Classification of GSM-R Network Operation Quality Evaluation Indicators Based on WDFCE Algorithm

Hang Zhang¹, Chaiyang Peng¹*, Liangde Ma², Yunru Zhao³ and Qing Liu³

¹School of Automation, Central South University, Changsha, Hunan 410083, China
²Infrastructure Inspection Research Institute, China Academy of Railway Sciences, Beijing 100081, China
³School of Computer Science and Engineering, Central South University, Changsha, Hunan 10083, China
Email: ⁰0904170313@csu.edu.cn

Abstract. This paper proposed a fuzzy clustering ensemble algorithm based on membership weighting and distance decision (WDFCE) and defined a method for determining the boundary threshold of indicators’ grade intervals. Base clusters and corresponding membership matrices were obtained by multiple fuzzy C-means (FCM) algorithm. Considering the reliability of each base cluster, the decision weighting method was introduced to weight the membership matrix and construct the cumulative distance matrix. The Density Peaks (DP) algorithm was used for clustering ensemble to obtain the final result. By defining the boundary zone, the boundary zone class density and the threshold possibility, the clustering result was combined to determine the grade interval of the indicators. In order to judge the advantages and disadvantages of the improved algorithm, the S_Dbw index with the best evaluation effect was selected from the various cluster validity evaluation indicators studied by domestic and foreign scholars. The results of the indicator S_Dbw for evaluating cluster validity indicated that the improved method we proposed is superior to the general clustering ensemble algorithm and it solved the problem of determining the boundary threshold of the indicator grade interval in the GSM-R evaluation system.

1. Introduction

GSM-R is a safe and efficient digital wireless communication system, which can provide plentiful services of data communication. In order to ensure the stable operation of GSM-R, it is necessary to make the GSM-R network run with good quality, and classify the operation quality can be better reflect the quality of GSM-R operation. At present, the GSM-R network QoS evaluation system, which is currently in effect in China is mostly refers to the European railway standards[1-2] or acquire the list of the span of testing indicators according to expert experience. However, the standard of grade division is in lack of a comparatively objective theoretical method to support. Therefore, this article does the research on how to divide the grade intervals of each indicator.

Many researchers used clustering algorithm to process and analyze the indicators’ data on the problem of quality evaluation. When Xing Xiaoqin[3] evaluates the quality of GSM-R communication, the K-means clustering algorithm is used to cluster the sample data. In this way, the cluster center of each indicator and the total number of each cluster can be obtained, which is used as the basis for the subsequent score calculation. Gao Kunlun et al.[4] put forward a kind of hierarchical evaluation system which bases on the grey clustering analysis to evaluate the network security situation. The network attack is divided into three levels: “Strong”, “Medium” and “Weak” by the grey clustering analysis. Classify each network attack so that the attacks can be qualified to calculate the value of the network...
security situation. Lin Siyu et al.[5] use gray clustering combined with rough set theory to evaluate the quality of service of wireless network. They classify the measuring and statistical results of key performance indicators for wireless network QoS to arrive at an assessment. On the improvement and research of clustering algorithms, Fei Bowen et al.[6] put forward the fuzzy clustering ensemble model based on distance decision. The fuzzy C-means (FCM) algorithm[7] is used to cluster the data samples multiple times in order to obtain \( m \) base clusters and \( m \) corresponding membership matrices. The accumulated distance matrix \( D \) is constructed by the distance decision method and the membership relationship. Then the density peaks algorithm is combined with the accumulated distance matrix to ensemble the base clusters to obtain the final clustering result. The algorithm has been verified by experiments to be better than the classical clustering ensemble models, but it does not consider the reliability of each base cluster. The results of clustering are susceptible to the negative impact of low-quality clustering, so the algorithm can still be improved. On the problem of determining the boundary threshold of indicators, this paper defines a kind of method based on the clustering results of indicators in the GSM-R network evaluation system.

The related indicators of GSM-R network service quality evaluation system lack objective and reasonable classification method and the existing clustering ensemble algorithms are improvable. Therefore, this paper proposes a fuzzy clustering ensemble algorithm based on membership weighting and distance decision (WDFCE) to divide the relevant indicators into grade intervals. Firstly, \( m \)-time fuzzy C-means (FCM) algorithm is performed on the sample dataset of the indicators, and \( m \) base clusters and \( m \) membership matrices are obtained. After that, measure the reliability and find the weight of each basic cluster, obtain the cumulative weighted membership matrix, and then get the final clustering result. Finally, the grade intervals of the indicators are determined according to the threshold possibility of each sample data.

The model of this paper considers the reliability of each base cluster based on the literature[8], uses the decision weighting method to find its weight and weights the membership. Through the Density Peaks algorithm and the clustering ensemble algorithm based on distance decision, a more objective and reasonable clustering result is obtained, and also defined a universally applicable method for dividing the boundary threshold of one-dimensional data and finally obtain the threshold of the indicators in the GSM-R evaluation system.

2. Fuzzy Clustering Ensemble Algorithm Based on Membership Weighting and Distance Decision (WDFCE) and Boundary Determination

2.1. Data Preprocessing

Sample data may have errors due to certain subjective and objective factors. Therefore, abnormal data culling of the data set is required to make the result more accurate. This article uses the classic Chauvenet Criterion to reject abnormal data.

In \( n \) measurements, if an error occurs less than half a time, it is rejected. That is, the confidence probability (Chauvenet Coefficient) \( \omega_n = 1 + 0.4 \ln(n) \). If \( |x_i - \bar{x}| > \omega_n S_x \), then \( x_i \) is rejected.

Where \( x_i \) is a sample data, \( \bar{x} \): the sample mean, \( S_x = \left( \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \right)^{\frac{1}{2}} \): the standard deviation of the sample.

That is, if the absolute value of the difference between a measured value and the average value is greater than the product of the standard deviation and the Chauvenet coefficient, the measured value is rejected.

2.2. Decision Weighting

Let \( \prod \) denote a set including \( M \) base clusters, written as: \( \prod = \{ \pi^1, \pi^2, \cdots, \pi^M \} \). Where \( \pi^m \) represents the \( m \)th base cluster in the cluster set \( \prod \).

Each base cluster is a clustering result of the data set \( X \). Each base cluster is obtained by the FCM algorithm, and includes several piles, which are recorded as: \( \pi^m = \{ C^m_1, C^m_2, \cdots, C^m_{C^m} \} \). Where \( C^m_i \) is the \( i \)th pile in the base cluster \( \pi^m \), \( C \): the number of piles in \( \pi^m \).
Every pile is a set of several data points. Represent the pile set of the whole base cluster as:
\[ c = \{ C_1, C_2, \cdots, C_n \} \]
Where \( C_i \) is the \( i \)th cluster in set \( c \), \( N_C = M \times C \): the total number of clusters in set \( c \).

In cluster ensemble, each base cluster can be seen as a set of several connection decisions. If \( x_i \) and \( x_j \) are divided into the same pile in the base cluster \( \pi^m \), then \( \pi^m \) makes a connection decision for \( x_i \) and \( x_j \). Thus, the number of connection decisions made by a pile \( C_i^m \in \pi^m \) is:
\[
\text{Decision}(C_i^m) = \frac{(|C_i^m| - 1)|C_i^m|}{2}
\]
(1) \(|C_i^m|\): the number of data points in \( C_i^m \).
Then the total number of connection decisions in the base cluster \( \pi^m \) is:
\[
\text{Decision}(\pi^m) = \sum_{i=1}^{C} \text{Decision}(C_i^m)
\]
(2)

Assign the credibility of a unit to each base cluster. Credibility is shared by all decisions within the base cluster. Then the reliability of each connection decision in the base cluster \( \pi^m \) is:
\[
\omega(\pi^m) = \frac{1}{\sum_{i=1}^{M} \text{Decision}(\pi^m)}
\]
(3)
The sum of the weights of all the base clusters is 1, that is:
\[
\sum_{i=1}^{M} \omega(\pi^m) = 1
\]
(4)

2.3. Construct Weighted Membership

After the base clustering is completed, the corresponding membership information matrix \( L_j \) is calculated by using the obtained membership matrix \( U_j \), \( j = 1, 2, \cdots, m \). That is:
\[
L_j = \{ l_{ij}^1, l_{ij}^2, \cdots, l_{ij}^C \} \quad \text{s.t.} \quad l_{ij}^l = \arg\max_{1 \leq k \leq C} u_{ik}
\]
(5)
Where \( l_{ij}^l \): the cluster number where the maximum membership is located, \( C \): the number of clusters.

The information matrix corresponding to the membership matrix \( U = \{ U_1, U_2, \cdots, U_m \} \) is \( L = \{ L_1, L_2, \cdots, L_m \} \). The distance matrix is constructed by taking a single membership matrix \( U_j \) and a corresponding information matrix \( L_j \) as an example. Take two numbers \( x_a, x_b \) from the data set \( X \), their membership matrices and corresponding information matrices are \((u_{ak}^j, l_{ak}^j), k = 1, 2, \cdots, C \) and \((u_{bk}^j, l_{bk}^j), k = 1, 2, \cdots, C \). Constructing the membership similarity matrix \( U_j^l \) from probability theory knowledge:
\[
U_j(x_a, x_b) = \begin{cases} 
\sum_{k=1}^{C} u_{ak}^j \times u_{bk}^j, & l_a^j = l_b^j \\
\sum_{k=1}^{C} (1 - u_{ak}^j) \times (1 - u_{bk}^j), & l_a^j \neq l_b^j 
\end{cases}
\]
(6)

Ensemble all membership matrices in the above way to obtain a accumulated membership similarity matrix \( U' = \{ \omega_1 \times U_1, \omega_2 \times U_2, \cdots, \omega_m \times U_m \} \), which is:
\[
U' = \sum_{j=1}^{m} \omega_j \times U_j^l
\]
(7)
Where \( \omega_j \) is the weight of the \( j \)th base cluster.

The accumulated membership similarity matrix represents the similarity between data samples. The greater the similarity between samples, the closer the distance between the two. Therefore the accumulated distance formula is:
\[
D = m \times (I - U')
\]
(8)
m: the number of base clusters, I: all 1 matrix.

2.4. Fuzzy Clustering Ensemble

The algorithm performs density peaks detection on the weighted membership of the base cluster. The top C data points with higher density were identified as the center of cluster. The specific process is: taking the accumulated distance matrix D obtained by the weighted membership as input data, and calculating the local density \( \rho_i \) and the distance \( \delta_i \) of the data sample to the higher density sample point. Assuming that the distance \( d_{ij} \) between the data sample points \( x_i \) and \( x_j \) is known, the local density \( \rho_i \) is:

\[
\rho_i = \sum_{j=1, j \neq i}^{n} \exp\left(\frac{-d_{ij}^2}{d_o^2}\right)
\]  

(9)

Where \( d_o (d_o > 0) \) represents the cutoff distance and needs to be specified by the user. The distance \( \delta_i \) of the data sample from the point of higher local density is:

\[
\delta_i = \begin{cases} 
\min_{j \in I_{k}^i} \{d_{ij}\}, & I_{k}^i \neq \emptyset \\
\max_{j \in I_{k}^i} \{d_{ij}\}, & I_{k}^i = \emptyset 
\end{cases}
\]  

(10)

Where \( I_{k}^i \) is the indicator set of the data set, which is specifically expressed as:

\[
I_{k}^i = \{k \in I_k: \rho_k > \rho_i\}
\]  

(11)

Therefore, for each data sample point \( x_i \) in the data set \( X \), the corresponding local density \( \rho_i \) and the distance \( \delta_i \) from the higher local density point can be obtained. Using the following formula:

\[
\gamma_i = \rho_i \times \delta_i
\]  

(12)

the density of each data sample point is obtained, and the first C data points with higher \( \gamma_i \) are selected as the cluster center. That is, when the data sample points have higher \( \rho_i \) and \( \delta_i \) at the same time, the data sample is regarded as the center point of the cluster. Finally, the remaining data points can be divided into clusters closest to it.

2.5. Determination of Boundary Threshold

**Definition 1** (Boundary Zone): Given data set \( X \) and distance radius \( d \), for \( \forall x_i \in X \), the boundary zone of the data sample point \( x_i \) is defined as follows: with \( x_i \) as the center line and two parallel lines with \( x_i \pm d \) as the boundary. The enclosed rectangular zone about \( x_i \) is called the boundary zone of \( x_i \).

**Definition 2** (Boundary zone class density): It is assumed that the data set \( X \) is divided into \( C \) clusters. For \( \forall x_i \in X \), the boundary zone of the data sample \( x_i \) contains the class \( C_k (k = 1, 2, \ldots, C; C \) is the number of clusters\). The total number of data sample points for each class is called the density \( \gamma_i \) of \( x_i \). The \( C_k \) class density of the data sample \( x_i \) is denoted as \( C_k \text{den}(x_i) \).

**Definition 3** (Threshold Possibility): A measure of the likelihood of a certain data sample \( x_i \) being a threshold. For \( \forall x_i \in X \), after calculating the density of each class of the boundary zone, select the first two classes \( C_{k1} \) and \( C_{k2} \) with higher density, then the threshold possibility of the data sample is:

\[
TP_{x_i} = \frac{C_{k1}\text{den}(x_i)}{C_{k2}\text{den}(x_i)}
\]  

(13)

The closer the value of \( TP_{x_i} \) is to 1, the more likely it is that \( x_i \) is the boundary threshold.

**Definition 4** (Boundary Threshold): A set of all boundary threshold possible points selected according to the threshold possibility within a certain range:

\[
BT = bt(T_1) \cup bt(T_2) \cup \cdots \cup bt(T_k)
\]  

(14)

Where \( k \) is the number of boundary threshold points or the number of clusters divided, \( T_k \): the value of each cluster’s boundary threshold, \( bt(T_k) \): the set of possible points of the boundary
threshold where the boundary value reaches a certain ratio ($r$) near the boundary value $T_k$, $BT$: a set of possible points for all boundary thresholds.

$$TP_{\text{threshold}} = 1 \ast r$$  \hspace{1cm} (15)

$r$ is the proportion of boundary division. The size of the boundary threshold point set can be changed by changing the value of $r$. If the number of points included in the boundary threshold point set is too large or too small, the accuracy of the boundary value will be affected.

In this paper, the boundary threshold of each class is obtained according to the above definition. For the data set $X$, the boundary zone of each data sample is obtained by the equation (16):

$$BZ_i = x_i \pm d$$  \hspace{1cm} (16)

Where $d$ is the distance set by the user, depending on the specific data set.

The plurality of data contained in the boundary zone may belong to any of the four classes of excellent, good, medium, and bad. Therefore the four classes of density for each sample data $x_i$ are:

$$C_k \text{den}(x_i) = n_{ik} \hspace{0.5cm} (k = 1, 2, 3, 4)$$  \hspace{1cm} (17)

Where $n_{ik}$ is the number of the data belonging to the class $k$ in the boundary zone $BZ_i$.

The first two classes with larger $n_{ik}$ in each boundary zone $BZ_i$ are selected as $C_1$ and $C_2$. Then the threshold of the sample data $x_i$ is:

$$TP_{xi} = \frac{C_1 \text{den}(x_i)}{C_2 \text{den}(x_i)}$$  \hspace{1cm} (18)

Set $TP_{\text{threshold}} = 1 \ast r$. When $TP_{xi} > TP_{\text{threshold}}$, $x_i$ may be the boundary threshold. Based on this, the boundary threshold point set $BT$ of the data set $X$ is obtained.

$BT$ includes multiple data that may be the boundaries of each cluster. Since each cluster center $c_k (k = 1, 2, 3, \cdots, C)$ has been obtained by cluster ensemble. Therefore, the set of left boundary threshold points of the class $c_k$ is:

$$BT_l = \{BT_i | c_{k-1} \leq BT_i \leq c_k \} \hspace{0.5cm} (k \geq 2)$$  \hspace{1cm} (19)

then the left boundary threshold of the class $c_k$ is $\text{avg}(BT_l)$.

Similarly, the set of right boundary threshold points of the class $c_k$ may be:

$$BT_r = \{BT_i | c_k \leq BT_i \leq c_{k+1} \} \hspace{0.5cm} (k \leq 3)$$  \hspace{1cm} (20)

then the right boundary threshold of the class $c_k$ is $\text{avg}(BT_r)$.

Therefore, the interval of the class $c_k$ is $(\text{avg}(BT_l), \text{avg}(BT_r))$. If $c_k$ is the "excellent" class, ie $k = 1$, the left boundary takes 0; if $c_k$ is the "poor" class ie $k = 4$, then the right boundary is taken $+\infty$.

2.6. Algorithm Flow

The whole algorithm flow is shown in Figure 1.
3. Results and Analysis

This paper selects two indicators of RxLevelUp and Transmission Delay in the GSM-R evaluation system as an application example of the algorithm of grade division and boundary determination. The RxLevelUp is derived from the Abis interface monitoring data of the Beijing-Guangzhou high-speed rail 99013 locomotive on September 5, 2018. The Transmission Delay is derived from the UDP test record table of the Beijing-Guangzhou high-speed railway Beijing West to Wuhan section on September 5, 2018. Considering that the above indicators are important and have a wide range of values, this paper selects the above two indicators to conduct experiments and determine the grade intervals.

3.1. Clustering Ensemble Result Analysis

In this paper, \( m = 4 \) is selected. That is, four FCM algorithms are used to find four base clusters and to ensemble the results. The weights of each base cluster are shown in Table 1.

|        | \( i = 1 \) | \( i = 2 \) | \( i = 3 \) | \( i = 4 \) |
|--------|------------|------------|------------|------------|
| RxLevelUp | 0.2405     | 0.2515     | 0.2540     | 0.2540     |
| Transmission Delay | 0.2393 | 0.2517 | 0.2645 | 0.2445 |

The clustering results obtained by running the FCM algorithm are different every time. Therefore, it is necessary to find the reliability of each base cluster, that is, the weight, and correct the base cluster, so that the final clustering result is more accurate and the credibility is more high.

Since the base cluster obtained by running the FCM algorithm is different every time, the weight obtained in the above table is only the result obtained by a certain running program, and the specific weight value is based on the result obtained by the user.

The four weighted base cluster are ensembled to get the center of the four clusters as shown in Table 2.

|        | Excellent | Good | Medium | Bad  |
|--------|-----------|------|--------|------|
| RxLevelUp(db) | -49 | -54  | -61    | -69  |
| Transmission Delay(ms) | 413 | 1155 | 2069 | 4686 |
The results obtained by grading the test data sets of the indicators are shown in Figure 2 and Figure 3.

![Figure 2] Clustering result of RxLevelUp

![Figure 3] Clustering result of Transmission delay

### 3.2. Cluster Model Validity Verification

This paper classifies the GSM-R evaluation indicators. The proposed method for determining the boundary threshold largely depends on the clustering algorithm that divides the one-dimensional data indicators. Considering the hierarchical fuzzy characteristics of GSM-R indicators and the limitations of a single algorithm, we propose an improved method on the basis of the fuzzy clustering ensemble model based on distance decision. The reliability of the basic clustering members is considered, and the clustering ensemble method by decision weighting is improved. We select two indicators and use the improved method to get the clustering results. Combining the study results of Liu Yanchi et al. [9] on 11 clustering results evaluation algorithms, we choose S_Dbw, which is the best clustering result evaluation algorithm, to verify the effectiveness of the clustering algorithm we proposed.

Maria Halkidi [10] proposed the S_Dbw algorithm to evaluate the validity of clustering results. According to her research results, the S_Dbw index is calculated for the clustering results to determine the optimal input parameter value. The research results show that the lower the value of the S_Dbw index, the better the clustering effect. The results obtained by comparing the distance-based clustering ensemble model with the clustering ensemble model in this paper are shown in Table 3.

| Table 3. Model validity comparison. |
|-------------------------------------|
| **RxLevelUp** | **Transmission Delay** |
| Model based on Distance | Model of this paper | Model based on Distance | Model of this paper |
| S_Dbw | 0.4224 | 0.3763 | 0.4721 | 0.3979 |

The smaller the S_Dbw index value is, the higher the effectiveness of the clustering ensemble model has. The experimental results in the above table show that the clustering ensemble model proposed in this paper is more effective than the Clustering ensemble model based on Distance.

### 3.3. Boundary Determination

According to the boundary determination method defined above, the boundary thresholds of the two indicators of RxLevelUp and Transmission Delay in the GSM-R evaluation system are shown in Table 4.
Table 4. Grade intervals of the indicators

|                | Excellent | Good  | Medium | Bad     |
|----------------|-----------|-------|--------|---------|
| RxLevelUp(db)  | (-51,0]   | (-58,-51] | (-64,-58] | (-∞,-64]|
| Transmission Delay (ms) | [0,784)  | [784,1612] | [1612,3372] | [3372,∞) |

Note: The boundary thresholds are rounded up according to standard practice.

4. Conclusion

We propose a boundary determination algorithm and a fuzzy clustering ensemble algorithm based on membership weighting and distance decision (WDFCE) to determine the grade interval of some indicators in the GSM-R network operation quality evaluation system.

In this paper, the FCM algorithm is used to cluster the data samples of each indicator to obtain multiple base clusters and corresponding membership matrices. The reliability of each base cluster is measured by decision weighting. We get the corresponding weights to weight the membership matrix. Then, transform the weighted membership matrix into an accumulated distance matrix using the distance decision in order to determine the distance between data samples. The distance information is introduced into the density peaks algorithm for cluster ensemble to obtain the final clustering result. We find the boundary thresholds by combining the final clustering result with the threshold possibility of the data. Finally, the grade interval is obtained.

In order to verify the performance of the clustering ensemble model proposed in this paper, we experimented with the test data set and compared the values of $S_Dbw$ of different algorithms. The experimental results show that the algorithm we proposed is more effective than other clustering ensemble models. Therefore, it can be proved that each indicator grade interval obtained in this paper has certain accuracy and practicality.

5. Acknowledgments

This work is supported by Research Project of Science and Technology Research and Development of China Railway Corporation (J2018G005).

6. References

[1] UIC SUBSET-093, ERTMS/ETCS-Class 1: GSM-R Interfaces Class 1 Requirements Version2.3.0, 2005.
[2] EEIG:04E117.ETCS/GSM-R Quality of Service-Operational Analysis, 2005.
[3] X.Q. Xing, Research on Key Technologies about Inspecting and Monitoring Data Analysis of GSM-R for High Speed Railway, China Academy of Railway Sciences, 2014.
[4] K.L. Gao, R.Z. Xu, Y.F. Wang, et al., A Study of Hierarchical Network Security Situation Evaluation System for Electric Power Enterprise based on Grey Clustering Analysis, International Conference on Computer Science and Service System, pp.1990-1995, 2011.
[5] S.Y. Lin, Z.D. Zhong and B. Ai, Quality of Service Evaluation Scheme for Heterogeneous Wireless Network, 2009 5th International Conference on Wireless Communications, Networking and Mobile Computing, pp.1-4, 2009.
[6] B.W. Fei, Y.F. Qiu, W.J. Liu, et al., Fuzzy Clustering Ensemble Model Based on Distance Decision, Journal of Electronics & Information Technology, vol.40, no.8, pp.1895-1903, 2018.
[7] J.C. Bezdek, R.J. Hathway, M.J. Sabin, et al., Convergence theory for fuzzy c-means: Counter-examples and repairs, IEEE Transaction on System, Man and Cybernetics, vol.17, no.5, pp.873-877, 1987.
[8] D. Huang, C.D. Wang, J.H. Lai, et al., Clustering ensemble by decision weighting, CAAI Transaction on Intelligent Systems, vol.11, no.3, pp.418-424, 2016.
[9] Y.C. Liu, Z.M. Li, H. Xiong, et al., Understanding and Enhancement of Internal Clustering Validation Measures, IEEE Transaction on Cybernetics, vol.43, no.3, pp.982-994, 2013.
[10] M. Halkidi, M. Vazirgiannis, A density-based cluster validity approach using multi-representatives, Pattern Recognition Letters, vol.29, no.6, pp.773-786, 2008.