ISSN 2522-3488 Vol. 7, Issue 8: 39-46, DOI: 10.25236/IJNDES.2023.070807

Method for Cleaning Outliers in Massive Data of Internet of Things Based on Hierarchical Clustering Algorithm

An Hailong

Shanghai Briup Technology Inc., Kunshan, Jiangsu, 215311, China
duck8815@126.com

Abstract: *In order to clean the outliers of massive data in the Internet of Things and improve the security of data mining and storage, a method of cleaning the outliers of massive data in the Internet of Things based on hierarchical clustering algorithm is proposed. Levenshtein matching method is used to construct the abnormal feature analysis model of massive data of the Internet of Things, and the method of attribute value correlation analysis is combined to classify the aggregate words of massive data of the Internet of Things, dynamic feature weighting method is used to match the equivalent attribute values of data abnormal values, evidential reasoning framework is used to realize hierarchical clustering of data, and the probability of abnormal value distribution of massive data of the Internet of Things is detected according to the matching probability of related attribute values. Combining the method of levenshtein and attribute value correlation analysis, hierarchical clustering of data is realized based on the method of aggregate word vector, and the outliers of data are cleaned according to the clustering results. The simulation results show that this method can improve the data purity, reduce the interference of abnormal data and reduce the computational complexity, and has better matching effect and wider applicability.*

Keywords: *Hierarchical clustering; Internet of things; Massive data; Outlier cleaning*

1. Introduction

With the continuous development of information technology and big data, the Internet of Things, as a new network structure, is widely used in the fields of data monitoring and information perception. Through the networking technology of the Internet of Things and wireless sensor networking, it is especially proposed in the "Twelfth Five-Year Plan" of the Internet of Things issued by the Ministry of Industry and Information Technology. Information processing technology (including mass data storage, data mining, intelligent analysis of images and videos), information perception technology, information transmission technology and information security technology are listed as four technological innovation projects[1]. At present, with the large-scale development of Internet of Things technology, the data scale in the Internet of Things system is increasing, so it is necessary to focus on the research on the method of cleaning the abnormal values of mass data in the Internet of Things, mine highly reliable association rule parameters from data with quality defects, and realize dynamic monitoring of abnormal data through data collection, transmission and processing[2].

In the research on the abnormal value cleaning of massive data in the Internet of Things, data evaluation, data cleaning, data monitoring and error warning are realized through the analysis of data quality problems, data improvement and management[3]. At present, the methods for cleaning the abnormal value of massive data in the Internet of Things mainly include entity analysis, attribute value matching and missing value filling, etc. In reference [4], an abnormal data detection and cleaning method based on duplicate record detection is proposed. The regular entity matching method is used to detect the abnormal data of repeated records, and the attributes are evaluated according to the information gain on the training data set to improve the cleaning ability of abnormal data. However, this method has poor reliability in dynamic detection of abnormal data. In reference [5], an outlier cleaning method based on the best combination of classification attributes is proposed. The entity matching method based on probability model can obtain the dynamic parameters of data and improve the classification ability of abnormal data, but this method has poor dynamic matching ability for network data. In reference [6], an algorithm for abnormal data detection and entity matching of Internet of Things data based on adaptive neural network learning is proposed, which realizes abnormal detection of massive data through word vector feature matching in deep learning, but the matching effect of this method depends on the selection of initial parameters.

Aiming at the disadvantages of traditional methods, this paper proposes a method of cleaning outliers in massive data of Internet of Things based on hierarchical clustering algorithm. Levenshtein matching method is used to construct the abnormal feature analysis model of massive data in the Internet of Things, and the method of attribute value correlation analysis is combined to classify the aggregated words of massive data in the Internet of Things, and the dynamic cleaning of abnormal data values is realized based on support vector machine and decision tree. Finally, the experimental test shows the superior performance of this method.

2. Abnormal data analysis and related word classification of massive data in Internet of Things

2.1 Internet of Things massive data abnormal data analysis

In order to realize the anomaly detection of massive data in the Internet of Things, unsupervised deep learning method is adopted to establish the feature analysis model of outliers of massive data in the Internet of Things and build a network learning model[7], as shown in Figure 1.

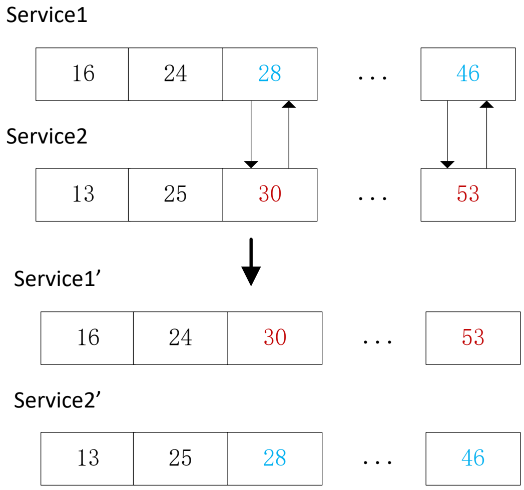

Figure 1: Dynamic learning model of Internet of Things data

When there is abnormal data, it is encoded twice, and the encoder learns the characteristic analysis method of time series, and optimizes its own objective function through the generator and discriminator, so as to obtain the diversity learning function of the generated samples of Internet of Things data, which is expressed as:

$$\min_{G}\max_{D} V(D,G) = E_{x\sim p_x}[lb(D(\boldsymbol{x}))] + E_{z\sim p_z}[lb(1-D(G(z)))] \tag{1}$$

Where, the minimax game function is expressed as $V(D,G)$, which has undergone twice coded abnormal fuzzy mapping on massive data, and is expressed as $\varepsilon : x \to z$. The discriminant is trained by Wasserstein-1 distance, and the time series analysis model of abnormal data is constructed. The penalty function is constructed by the sample generator, and the data is coded and mapped to the potential space again, so as to obtain the entity matching model based on learning[8]. Through the dependence analysis of conditional function, the template function of automatic tuple matching in large-scale relational data is expressed as:

$$\begin{aligned} p\left[s\middle|\ \boldsymbol{c}_{\mathrm{d}}\right] &= \prod_{w\in s} p\left(w\middle|\ \boldsymbol{c}_{\mathrm{d}}\right) \\ &= \prod_{w\in s}\left(\alpha p(w)+(1-\alpha)\exp\left(\tilde{c}_{\mathrm{d}},\boldsymbol{v}_{w}\right)/Z_{\tilde{c}_{\mathrm{d}}}\right) \end{aligned} \tag{2}$$

Where, S is a similarity measurement parameter, c_{d} is a neighbor missing distribution function, α is a neighbor tuple, $p(w)$ is a matching rule function, $Z_{\tilde{c}_{\mathrm{d}}}$ is a neighbor tuple, and R_1 and R_2 are detailed matching rule for mining. Given the association rule items from the distribution areas of two mass data target attribute sets, Y_1, Y_2 are expressed as: the target ordered pair set with equivalent attribute values is S:

$$\mathbf{S}=\left\{<y_i, y_j>\mid<y_i, y_j>\in Y_1\times Y_2\wedge y_i\simeq y_j\right\}, \tag{3}$$

Where, y_i, y_j represent the entity object set of mass data distribution in the Internet of Things. When $Y_1=Y_2\wedge R_1=R_2$ is met, an abnormal feature analysis model of mass data in the Internet of Things is constructed by using levenshtein matching method, and the aggregation words of mass data in the Internet of Things are classified into multiple related data by combining with attribute value correlation analysis method[9], so that the problem of equivalent attribute value is transformed into the problem of matching equivalent attribute values in the data table. According to Bayesian theory, the conditional probability density of basic model matching through related attributes is:

$$P\left(y_1\simeq y_2\right)=\frac{P\left(y_1\simeq y_2\mid x_1\simeq x_2\right)}{P\left(x_1\simeq x_2\mid y_1\simeq y_2\right)}\cdot P\left(x_1\simeq x_2\right) \tag{4}$$

Where, $x_1\in X\left[y_1\right]$, $x_2\in X\left[y_2\right]$, $P\left(x_1\simeq x_2\right)$ represent generalized correlation analysis to approximate the two conditional probabilities on the right. Conditional probabilities are essentially obtained by matching the existing attributes of related entities X and Y, and the generalized correlation eigenvalues of abnormal data of massive data in the Internet of Things are obtained:

$$f(Y,X)=\frac{P\left(y_i\simeq y_j\mid x_i\simeq x_j\right)}{P\left(x_i\simeq x_j\mid y_i\simeq y_j\right)} \tag{5}$$

To sum up, the matching probability of ordered pairs is estimated as $<y_1, y_2>$, and the dynamic feature weighting method is used to match the equivalent attribute values of data outliers[10].

2.2. Related word clustering

The hierarchical clustering of data is realized by using the framework of evidential reasoning, and the probability of outlier distribution is detected according to the matching probability of related attribute values[11]. The matching probability can be expressed as

$$P\left(y_1\simeq y_2\right)\approx f(Y,X)\cdot P\left(x_1\simeq x_2\right) \tag{6}$$

Set two attribute values of massive data of the Internet of Things to match if and only if they are completely equal, and get the correlation factor of Y and X:

$$f(Y,X)=\frac{P\left(y_i\simeq y_j\right)}{P\left(x_i\simeq x_j\right)} \tag{7}$$

The extended model not only considers the same attribute values, but also extends the basic model to obtain the similarity between the ordered pairs of related words:

$$P\left(x_1\simeq x_2\right)=sim\left(x_1, x_2\right) \tag{8}$$

Calculating the average similarity of outliers in massive data of Internet of Things

$$P\left(x_{1i} \approx x_{2j}\right) = \frac{\sum_{i,j} sim\left(x_{1i}, x_{2j}\right)}{\left|X\left[y_1\right]\right| \cdot \left|X\left[y_2\right]\right|} \tag{9}$$

Where, $x_{1i} \in X\left[y_1\right], x_{2j} \in X\left[y_2\right], sim\left(x_{1i}, x_{2j}\right)$ represents the distribution levenshtein, average similarity and matching order even value of tuples within equivalence classes of massive data in the Internet of Things:

$$P\left(x_{ai} \approx x_{aj}\right) = \frac{\sum_{i,j} sim\left(x_{ai}, x_{aj}\right)}{\left|X\left[y_a\right]\right| \cdot \left(\left|X\left[y_a\right]\right| - 1\right) / 2} \tag{10}$$

By using the method of attribute value correlation analysis, multiple attribute values are spliced into one attribute value, and the related word analysis model of massive data in the Internet of Things is constructed. The problem of attribute value matching is modeled as a classification problem, which improves the dynamic cleaning ability of outliers[12].

3. Cleaning optimization of outliers in massive data

3.1. Massive data outlier filtering

On the basis of the above-mentioned model of abnormal data analysis and related word classification of massive data in the Internet of Things, an abnormal value filtering analysis model of massive data is constructed, and hierarchical clustering of data is realized by using the framework of evidence reasoning, and the probability of abnormal value distribution of massive data in the Internet of Things is detected according to the matching probability of related attribute values[13]. By using the method of conditional correlation factor analysis, the conditional correlation factor between abnormal value attribute value X and attribute value Y is obtained as follows:

$$ccf\left(X, y_a\right) = \frac{P\left(x_{ai} \approx x_{aj}\right)}{P\left(x_i \approx x_j\right)} \tag{11}$$

Where, $x_{ai}, x_{aj} \in X\left[y_a\right]$, $x_i, x_j \in X[R]$, describes the outlier component that measures the performance difference of candidate attributes. Given the target attribute Y and a candidate association attribute X, the conditional association density between the normal value and the outlier of massive data is obtained. Through the weighted calculation of $ccf\left(X, y_a\right)$:

$$ccf(X, Y) = \sum_{y_a \in \mathbf{Y}} \left(w_a \cdot ccf\left(X, y_a\right)\right) \tag{12}$$

Where, $y_a \in \mathbf{Y}$ is a differential attribute value, and the ordered even equivalent probability of outlier distribution of massive data is described as:

$$P\left(y_i \simeq y_j\right) = f\left(Y, X_1\right) \cdot P\left(x_{1i} = x_{1j}\right) \tag{13}$$

The performance difference expression of outliers in massive data is:

$$\left| f\left(Y, X_1\right) \cdot P\left(x_{1i} = x_{1j}\right) - f\left(Y, X_2\right) \cdot P\left(x_{2i} = x_{2j}\right) \right| \tag{14}$$

Normalize the above formula, use dynamic feature weighting method to match the equivalent attribute values of data outliers, and use evidential reasoning framework to realize hierarchical clustering of data[14], and get the normalized clustering model parameters as follows:

$$D_{\left(y_i, y_j\right)}\left(X_1, X_2\right) = \left| 1 - f\left(X_1, X_2\right) \cdot \frac{P\left(x_{2i} = x_{2j}\right)}{P\left(x_{1i} = x_{1j}\right)} \right| \tag{15}$$

The algorithm of generating attribute combination is used, and the hierarchical clustering algorithm is adopted to obtain the correlation factor, which is recorded as

$$f\left(X_1\left[y_i\right],X_2\left[y_j\right]\right)=\frac{P\left(x_{1i}=x_{1j}\right)}{P\left(x_{2i}=x_{2j}\right)} \tag{16}$$

Based on each selected attribute, the abnormal eigenvalue of massive data is estimated, and the logarithm of conditional correlation factor is obtained, and then the filtered output of abnormal value is:

$$D_{\left(y_i,y_j\right)}\left(X_1,X_2\right)=\left|1-\frac{f\left(X_1,X_2\right)}{f\left(X_1\left[y_i\right],X_2\left[y_j\right]\right)}\right| \tag{17}$$

Wherein, $f\left(X_1,X_2\right)$ represents the matching function of correlation attributes, and assigns a correlation factor to each element to realize the filtering of outliers in massive data.

3.2. Cleaning of massive data outliers

In outlier cleaning, it is necessary to characterize the aggregate word vector and evaluate the joint distribution, and get the formula for calculating the feature distribution of the aggregate word vector between the outliers of $\mathbf{D}\left(f_i,y_j\right)$, $f_i\in\mathbf{F}$, $y_j\in\mathbf{Y}$ massive data and the two target attribute values.

$$D_{y_j}=\sum_{r_k\in R\left[y_j\right]}v_k, \tag{18}$$

Wherein, v_k represents the word vector representation from the massive data combination $r_k\left[X\right]$, and the similarity of the corresponding aggregation feature distribution can be obtained by calculating the similarity of the aggregation word vector of the massive data outliers, which is described as:

$$cohesion\left(y_1,y_2\right)=\frac{D_{y_1}\cdot D_{y_2}}{\|D_{y_1}\|_2\cdot\|D_{y_2}\|_2} \tag{19}$$

Using standardized factor expression and feature reconstruction, the abnormal value distribution of massive data of the Internet of Things in the whole corpus is obtained, and through the aggregation layer, the vector of node features is formed:

$$U_{y_i}=\frac{D_{y_i}}{\|D_{y_i}\|_2} \tag{20}$$

Combined with the probability detection results of outlier distribution, the hierarchical clustering of data is realized based on the method of levenshtein and attribute value correlation analysis, and the outliers are cleaned according to the clustering results[15].

4. Experimental test

The hybrid programming of C++ and Matlab 7 is used to process the clustering algorithm of outlier information of massive data in the Internet of Things. The algorithm is tested on the DBLP data set, and VCA infers the equivalence of journal through the author attribute. In the first case, we adjust the data set size from 200k to 800k K. In Hadoop cloud platform, the structural model of the Internet of Things massive data abnormal value information database is constructed. Based on the similarity of set word vectors and the matching method of equivalent attribute values, the initial sample size is 1200, and the fuzzy coefficient m is set to 2. In this paper, the equipment status data of the Internet of Things monitoring platform generated by AGV in the automated terminal includes the information of the equipment's running state, power, running speed and running power at a certain moment, which belongs to typical Internet of Things data and is divided into 9 data sets.

Table 1: Data set of Internet of Things

Data sets	Data scale	Data size /GB
IoT1	231703	10.22
IoT2	216056	13.16
IoT3	309858	11.31
IoT4	205425	8.48
IoT5	237773	9.30
IoT6	380845	14.20
IoT7	224361	14.71
IoT8	243914	7.44
IoT9	322931	11.86

According to the above test data and environment, the abnormal values of massive data of the Internet of Things are cleaned, and the aggregation words of massive data of the Internet of Things are classified into multiple related data by combining the method of attribute value correlation analysis, as shown in Figure 2.

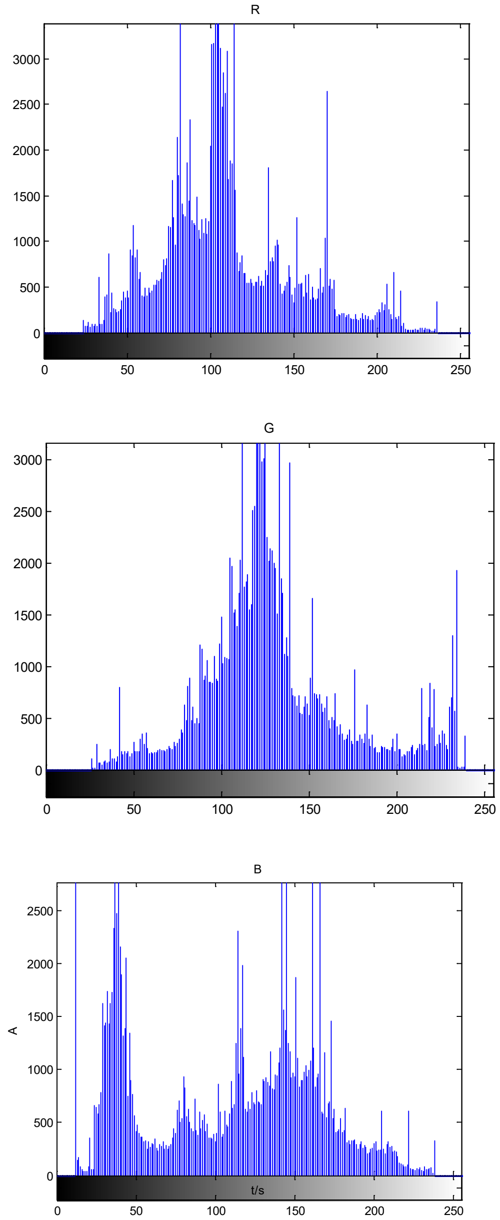

Figure 2: Internet of Things massive data classification

According to the classification result, hierarchical clustering of data is realized, and the clustering result is shown in Figure 3.

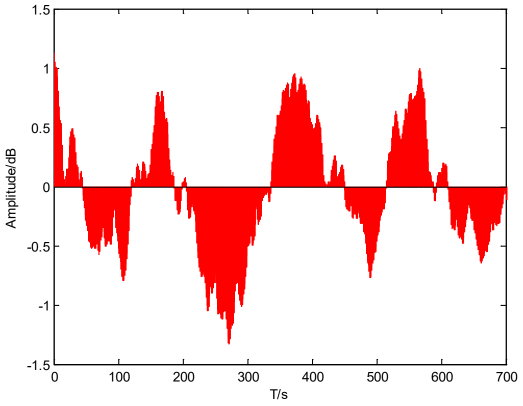

Figure 3: Data Hierarchical Clustering Results

According to the clustering result in Figure 3, outlier cleaning is realized, and the cleaning result output is shown in Figure 4.

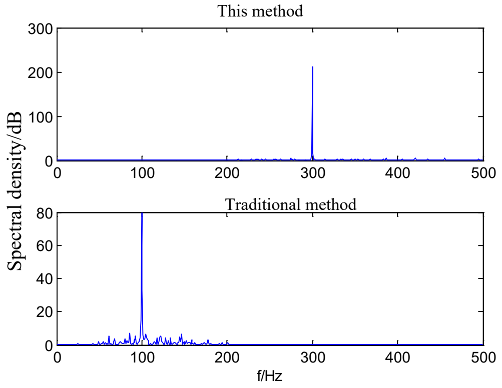

Figure 4: Data Outlier Cleaning Output

Analysis of Figure 4 shows that this method can effectively clean the outliers of massive data in the Internet of Things, and the feature spectrum clustering is good. The complexity of cleaning the outliers of massive data in the Internet of Things by different methods is tested, and the comparison results are shown in Figure 5.

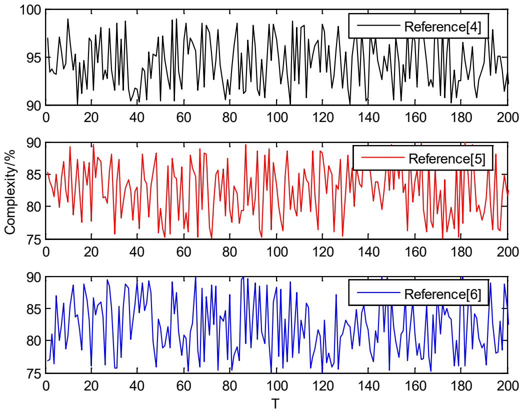

Figure 5: Performance comparison

By analyzing Figure 5, it is known that this method is used to clean the abnormal values of massive data in the Internet of Things, which improves the purity of data, reduces the interference of abnormal data and reduces the computational complexity.

5. Conclusions

In this paper, a method of cleaning outliers in massive data of Internet of Things based on hierarchical clustering algorithm is proposed. An unsupervised deep learning method is adopted to establish an analysis model of outlier characteristics of massive data in the Internet of Things, a penalty function is constructed by a sample generator, the data is encoded and mapped to a potential space again, a learning-based entity matching model is obtained, hierarchical clustering of data is realized by using an evidential reasoning framework, the probability of outlier distribution of massive data in the Internet of Things is detected according to the matching probability of related attribute values, and hierarchical clustering of data is realized by using an evidential reasoning framework. According to the matching probability of related attribute values, the probability of abnormal value distribution of massive data in the Internet of Things is detected, hierarchical clustering of data is realized based on the method of aggregate word vectors, and the abnormal values of data are cleaned according to the clustering results. The research shows that this method improves the clustering and secure storage access ability by detecting the anomaly of massive data in the Internet of Things.

References

[1] Zhang Binru. A deep learning approach for daily tourist flow forecasting with consumer search data [J]. Asia Pacific Journal of Tourism Research, 2020, 25(3): 323-339.

[2] Xuejian ZHAO, Hao LI, Haotian TANG. Recommendation rating prediction algorithm based on user interest concept lattice reduction[J]. Journal of Computer Applications, 2023, 43(11): 3340-3345.

[3] Zhuangzhuang XUE, Peng LI, Weibei FAN, Hongjun ZHANG, Fanshuo MENG. Multiple clustering algorithm based on dynamic weighted tensor distance[J]. Journal of Computer Applications, 2023, 43(11): 3449-3456.

[4] WANG J, WANG X, YU G, et al. Discovering multiple co-clusterings with matrix factorization[J]. IEEE Transactions on Cybernetics, 2021, 51(7): 3576- 3587.

[5] ZHOU Y, YU F R, CHEN J, et al. Cyber-physical-social systems: a state-of-the-art survey, challenges and opportunities[J]. IEEE Communications Surveys and Tutorials, 2020, 22(1): 389- 425.

[6] OU Q Y, ZHU E. Multi-kernel clustering algorithm based on compressed subspace alignment[J]. Computer Engineering and Science, 2021, 43(10): 1730- 1735.

[7] YAN J Z, CHEN H, LI Y. Improved fuzzy C-means clustering validity index[J]. Computer Engineering and Applications, 2020, 56(9): 156- 161.

[8] Xin Yu, Yang Jing, Tang Chuheng, Ge Siqiao. An Overlapping Semantic Community Detection Algorithm Based on Local Semantic Cluster. Journal of Computer Research and Development, 2015, 52(7): 1510-1521.

[9] WU Jiang, TANG Chang-jie , LI Taiyong, CUI Liang. Sentiment analysis on Web financial text based on semantic rules. Journal of Computer Applications, 2014, 34(2): 481-485.

[10] ZHANG Deng-yi, WU Wen-li, OUYANG Chu-fei. Approximating Query with Semantic-Based Measure on RDF Graphs. Chinese Journal of Electronics, 2015, 43(7): 1320-1328.

[11] XU Ying, ZENG Shuiling, WU Wenyuan. Complex Morphological Bidirectional Associative Memory Network and Its Performance Analysis[J]. Information and control, 2015, 44(3): 270-275.

[12] Xuewen LIU, Jikui WANG, Zhengguo YANG, et al. Imbalanced data classification algorithm based on ball cluster partitioning and undersampling with density peak optimization[J]. Journal of Computer Applications, 2022, 42(5): 1455-1463.

[13] LIU Q, ZHAI J W, ZHANG Z Z,et al. A survey on deep reinforcement learning[J]. Chinese Journal of Computers, 2018, 41(1):1-27.

[14] Zhao Yali, Yu Zhengtao, Guo Junjun, etc. Cross-language emotion classification model based on emotional semantic confrontation [J]. Computer Engineering and Science, 2023, 45 (02): 338-345.

[15] Xiangyu LUO, Ke YAN, Yan LU, Tian WANG, Gang XIN. Nonuniform time slicing method based on prediction of community variance [J]. Journal of Computer Applications, 2023, 43(11): 3457-3463.