Research Article

Abnormal Data Detection in Sensor Networks Based on DNN Algorithm and Cluster Analysis

Min Li1 and Anil Sharma2

1Xinxiang Vocational and Technical College, Xinxiang, Henan 453006, China
2Department of Computer Science, Faculty of Technology, Debre Tabor University, Debre Tabor, Ethiopia

Correspondence should be addressed to Anil Sharma; anilsharma@dtu.edu.et

Received 10 December 2021; Revised 19 January 2022; Accepted 14 March 2022; Published 30 March 2022

Academic Editor: Pradeep Kumar Singh

In order to solve the abnormal behaviors in wireless sensor networks, such as attacks, intrusions, node failures, and data anomalies, a data anomaly detection method for sensor networks based on DNN algorithm and cluster analysis is proposed. The deep neural network is introduced into the wireless sensor network, and each wireless sensor data is described by the neuron to construct the neural network element model. The traditional neural network model is improved, and the neural network model of the wireless sensor is used to realize the fusion and extraction of the data collected by the wireless sensor network. The clustering technology is used to complete the abnormal data judgment of the nodes, and at the same time, the spatial correlation of the sensing data between neighboring nodes is used to filter the noise data, extract the abnormal event information, and assist the system decision-making. The results show that when the number of neighbor nodes increases, the number of nodes with similar physical locations in the optimal neighbor node set gradually increases, and the accuracy rate is also continuously improved. However, if too many nodes are selected, the algorithm will use the data of nodes with a longer physical distance to vote, which will lead to an increase in the error rate. Therefore, when determining the number of neighbor nodes, 25%-30% of the total number of nodes in the positioning scenario can be selected. Therefore, the algorithm can effectively detect and distinguish environmental noise and abnormal events in the network.

1. Introduction

With the development and progress of communication technology, embedded computing technology and sensor technology, sensor networks with perception, computing, and communication capabilities increasingly show broad application prospects. The wireless sensor network comprehensively utilizes sensor technology, embedded computing technology, distributed information processing technology, and communication technology to monitor, perceive, and collect various environmental information in the network distribution area in real time, process this information, and transmit it to the users who need this information, so as to make scientific and reasonable decisions [1]. When the wireless sensor network (WSNS) is affected by external events (such as forest fires, geological disasters, and air pollution) or the sensor node itself has software or hardware failures (such as software defects, insufficient battery power, and electromagnetic interference), the measurement data of sensor nodes will be abnormal [2]. Real-time and efficient detection of abnormal data in the sensor network is of great significance for both the early warning and prevention of external emergencies and the health monitoring of the sensor network itself [3]. The data processing flow of the wireless sensor network is shown in Figure 1. For the clustering structure of wireless sensor networks, an agent-based intrusion detection system was designed by Dong et al. An IDS agent with two different agents is deployed on each node of the network. One is a local detection agent, and the other is a global detection agent. They perform different detection missions. Based on the Bluetooth communication technology, the Bluetooth scattering network formation algorithm TPSF is used to build the sensor network cluster layer and complete the task assignment of different agent. The TPSF algorithm is improved, limiting the role of the nodes, reducing the complexity of the nodes, and
enabling the IDS agent to work effectively and improve the safety coefficient of the node [4].

The localization of wireless sensor network nodes is the basis and prerequisite for many applications, and the ranging technology based on received signal strength indication (RSSI) is the more commonly used positioning method. However, due to the large amount of reflection, scattering, and diffraction in signal transmission, the abnormal problems of sensing data in node positioning are inevitable. For example, changes in the environment, occlusion of obstacles, and personnel flow in network communication often lead to jumps in the signal strength values between nodes [5]. Other examples are the small size and limited energy of the node. When the node has low energy or is damaged by natural changes and man-made damage, the change of RSSI value is more difficult to estimate. This makes the ranging technology based on received signal strength have large errors [6]. If such information cannot be detected in time, it will not only consume limited node communication energy but also limit the positioning accuracy in the network. In addition, the abnormal data in practical applications is mainly divided into two categories: environmental noise and abnormal events. The event data often contains information that people are interested in or can assist in decision-making [7].

2. Literature Review

At present, the methods widely used in this field mainly include techniques based on statistical models, proximity, and clustering. Statistics-based methods are the oldest, but they are not suitable for nonnormally distributed data. The method based on proximity is more complicated, and the method based on clustering is more dependent on the number of clusters used. Premkumar and Sundararajan used the kernel density method to estimate the distribution of perceptual data in the sliding window and determined the abnormal data by the spatial density of the data, and the calculation was complicated [8]. Yang et al. combined the filter and kernel density to calculate the confidence value of each sampled value, which is easy to implement, but with low detection rate [9]. Manikandan and Chinnadurai proposed a recognition method of environmental data flow significantly deviating from historical patterns. This method is an autoregressive data-driven model based on the data stream and its prediction interval. It has the characteristics of fast execution, large amount of data, and having no need to classify outliers in advance [10]. Singh and Chen proposed an abnormal data detection algorithm based on widening histograms, which clustered dynamic sensing data in the network into widened histograms through data clustering to accurately detect abnormal data, avoiding unnecessary data transmission [11]. Li et al. carried out anomaly detection based on the correlation coefficient changes between the predicted flow sequence and the actual flow sequence of the WSNS node, and the experimental results showed the effectiveness of the method [12]. Priyadarshi and Gupta used a neural network multilayer perceptron model, combined with rolling learning-prediction mechanism, to propose a method of using historical data modeling to estimate the current data and detect outliers according to the difference between the estimated value and the actual measured value [13]. Wang et al. proposed a data prediction
On the basis of the above studies, the spatiotemporal correlation method is improved in this paper. On the basis of pre-existing methods, Syarif et al. proposed that the artificial neural network is a parallel distributed processor formed by the interconnection of processing units and their undirected signal channels called connections [1].

On the basis of the above studies, the spatiotemporal correlation method is improved in this paper. On the basis of previous research, the clustering technology replaces the temporal correlation method, and an outlier detection technique based on cluster analysis and spatial correlation (ODCASC) is proposed. The algorithm first uses clustering technology to determine whether the node sampling data is abnormal. If it is abnormal, the information of neighbor nodes is used to distinguish the nature of abnormal data. The complexity of the ODCASC algorithm is low, and the experimental data analysis in the indoor scene confirms that the algorithm can effectively solve the problem of the spatiotemporal correlation method and realize the detection and analysis of abnormal data in positioning.

### 3. Research Methods

#### 3.1. Outlier Detection Technique Based on Cluster Analysis and Spatial Correlation (ODCASC)

The ODCASC outlier detection algorithm is mainly divided into three stages: node self-judgment, neighbor node judgment, and node self-determination. The node first starts from the similarity of the sampled data in the sliding window and uses clustering technology to determine whether the data in the window is abnormal. If there is an abnormal value, it uses the similarity of the RSSI (Received Signal Strength Indication) vector received between nodes to find neighbor nodes. Then compare the current node perception data and the data in the corresponding window of the neighbor node to see if the change trend is similar. If more than 50% of the neighbor nodes are similar to it, the current sampling data is considered to be an abnormal value of the event; otherwise, it is an abnormal value of noise. From the implementation process of the ODCASC algorithm, it can be seen that the clustering technology and the spatial correlation method essentially use the concept of correlation, but the scope of the applied data is not the same [10]. Clustering technology is mainly used to find outliers in the data correlation of the node itself, and spatial correlation is mainly used to determine the abnormal data correlation between nodes.

#### 3.1.1. Node Self-Judgment

When a node receives new perception data, it will enter the self-judgment stage to judge whether the data is abnormal. Considering the limited memory space of nodes, a sliding window is maintained at each node to store the data arrived in the last $\Delta t$. Many detection methods use time-series correlation technology at this stage; that is, compare the similarity between the data in the current window and the historical data in the previous window to determine the outliers. However, the detection rate of this technology is usually closely related to factors such as the sampling window and the similarity threshold. When the value of the impact factor is different, the detection rate often differs greatly [17, 18]. Next, we do a set of experiments to verify the sensitivity of temporal correlation theory to threshold. In the $9 \times 7$ m positioning area, place 15 unknown nodes and one reference node, set obstacles in the scene after a period of observation, and use the spatiotemporal correlation method to detect. Only the timing similarity threshold changes, and the remaining thresholds remain unchanged. The test result is shown in Figure 2. It

![Graph showing abnormal data detection rate vs threshold values for time-series autocorrelation](image)

**Figure 2:** The relationship between abnormal value detection rate and threshold.

### Table 1: Input neurons of the training sample set.

| Time slot | S1 | S2 | S3 | S4 | S5 | S6 |
|-----------|----|----|----|----|----|----|
| Week 1    | 0.63 | 0.22 | 0.14 | 0.34 | 0.01 | 0.45 |
| Week 2    | 0.62 | 0.21 | 0.15 | 0.23 | 0.01 | 0.44 |
| Week 3    | 0.70 | 0.24 | 0.11 | 0.25 | 0.01 | 0.38 |
| ...       | ... | ... | ... | ... | ... | ... |

### Table 2: Output neurons of the training sample set.

| Time slot | S1 | S2 | S3 | S4 |
|-----------|----|----|----|----|
| Week 1    | 0.11 | 0.10 | 0.45 | 0.20 |
| Week 2    | 0.01 | 0.08 | 0.65 | 0.22 |
| Week 3    | 0.12 | 0.21 | 0.42 | 0.11 |
| ...       | ... | ... | ... | ... |
can be seen that the size of the threshold has a great impact on the detection rate, and in any scenario, it is impossible to know the changing law between the data in advance, and it is difficult to value the appropriate factors. Therefore, the use of time series correlation method has certain limitations.

In the positioning scenario, the processed data set is the RSSI vector from the node to be located to each reference node. From the perspective of clustering, for each node to be located, the number of clusters in the data set should be the same as the number of reference nodes. Therefore, at this stage, cluster detection technology is used to avoid the influence of uncertain factors in time series method.

Taking the node \( n_i \) as an example, the sliding window data from \( n_i \) to the reference node \( j \) at the time \( t \) is \( R_{ij}(t) = \{ r_{ij}(t - \Delta t + 1), r_{ij}(t - \Delta t + 2), \ldots, r_{ij}(t) \} \) and represents the received data from \( n_i \) to the node \( j \) at the time \( t \), and there are a total of \( W \) data in \( \Delta t \), that is, the sliding window size. Suppose point \( O \) is the center point of \( R_{ij}(t) \), establish a scoring function \( Z \) according to the distance from each data in the window to point \( O \), and evaluate the abnormality of each data. Then, set the sum of the absolute error (SAE) as the objective function. Starting from the perceptual data corresponding to the maximum score, if the SAE of this group of data can be significantly improved after deleting this value, this data is abnormal data [11]. Proceed from large to small until SAE changes little when deleting a data. The specific calculations are shown in formulas (1) to (4).

Considering that the fluctuation range of the sensing data under normal conditions is 0~5 dB, in the experiment, the change of SAE less than 5 dB is defined as little change in SAE, which solves the problem that the similarity threshold in terms of time correlation is difficult to determine.

\[
d(r_{ij}(m), r_{ij}(n)) = \sqrt{(r_{ij}(m) - r_{ij}(n))^2}, \quad m, n = 1, \ldots, W, \quad \Delta t,\quad (1)
\]

\[
O = \arg \min_{m} \left\{ E_m | E_m = \sum_{m=1}^{W} d(r_{ij}(m), r_{ij}(n)) \right\}, \quad (2)
\]

\[
Z(m) = \frac{d(r_{ij}(m), O) - \text{median}(D)}{\text{median}(D)}, \quad (3)
\]

\[
\text{SAE} = \sum_{m=1}^{W} d(r_{ij}(m), O). \quad (4)
\]

3.1.2. Node Self-Determination. Nodes need to make decisions on the nature of the outliers. If more than 50% of the neighbors participating in the voting at 1, then the abnormal event occurs, and the abnormal cause can be analyzed according to the demand; otherwise, it is the noise data. After the decision, the average of the remaining perceptual data in the sliding window replaces the perceptual data at the current moment and realizes the correction of the abnormal data.

3.2. Anomaly Data Detection Method of Wireless Sensor Network Based on Deep Neural Networks (DNN)

3.2.1. Deep Neural Network Training. After determining the input neuron and output neuron, the next step is to train the sample set of the neural network model to determine the network direction and finally realize the prediction [19]. Based on the real-time data of a certain area through the wireless sensor network, take the 10-week observation data of this area as a sample for training and learning. The maximum number of training times is 2 million times, and the network convergence error square sum index is \( 0.5 \times 10^{-6} \), the initial learning rate is 0.01, and the momentum constant is 0.9. Input neurons and output neurons of some samples are listed in Tables 1 and 2.

Perform repeated iterative neural network training according to the above settings, and stop when the convergence meets the set network connection weights and thresholds.

4. Results and Discussion

This paper uses the built ZigBee-based wireless sensor network indoor positioning platform to analyze the ODCASC algorithm in actual scenarios. The realization of the positioning system adopts the sensor node of the IRIS model, which adopts the AtmelRF230 wireless transceiver and the At-mega128l microprocessor conforming to IEEE802.15.4. The internal program of the node is written in the nesC language on the basis of the Tiny OS operating system [20, 21].

In the experiment, the emission energy of the node is set to 8 dBm, and the experiment scene is the indoor laboratory of the college building, with \( 9 \times 7 \) m in size. A total of 4 reference nodes and 15 nodes to be located are placed. In order to evaluate the performance of the algorithm, the experiment set up two test scenarios: (a) the frequent walking of personnel causes abnormal noise of perceived data, and three curves represent the walking route of personnel; (b) an obstacle is set in front of a reference node to simulate the abnormal event of reference node failure.

Figure 3: The effect of sliding window size.

\[
O = \arg \min_{m} \left\{ E_m | E_m = \sum_{m=1}^{W} d(r_{ij}(m), r_{ij}(n)) \right\}, \quad (2)
\]

\[
Z(m) = \frac{d(r_{ij}(m), O) - \text{median}(D)}{\text{median}(D)}, \quad (3)
\]

\[
\text{SAE} = \sum_{m=1}^{W} d(r_{ij}(m), O). \quad (4)
\]
the five-pointed star symbol represents the reference node, and the square symbol represents the node to be located.

In the test, the detection rate (DR) and false alarm rate (FAR) are used as the metrics for the detection performance of the algorithm. At the same time, the correct rate (RR) and error rate (WR) are used as the metrics of algorithm to distinguish between events and noise. In formula (5), $X$ represents the data set, $Y$ represents the abnormal data set in $X$, $Y'$ represents the abnormal data set detected by the algorithm, $Y_1$ represents the noise data set in $Y$, $Y_2$ is the event abnormal data set in $Y$, $Y'_1$ is the noise data set detected in the algorithm, and $Y'_2$ is the detected event abnormal data set.

$\text{DR} = \frac{|Y \cap Y'|}{|Y|}$, $\text{RR} = \frac{|Y_1 \cap Y'_1|}{|Y_1|}$, or $= \frac{|Y_2 \cap Y'_2|}{|Y_2|}$. (5)

4.1. Analysis of Outlier Detection Rate. The timing similarity threshold and the size of the sliding window have always been the primary factors affecting the detection rate and false alarm rate of various outlier detection algorithms. The ODCASC algorithm does not involve the problem of similarity threshold in the node self-judgment stage, avoiding the impact of the threshold. Considering that this part is mainly for the detection rate of outliers, there is no need to distinguish the nature of outliers. The following first uses the measured data to investigate the influence of the window size $W$ and compares the ODCASC algorithm with the statistics algorithm, Hampel-KDE, and STCOD algorithm, where the corresponding thresholds of Hampel-KDE algorithm and STCOD algorithm are fixed. It can be seen from Figure 3 that the STCOD algorithm, which is an algorithm based on spatiotemporal correlation, is most affected by the window size. When the timing similarity is constant, the larger the sliding window, the more historical data is used, the greater the similarity of the front and rear window data, and reduce the sensitivity to the abnormal judgment of new sampling values, the lower the sensitivity to the abnormal judgment of the new sampled value. A statistics algorithm is not suitable for small sample data. As the number of samples in the window increases, the false alarm rate decreases slightly. The Hampel-KDE and ODCASC algorithms are hardly affected by the window size, but the Hampel-KDE algorithm has the highest false alarm rate. Therefore, no matter how the window size changes, the ODCASC algorithm can maintain a high detection rate and a low false alarm rate, and the detection performance is always close to the best detection result.
4.2. Differentiation Analysis of the Nature of Outliers. The following experiment is mainly to analyze the outlier property discrimination of ODCASC algorithm in the second and third stages of the algorithm. The influencing factors in this part are the number of neighbor nodes and the similarity threshold. The selected value will not affect the detection rate and false alarm rate of outliers but will only affect the property discrimination of outliers. This paper first uses the data of one of the reference nodes received by the node in the environment with abnormal noise to analyze the influence of the number of neighbor nodes. It can be seen from Figure 4 that when the number of neighbor nodes is small, because the indoor environment is more complicated, the neighbor nodes selected according to the similarity of the RSSI vector may not be in the adjacent area of the node’s physical location, which may cause misjudgment. With the continuous increase of the number, the number of nodes with similar physical locations contained in the optimal neighbor node set gradually increases, and the accuracy rate is also continuously improved. However, if too many nodes are selected, the algorithm will use the data of nodes with a longer physical distance to vote, which will lead to an increase in the error rate. Therefore, when determining the number of neighbor nodes, 25%-30% of the total number of nodes in the positioning scenario can be selected.

Figure 5 shows the abnormal discrimination of experimental data in environmental noise when the optimal neighbor node set is not used, and Figure 6 shows the corresponding discrimination when the optimal neighbor node set \((k = 4)\) is used.

In contrast, the correct discrimination rate and false discrimination rate of noise in Figure 5 change greatly with the increase of the similarity threshold, and the discrimination rate of abnormal events is basically unchanged. However, the variation range of the noise discrimination rate in Figure 6 is small, and the accuracy rate is improved as a whole. Therefore, the optimal neighbor node set method can effectively suppress the sensitivity of the outlier discrimination rate to the similarity threshold and improve the correct discrimination rate. In this case, the similarity threshold has little effect on the algorithm to correctly distinguish the nature of outliers. It is helpful to set the corresponding threshold size when the data change law in the positioning scene is not known.

5. Conclusion

Through the wireless sensor network, users can obtain real-time monitoring data of various different areas. It is one of the important technologies of the next-generation communication network technology. The inherent characteristics of the wireless signal propagation environment make the abnormal problem of perception data inevitable and seriously affect the positioning accuracy of nodes and targets. This paper proposes a sensor network data anomaly detection based on DNN algorithm and clustering analysis, which can use existing wireless sensor network resources to achieve high-speed data transmission and provide users with real-time data prediction needs. This algorithm can not only be used in indoor and outdoor positioning scenes, with less effect on the detection performance by several factors, but also can distinguish the nature of abnormal data in the scene. Based on the spatiotemporal correlation method, cluster analysis is used to replace the time series analysis, which effectively compensates for the deficiencies of the time series correlation method. Experimental results confirm that the detection rate and false alarm rate of the algorithm have reached satisfactory results. With the development of information technology and communication technology, people’s requirements for data acquisition and data transmission rate are increasing, and the traditional communication technology has been difficult to meet people’s needs. A wireless sensor network is the product of achieving this goal and requirement. Through the wireless sensor network, users can obtain the monitoring data of various different regions in real time, and it is one of the important technologies of the next generation of communication network technology. Wireless sensor networks have been applied in many fields and have achieved good results.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] A. Syarif, A. Abouaissa, L. Idoumghar, P. Lorenz, R. Schott, and G. S. Staples, "New path centrality based on operator calculus approach for wireless sensor network deployment," *IEEE Transactions on Emerging Topics in Computing*, vol. 7, no. 1, pp. 162–173, 2019.

[2] N. Goyal, M. Dave, and A. K. Verma, "Data aggregation in underwater wireless sensor network: recent approaches and issues," *Journal of King Saud University - Computer and Information Sciences*, vol. 31, no. 3, pp. 275–286, 2019.

[3] H. Yang, F. Li, D. Yu, Y. Zou, and J. Yu, "Reliable data storage in heterogeneous wireless sensor networks by jointly optimizing routing and storage node deployment," *Tsinghua Science & Technology*, vol. 26, no. 2, pp. 230–238, 2021.

[4] R. H. Dong, H. H. Yan, and Q. Y. Zhang, "An intrusion detection model for wireless sensor network based on information gain ratio and bagging algorithm," *International Journal of Network Security*, vol. 22, no. 2, pp. 218–230, 2020.

[5] M. Kanthimathi, R. Amutha, and K. S. Kumar, "Energy efficiency analysis of differential cooperative algorithm in wireless sensor network," *Cluster Computing*, vol. 22, no. S4, pp. 9837–9845, 2019.

[6] E. H. Houssein, M. R. Saad, K. Hussain, W. Zhu, and M. Hassaballah, "Optimal sink node placement in large scale wireless sensor networks based on Harris’ hawk optimization algorithm," *IEEE Access*, vol. 8, no. 99, pp. 19381–19397, 2020.

[7] P. Rama and S. Murugan, "Localization approach for tracking the mobile nodes using FA based ANN in subterranean wireless sensor networks," *Neural Processing Letters*, vol. 51, no. 2, pp. 1145–1164, 2020.
[8] M. Premkumar and T. Sundararajan, “DLDM: deep learning-based defense mechanism for denial of service attacks in wireless sensor networks,” *Microprocessors and Microsystems*, vol. 79, no. 8, article 103278, 2020.

[9] L. Yang, Y. C. Eldar, H. Wang, and H. Qian, “An ADMM-ResNet for data recovery in wireless sensor networks with guaranteed convergence,” *Digital Signal Processing*, vol. 111, no. 3, article 102956, 2020.

[10] S. Manikandan and M. Chinnadurai, “Effective energy adaptive and consumption in wireless sensor network using distributed source coding and sampling techniques,” *Wireless Personal Communications*, vol. 118, no. 2, pp. 1393–1404, 2021.

[11] P. Singh and Y. C. Chen, “Sensing coverage hole identification and coverage hole healing methods for wireless sensor networks,” *Wireless Networks*, vol. 26, no. 3, pp. 2223–2239, 2020.

[12] H. Li, W. Dong, Y. Wang, Y. Gao, and C. Chen, “Enhancing the performance of 802.15.4-based wireless sensor networks with NB-IoT,” *IEEE Internet of Things Journal*, vol. 7, no. 4, pp. 3523–3534, 2020.

[13] R. Priyadarshi and B. Gupta, “Area coverage optimization in three-dimensional wireless sensor network,” *Wireless Personal Communications*, vol. 117, no. 2, pp. 843–865, 2021.

[14] T. Wang, C. Lv, X. Jin, F. Chen, and C. Wang, “A secure and verifiable continuous data collection algorithm in wireless sensor networks,” *Wireless Personal Communications*, vol. 119, no. 3, pp. 2265–2285, 2021.

[15] A. Boukerche, Q. Wu, and P. Sun, “Efficient green protocols for sustainable wireless sensor networks,” *IEEE Transactions on Sustainable Computing*, vol. 5, no. 1, pp. 61–80, 2020.

[16] J. John and P. Rodrigues, “MOTCO: multi-objective Taylor crow optimization algorithm for cluster head selection in energy aware wireless sensor network,” *Mobile Networks & Applications*, vol. 24, no. 5, pp. 1509–1525, 2019.

[17] D. Manju, S. Singh, S. Kumar, A. Nayyar, and L. Mostarda, “Proficient QoS-based target coverage problem in wireless sensor networks,” *IEEE Access*, vol. 8, no. 1, pp. 74315–74325, 2020.

[18] K. Hamouid, S. Othmen, and A. Barkat, “Lstr: lightweight and secure tree-based routing for wireless sensor networks,” *Wireless Personal Communications*, vol. 112, no. 3, pp. 1479–1501, 2020.

[19] S. A. Elsaid and N. S. Albatati, “An optimized collaborative intrusion detection system for wireless sensor networks,” *Soft Computing*, vol. 24, no. 16, pp. 12553–12567, 2020.

[20] B. Al-Ghamdi, M. Ayaida, and H. Fouchal, “Performance evaluation of scheduling approaches for wireless sensor networks,” *Wireless Personal Communications*, vol. 110, no. 3, pp. 1089–1108, 2020.

[21] J. Du, B. Kang, and Y. Han, “Improvement on a biometric based user authentication scheme in wireless sensor networks using smart cards,” *Wuhan University Journal of Natural Sciences*, vol. 25, pp. 68–74, 2020.