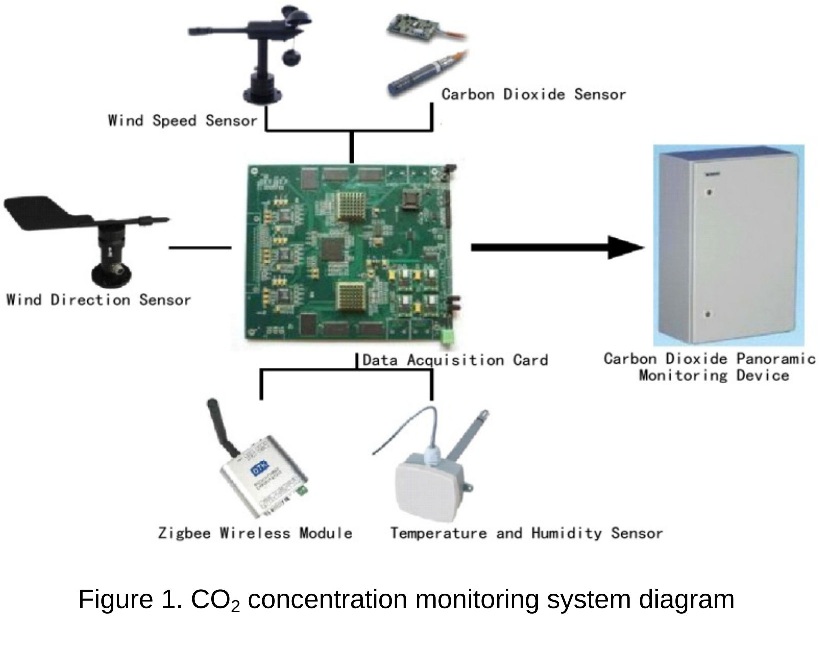


Figure 1. CO_2 concentration monitoring system diagram

2.3. Experimental Set-up

To identify the spatial–temporal characteristics of the CO_2 leakage, we designated eight sensors at equidistance of the leakage source, and each sensor is located at the same height as the leakage sources. The layout of every monitoring sensor is shown in Figure 2.

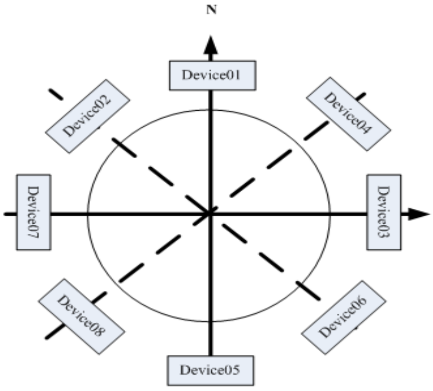


Figure 2. Sensor layout map

As the leakage rate is 20 L/min and the leakage time is 15 min, the variation of each sensor is shown in Figure 3.

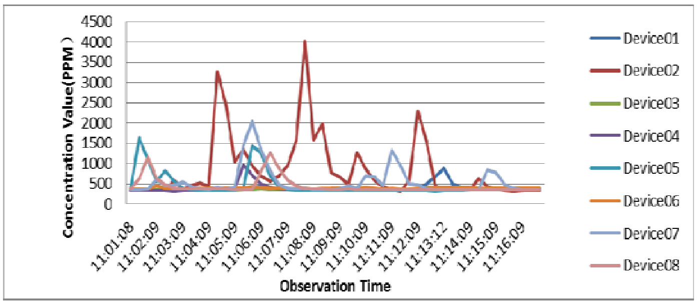


Figure 3. Concentration variation of each sensor

Figure 3 shows that, when leakage occurs, only part of the sensors' concentration levels changes significantly, and the concentration does not increase continuously but volatilely, while the difference of concentration detected by the remaining sensors is minimal.

2.4. Particularity of CO2 Anomaly Judgment

Many scholars provide different definitions of spatial–temporal anomaly. The CO_2 concentration data obtained by WSN monitoring are slightly different from those obtained by general STAD. First, the sensor data belong to the data stream. Second, CO_2 concentration is decided by diffusion, and the diffusion of CO_2 is influenced by wind speed, wind direction, and other factors. Moreover, the responses of different sensors vary. Therefore, the particularity of a CO_2 data stream is as follows:

1) Data stream

The data stream has many features, such as large and continuous amounts of CO_2, rapidity, unpredictability, infrequent scanning, and concept drift characteristics [13]. In general, the researchers proposed landmark window, sliding window, and attenuation models according to the scope of different time ranges to reduce storage and computational costs.

2) Abnormality feature

The data stream cannot effectively identify abnormal leakage through a single sensor analysis of time series data at a certain moment or through an adjacent sensor data analysis because single sensor abnormality may be caused by equipment failure. In addition, the response of each sensor to concentration is different. In general, the CO_2 leakage caused by abnormality has certain global, durability, and fuzziness features.

2.5. Definition of CO_2 Leakage Anomaly

Definition 1. Sliding window: We chose a sliding window model to represent the data stream, and assuming that the window lengths is *W*, we used *W* as the time interval. The observation value in *W* time can be expressed as time series $S_W=<s_1=(c_1,t_1), s_2=(c_2,t_2),\ldots, s_w=(c_w,t_w)>$, where s_i represents the value c_i at the moment t_i. The schematic for the sliding window is shown in Figure 4.

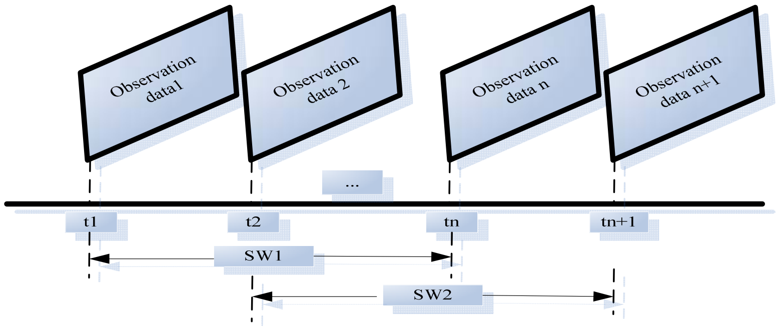


Figure 4. Schematic of the sliding window

Definition 2. CO_2 time series abnormal points: Given a time series within the sliding window, $S_W=<s_1=(c_1,t_1), s_2=(c_2,t_2),\ldots, s_w=(c_w,t_w)>$. If the newly derived observation value c_j exceeds the threshold, then that point is abnormal.

Definition 3. CO_2 leak anomalies: We determined the eigenvalue of the sliding window and the neighbor node *n* of the sensor to be detected. We obtained the classification of each node as a result of the fuzzy clustering algorithm. The probability of the abnormal leakage can be represented as the ratio of the number of sensors with the same class as the node to be detected among all the nodes. The ratio of the threshold *T* is expressed as follows:

$$Dev = \frac{count(C)}{nT} * 100\% , \tag{1}$$

Where *Count (C)* represents the number of sensor nodes in the same class as the node to be detected and *T* is the threshold. Sensor nodes are evenly distributed, so the concentration of about 50% of sensors downwind are affected; therefore, on the basis of prior experience, we set the value of *T* to 50%.

3. Algorithm Definition and Main Ideas

Considering the particularity of CO_2 leakage, we adopted fuzzy clustering algorithm [14] to analyze the spatial–temporal correlation measurements of each sensor to effectively identify anomalies.

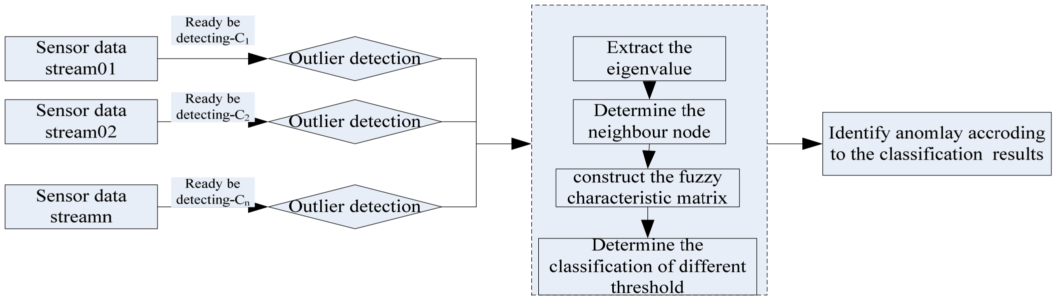


Figure 5. STAD process based on fuzzy clustering

Taking into account the lightweight requirements of sensor anomaly detection, we divided the algorithm into two phases. The first stage uses the sliding window to identify the abnormal points. At the second stage, the neighbor node is determined by extracting the eigenvalue of the sliding window. The fuzzy characteristic matrix for the eigenvalue of the sliding window specifies its neighbor nodes. We used fuzzy clustering to identify abnormal leakage probability. The process of anomaly detection is shown in Figure 5.

3.1. Abnormal Point Determination of Time Sequence

The data stream of CO_2 exhibits a strong seasonal feature; to improve the accuracy of detection threshold, adaptive problems should be considered. By analyzing the time-varying characteristics and distribution feature of the CO_2 monitoring data, we concluded, on the basis of the Chebyshev theorem of large numbers and central limit theorem, that the CO_2 concentration of the data stream in the fixed sliding window conform to normal distribution (proof omitted). The newly derived observation value in the sliding window can determine its threshold value according to the following rules.

3σ rules: If we suppose that $X \sim N(\mu,\sigma^2)$, then the probability of the normal observed values distributed in $(\mu-3\sigma,\mu+3\sigma)$ should be 99.74%. Among them, μ is the mean value of the window and σ is the standard deviation of the data within the window.

Through 3σ rules, we observed that the change in the threshold value, accompanied by the change in the mean and standard deviation, has strong adaptability.

3.2. Spatial–temporal Abnormal Judgment

As shown in the preceding analysis, the leakage determination is unusual. The eigenvalue of the multiple sensors is needed for the spatial and temporal correlation analysis. The discussion on selecting the eigenvalue, determining the neighbor nodes, and STAD based on fuzzy clustering is as follows.

1) Selecting the abnormal eigenvalue

To determine the sequence similarity degree of the adjacent nodes for the anomaly detection sensor, we used the distance of the corresponding measurements between the nodes [2]. However, we observed that CO_2 leakage caused by the change in the sensor does not have a one-to-one relationship, as shown in Figure 6.

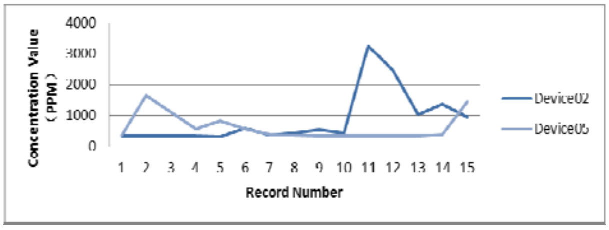


Figure 6. Comparison chart of observation data from different sensors

Figure 6 shows that directly using the observation value can cause a large error DR. Considering the characteristics of the observed value of CO_2 leakage, we chose the mean value of concentration to describe the change characteristics of the observation values within the sliding window, which can smoothen the influence of the instantaneous changes to a certain degree.

2) Determining the neighbor node

The voting decision was used to identify the neighbor node [15]. A Voronoi diagram was used to determine the adjacent node [16]. To simplify the calculation, in this paper, we determined the neighbor node based on the fact that the Euclidean distance is less than a fixed value *K*. *K* is set according to the distance of a sensor. The sensor node to be detected is *O*, the coordinate position is (*x*, *y*), the neighbor node set to be determined is $X = \{X_1, X_2,..., X_n\}$, the coordinates of the X_i is (x_i, y_i), and the distance between the nodes X_i and *O* is defined as follows:

$$dist\ (X_i, O) = \sqrt{(x_i - x)^2 + (y_i - y)^2}\ . \qquad (2)$$

If the distance of dist (X_i, O) is less than K, then it becomes the Kth neighbor of X_i when the value is O.

3) STAD based on fuzzy clustering

The data collected by sensor nodes tend to have certain spatial correlations. Generally, the relevance of space refers to the data of the nodes related to the close physical location change approximation. However, CO_2 leakage caused by the spatial correlation has a certain particularity because the diffusion of CO_2 in the atmosphere is not evenly distributed. In principle, it is spread downwind, but because of the influence of atmospheric turbulence, the wind is not stable and the concentration change of the response of each sensor also differs. No strict mathematical formulas are used to determine the leakage. The sensors also cannot identify the status of leakage strictly according to the relationship between distance on the basis of the neighbor nodes, so we need to combine the fuzzy theory and spatial–temporal variation characteristics of CO_2 diffusion to judge the leakage probability. This study uses fuzzy clustering methods to conduct the STAD. The steps are described briefly as follows:

a) The data are preprocessed.

$$x_{ij}^{'} = \frac{x_{ij} - \overline{x_j}}{\sigma_j},\ i = 1,2,...,n,\ j = 1,2,...,m \qquad (3)$$

$$\overline{x_j} = \frac{1}{n}\sum_{i=1}^{n} x_{ij},\ \sigma_j = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \overline{x_j})}\ , j = 1,2,...,m$$

b) The coefficient of similarity between samples or variables is calculated, and the fuzzy similar matrix is constructed.

$$r_{ij} = \frac{\sum_{k=1}^{m}\left|x_{ik} - \overline{x_i}\right| \cdot \left|x_{jk} - \overline{x_j}\right|}{\sqrt{\sum_{k=1}^{m}(x_{ik} - \overline{x_i})^2} \cdot \sqrt{\sum_{k=1}^{m}(x_{ik} - \overline{x_j})^2}} \qquad (4)$$

$$\overline{x_i} = \frac{1}{m}\sum_{k=1}^{m} x_{ik},\quad \overline{x_j} = \frac{1}{m}\sum_{k=1}^{m} x_{jk}$$

c) The fuzzy arithmetic is used to transform and synthesize the similar matrix. The fuzzy equivalence matrix is generated.

d) The fuzzy clustering is conducted on the basis of the different levels of interception for the fuzzy equivalence matrix.

4. Steps of the Algorithm

4.1. Overview of the Algorithm

The 3σ rules were used for each sensor data stream to detect the outliers. After the outliers were found, formula (2) was used to determine the neighbor nodes, and the correlation coefficients of the eigenvalue were calculated by formula (4). Then, one can judge whether any abnormal mode occurs according to definition 3.

4.2. Steps of the Algorithm

1) Outlier detection

Algorithm input: the point of C_i to be detected

Algorithm output: whether the point of C_i is an outlier

The algorithm steps are shown as follows:

a) The information of each sliding window (WID_i, μ_i, σ_i) is calculated and recorded.

b) The variance and the mean value is used to calculate $(\mu - 3\sigma, \mu + 3\sigma)$.

c) The observed value of C_i to be detected is read.

d) If C_i is not in the interval of $(\mu-3\sigma,\mu+3\sigma)$, it is judged as an abnormal point.

(2) Abnormal pattern recognition

Algorithm input: the coordinate position (*x*, *y*) of the abnormal sensor; the sliding window number of *m*; and the threshold value

Algorithm output: abnormal leakage

a) Formula 2 is used to determine the neighbor node.

b) The fuzzy characteristic matrix with *n* nodes and the eigenvalue of *m* sliding windows are established.

c) A series of transformations is conducted on the fuzzy characteristics matrix, and this matrix is transformed into a fuzzy equivalence matrix.

d) The fuzzy equivalence matrix is classified according to λ .

e) The probability of abnormal leakage is determined according to formula 1.

5. Experimental Verification and Analysis

Considering that no standard database is currently available for the CO_2 leakage test, we analyzed the detection results of the algorithm with the real datasets of the CO_2 leakage to verify the efficiency of the STAD algorithm in this study.

5.1. Experimental Setup Description

The experimental dataset used the field to simulate the leakage data. The experimental site is at the open square without any buildings. The monitoring date was September 21, 2014. The leakage rate was 20 L/min and the leakage lasted for 15 min. The height of the leakage source was 1 m. The height of the monitoring point was also 1m. The sensors collected data once every 20 s. The range of the wind speed was from 0 m/s to 4.5 m/s. The sensors were 0.6 m away from the leakage source. The sensor distribution is shown in Figure 2. Figure 7 depicts each monitoring data curve for each of the eight sensors.

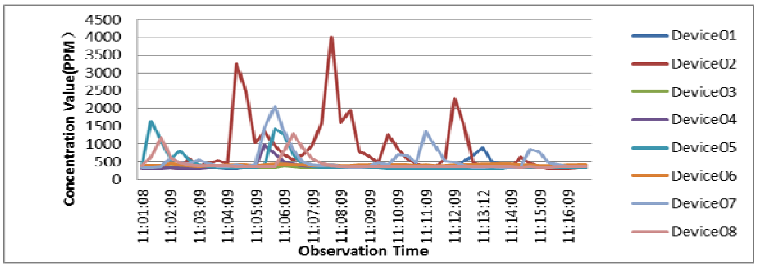


Figure 7. CO_2 Concentration curve

5.2 Data Processing

As an example based on sliding window length *L* of 10, the number of windows that need to be detected is 10 and the number *n* of nodes equals 8. The classification results of the fuzzy clusters obtained are shown as follows.

Table 1. Classification results (M = 10, L = 10)

Threshold	Number	Specific Category
λ =0.5	1	{device01,device02,device03,device04,device05,device06,device07,device08}
λ =0.6	1	{device01,device02,device03,device04,device05,device06,device07,device08}
λ =0.7	1	{device01,device02,device03,device04,device05,device06,device07,device08}
λ =0.8	3	{device01},{device02,device04,device05,device08},{device03,device06,device07}
λ =0.9	4	{device01},{device02,device04,device05,device08},{device03,device06},{device07}

The results in Table 1 show that we cannot achieve the classification effect when the value of is λ too small. The greater the value of is λ, the more accurate the classification is. When *L* is 10 and *M* is 10, the test results are the same when *L* and *M* are respectively equal to 0.8 and 0.9. To compare the detection results, we evaluate the performance of the algorithm with the accuracy of the classification results. When the classification number is 3, the abnormal sensor nodes in this experiment are devices 02, 04, 05, and 08.

We adopted DR and FPR, which are commonly used in anomaly detection as an index to measure the performance of the algorithm.

5.3. Result Evaluation

1) DR

The complexity of the algorithm is determined by the length and the number of the sliding windows included in the calculation. Thus, we analyzed the length and number.

First, we analyzed the lengths of the sliding windows. Because this factor can increase the complexity of the calculation, our analysis shows that 100% of the anomalies can be identified when the sliding window length of *L* is 10 and the number of *M* windows is equal to 10. Therefore, to reduce the computational complexity and compare the DR of the algorithm, we appropriately decreased the sliding window lengths to 6 and 10.

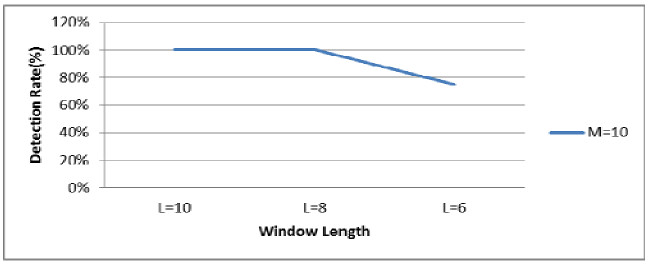
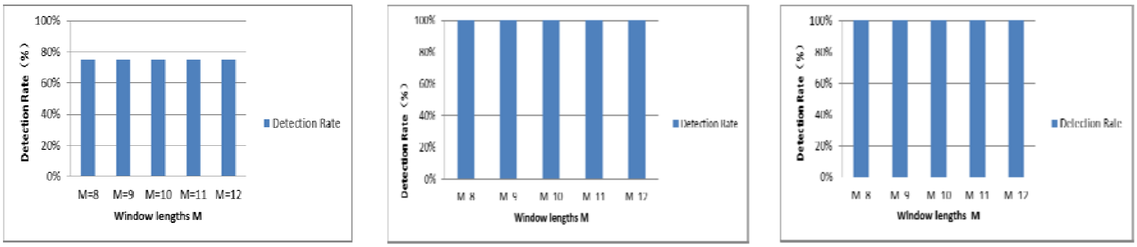


Figure 8. DRs of different window lengths when *M* is equal to 10

Figure 8 depicts that the DR was 100% when the lengths of the sliding windows were 10 and 8. The DR dropped to 75% when the lengths of the sliding windows were changed to 6. This finding shows that reducing the computational complexity is conducted at the expense of the DR.

Second, we analyzed the number of sliding windows. The following compares the DR when the sliding window number ranges from 8 to 12 and the lengths of windows are 6, 8, and 10.



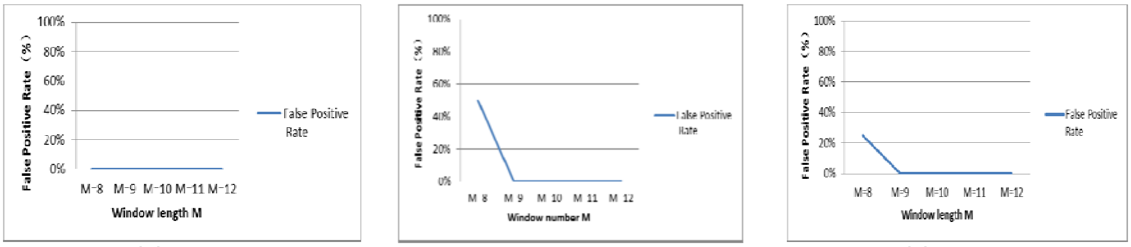
(a) *M*=6 (b) *M*=8 (c) *M*=10

Figure 9. Comparison of the DR under different window numbers and lengths

Figure 9 shows that the inspection DR of the event anomaly detection based on the fuzzy clustering is higher as a whole. Figure 9(a) depicts that the DR decreased slightly when *M* takes 6 as its value. Figure 9(b) and 9(c) depict that the DR can reach 100% when *M* is greater than 8. This finding suggests that the lengths of the sliding windows should have a value of 8 or larger in such a situation; no difference exists in the DR in terms of the number of windows.

2) FPR

The FPRs shown in Figure 10 are when the number of windows is from 8 to 12 and the lengths of the sliding windows are 6, 8, and 10.



(a) M=6 (b) M=8 (c) M=10

Figure10. Comparison of the FPRs under the different numbers of windows and different lengths

Figure 10 depicts that the FPR of the event anomaly detection based on fuzzy clustering is nearly 0 when the number of windows is larger than 10; when the number of windows is 8, the FPR is higher. The FPR increases significantly when the length of the sliding windows is 6.

In summary, the algorithm has higher DR and lower FPR as a whole for event anomaly detection. For the anomaly detection under these experimental conditions, considering the dual demand of the DR and the FPR, we suggest that the lengths of windows should be 8 or greater and the number of the detected windows should be 10 or greater.

6. Conclusion

The traditional detection method, which neglects the feature of observation values and usually adopts the static threshold method, may cause the FPR to be too high. Considering the temporal and spatial characteristics of CO_2 leakage, we proposed a STAD algorithm based on fuzzy clustering. The algorithm is divided into two stages. First, the abnormal points for every sensor are identified using 3σ rules. Second, the eigenvalue of the sliding window are extracted to create a model based on the fuzzy equivalence model to obtain the classification results under different thresholds. This method allows the identification of the abnormal probability. This algorithm extends the application scope of the fuzzy clustering algorithm. The experimental results show that the algorithm has a high DR and a low FPR. As a result of the limited conditions, the number of the simulation nodes is less and the parameter selection is too simple, which only verified the performance of the method initially. These findings should be verified on platforms with a larger number of sensor nodes.

Acknowledgements

The paper was supported by the 12th Five Year science and technology support Plan No. 2011BAC08B03, a project funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions under Grant No. SZBF2011-6-B35.

References

[1] McDonald Dylan, Sanchez Stewart Madria Sanjay, et al. A Survey of Methods for Finding Outliers in Wireless Sensor Networks. *Journal of network and systems management*. 2015; 23(1): 163-182.

[2] Shahid N, IH Naqvi, SB Qaisar. Characteristics and classification of outlier detection techniques for wireless sensor networks in harsh environments: a survey. *Artificial Intelligence Review*. 2015; 43(2): 193-228.

[3] Huanhuan C, Jian H, Kai W, et al. Research on WSNs fault detection method based on node similarity. *Transducer and Microsystem Technology*. 2014; 33(4): 10-13.

[4] Shahid N, IH Naqvi, SB Qaisar. Real Time Energy Efficient Approach to Outlier & event detection in wireless sensor networks. *IEEE International Conference on Communication systems*. 2012; 162-166.

[5] Qiuyan Y. Research on Mining Method of Mine Probabilistic Stream Data. *China University of Mining and Technology*. 2011.

[6] Zhiyuan L, Qiuzhi Z, Yongkun W, et al. Wavelet Analysis-Based Real-time Anomaly Detection Algorithm for Wireless Sensor Network. *Journal of Nanjing Normal University (Natural Science Edition).* 2014; 37(1): 87-92.

[7] Kim Hongyeon, Min Jun-Ki. An Energy-Efficient Outlier Detection Based on Data Clustering in WSNs. *International Journal of Distributed Sensor Networks*. 2014.

[8] Zhenghong X, Zhanfu X, Zhiyang C. An Anomaly Detection Approach Based on Traffic Prediction and Correlation Coefficient for WSN. *Microelectronics & Computer*. 2009; 7: 214-216.

[9] Anrong X, Ming L. Anomaly reading detection algorithm in WSN. *Application Research of Computers*. 2010(9): 3452-3455.

[10] Qi T, Xuejun L. Outlier Time Series Detection Based on WSN. *Journal of Transduction Technology*. 2013(1): 95-99.

[11] Yu Genjian, Weng kunpeng. Intrusion detection technology of layered wireless sensor network based on Agent. *TELKOMNIKA Indonesian Journal of Electrical Engineering*. 2013; 11(8): 4238-4243.

[12] Sano Fuminori, Akimoto Keigo, Wada Kenichi. Impacts of different diffusion scenarios for mitigation technology options and of model representations regarding renewables intermittency on evaluations of CO_2 emissions reductions. *Climatic Change*. 2014; 123(3-4): 665-676.

[13] Kumar Ashok PM, Vaidehi V. Anomalous Event Detection in Traffic Video Based on Sequential Temporal Patterns of Spatial Interval Events. *KSII Transactions on Internet and Information systems*. 2015; 9(1): 169-189.

[14] Weilin Li, Pan Fu, Erqin Zhang. Application of fractal dimensions and fuzzy clustering to tool wear monitoring. *TELKOMNIKA Indonesian Journal of Electrical Engineering*. 2013, 11(1): 187-194.

[15] Yinghui Qiu, Chao Liu. Modelling and stimulation of target tracking and localization in wireless sensor network. *Technical Gazette*. 2014; 21(2): 233-245.

[16] Yan W, Yuchun P, Hui W. Based on Voronoi and Information entropy spatial Outliers Detection Algorithms. *Computer Engineering and Design*. 2010; 18: 3998-4000.