

Glass Cultural Relic Classification Based on Random Forest and Robust K-Means Algorithm

Yifei Wang^{*}, Fuyu Jiang, Dongyang He

Electrical and Mechanical Engineering, Lanzhou Jiaotong University, Lanzhou, Gansu, China, 730000

^{*} Corresponding Author Email: 127230384464a@sina.com

Abstract. The ancient glass relics have weathered to a certain extent under the influence of the environment, and there is a certain degree of chemical exchange between the glass relics and the external environment, which interferes with the classification of the glass relics. In order to better study the classification of glass relics. In this paper, the random forest algorithm is firstly used to determine the characteristic importance of various chemical components in high potassium glass and lead barium glass, and some chemical substances are sorted out, and then their classification rules are analyzed according to the selected chemical substances. Then, the elbow method was used to determine the number of two kinds of glass subclasses into 3 categories, and then the robust K-Means algorithm was used to divide the subclasses, and the characteristics of each subclass were analyzed and studied. Finally, the standardized mutual information NMI value was used to analyze the sensitivity, and the logic theory and classification results were used to evaluate the model.

Keywords: Random Forest, Robust K-Means algorithm, Classification of glass.

1. Introduction

The Silk Road has been an important channel for material and cultural exchanges between the East and the West since ancient times, and glass was an important basis for early trade.

One of them. China absorbs foreign manufacturing techniques and produces locally similar glass with different ingredients [1].

Ancient glass uses quartz sand as the main raw material, in order to reduce its melting point, often add saltpeter, grass ash, lead iron ore as flux, add limestone as a stabilizer. The different fluxes lead to the different chemical composition of glass: lead-barium glass contains higher PbO and BaO; The formation of high potassium glass requires high potassium flux such as plant ash.

Ancient glass is weathered by conditions, and the exchange of internal and external elements of glass leads to changes in composition proportions and affects judgment. Existing ancient glass related data, workers have been divided into high potassium, lead and barium two types.

In this paper, we noticed that Qiu Yang [2], Huang Mei [3] and others used random forest to rank the data according to the importance characteristics and then carried out the research. Therefore, this paper referred to this method, calculated and ranked the importance of chemical content, so as to explore the classification law.

Yang Youjian et al. used kmeans algorithm to classify the rainstorm events in a certain place into six categories [4], but when dealing with data with noise points, we need an algorithm with better stability such as robust k-means [5].

2. Classification law of high potassium - lead barium glass

The data used in this paper include the weathering of cultural relics and the percentage of certain chemicals, as well as the color and ornamentation of cultural relics.

2.1. The basic fundamental of Random Forest

A random forest is a classifier that contains multiple decision trees whose output class is determined by the mode of the class output by the individual trees. Bagging algorithm is often used

in decision trees learning. Given a training set $X = x_1, \dots, x_n$ and the target $Y = y_1, \dots, y_n$. Use bagging to repeat (R times) samples from the training set with replacement, and then train the tree model on samples. After training, the prediction for an unknown sample x can be achieved by averaging the predictions of all the individual regression trees on x .

$$\hat{f} = \frac{1}{R} \sum_{n=1}^R f_b(x') \quad (1)$$

Using bagging reduces the variance without increasing the bias, however, it means that the decision tree is very sensitive to noise. But for multiple tree models, you don't need to worry about this as long as the numbers aren't correlated. The decision tree structure is shown in Figure 1.

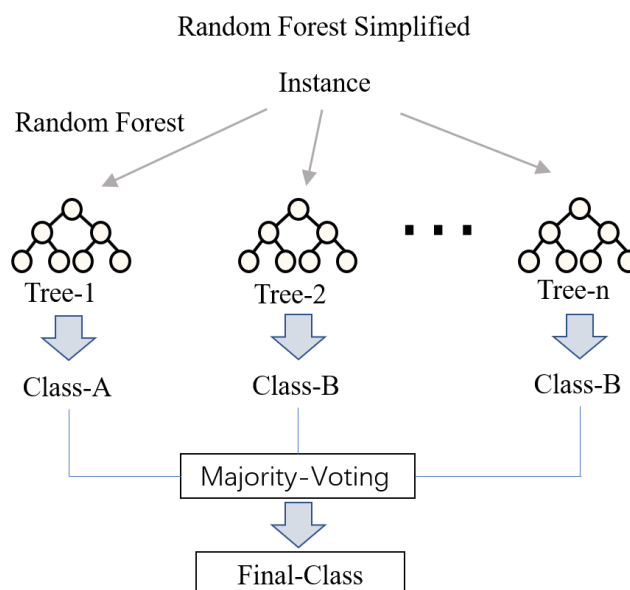


Figure 1. Decision tree structure diagram

2.2. Screening of classified substances

In this paper, the random forest model is used to obtain several substances with obvious effect on the classification of high potassium - lead barium glass [2-3].

In the random forest model, feature importance is an important part of machine learning, and its feature is an independent and observable attribute or characteristic of the observed object in the model. Feature importance also measures the contribution of each input feature to the model prediction, which can highlight the degree of correlation between the data feature and the target. The random forest training results are shown in Table 1.

Table 1. Model evaluation Results

	Accuracy rate	Rate of recall	Rate of accuracy	F1
Set of training	1.00	1.00	1.00	1.00
Cross validation set	0.96	0.96	0.94	0.95
Cross validation set	1.00	1.00	1.00	1.00

(1) Accuracy: Ratio of correct samples to total samples.

(2) Recall: The prediction is the ratio of the number of positive samples to the number of actually positive samples.

(3) Precision: The ratio of the number of actual positive samples to the number of predicted positive samples.

(4) F1: Harmonic mean of precision and recall. Recall and precision affect each other and restrict each other. In the actual process, the higher the recall rate, the lower the precision rate. The higher the precision, the lower the recall. In order to control them in a reasonable range, F1 can be used to measure.

The feature importance of each index is shown in Figure 2.

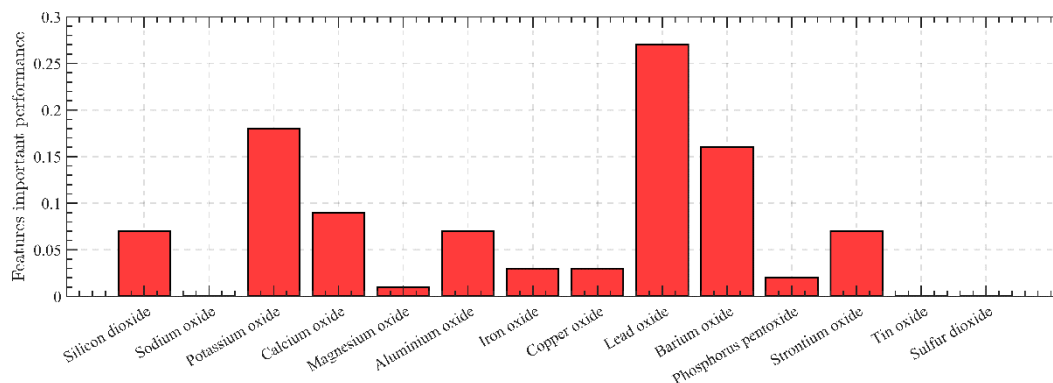


Figure 2. Importance of feature

2.3. Summarizing classification rules

The interval ranges corresponding to each index for high potassium glass and lead barium glass are shown in Table 2.

Table 2. Comparison of the range of each indicator

	Comparison of the range of each indicator	High potassium glass	Lead barium glass
		Range of interval	
Lead oxide	0.27	0-1.62	9.3-70.21
Potassium oxide	0.18	0-14.52	0-1.41
Barium oxide	0.16	0-2.86	0-35.45
Calcium oxide	0.09	0-8.70	0-6.40
Silicon dioxide	0.07	59.01-96.77	3.72-75.51
Aluminium oxide	0.07	0.81-11.15	0.45-13.65
Strontium oxide	0.07	0-0.12	0-1.12
Iron oxide	0.03	0-6.04	0-4.59
Copper oxide	0.03	0-5.09	0-10.57
Phosphorus pentoxide	0.02	0-4.5	0-14.13
Magnesium oxide	0.01	0-1.98	0-2.73
Sodium oxide	0.00	0-3.38	0-4.66
Tin oxide	0.00	0-2.36	0-1.31
Sulfur dioxide	0.00	0-0.47	0-15.95

Table 2 shows that the characteristics of lead oxide, potassium oxide and barium oxide are of high importance. It can be seen that the contents of these three chemical components have a great correlation with the classification results of high potassium glass and lead-barium glass. Calcium oxide, silicon dioxide, aluminum oxide and strontium oxide are less important than the first three chemical components, and have less influence on glass classification. The feature importance of the four chemical components of iron oxide, copper oxide, phosphorus pentoxide and magnesium chloride is very small, and the influence on the glass classification results is negligible. Sodium oxide, tin oxide and sulfur dioxide have almost no effect on the classification of glass.

Based on the analysis of the corresponding ranges of the main chemical components that affect the content of high potassium glass and lead-barium glass, the conclusions are as follows:

(1) Lead oxide: high potassium glass may contain less of this component, and lead-barium glass has a higher content of this chemical component, which is one of the main components of glass. Higher, one of the main components of glass.

(2) Potassium oxide: The content of this chemical in high potassium glass is generally higher than that in lead-barium glass.

(3) Barium oxide: The content of this chemical in lead-barium glass is generally higher than that in high potassium glass.

There are many factors that affect the classification of glass. In addition to the difference in the content range of various chemical components in glass, the range of weathering, decoration, color and total content is also an important basis for glass classification. The classification law is shown in Table 3.

Table 3. Classification rules of other indicators

Type of glass	Weathering or not	grain	color
High potassium glass	weathering	B	Blue, green
	undifferentiated	A, C	Blue green, light blue, dark blue
Lead barium glass	weathering	A, C	Black, blue-green, light blue, dark blue, light green, dark green, purple
	undifferentiated	A, C	Dark blue, light blue, dark green, light green, self, green

3. Rules for partitioning subclasses

In this paper, the Robust K-means algorithm is used for subclass partitioning.

3.1. Selection of partitioning metrics

This paper involves a total of 14 chemical indicators, but due to the small sample data of glass cultural relics, some chemical indicators play a minimal role in the classification, reducing the operation efficiency and accuracy of the algorithm, so it is necessary to select the index, and select the most effective discriminant index for classification.

Because of the existence of subclasses, the chemical composition of the main class changes in a large interval, the more subclasses, the larger the variation interval and the larger the fluctuation, on the contrary, the smaller the variation interval. The proportion of one or several chemicals before different subclasses is quite different, which is also the reason for the generation of subclasses, and subclasses tend to diversify the chemical composition of the main class.

Based on the above characteristics of the subclass, we select indicators. Usually, the indicators to measure the degree of data change include change interval, variance, standard deviation, etc.

A chemical composition of the main type of glass varies greatly, indicating that it may have an important role in dividing the sub-categories, but not absolutely, such as silica and other components, the overall proportion has always been very high, a slight change will produce a large change interval and variance, although the change interval and variance may be large, but the overall proportion of change may be small. On the contrary, some chemical components, such as magnesium oxide and lead oxide, have very small average values, and even if the proportion of variation is large, the variation interval and variance are still small. Since the proportion averages of different components are very different, which brings "unfairness" phenomenon, we believe that it is inappropriate to use the change interval or variance and standard deviation as the judgment index, and there are loopholes. If the proportion of each component is mapped in the same interval [6], this "unfair" phenomenon can be seen, and then the variance is used to measure the fluctuation of the data, and it is much more reasonable to use it as a discriminant index.

The metrics are selected as follows:

(1) Data mapping: the content ratio data of silicon dioxide, sodium oxide, potassium oxide, calcium oxide, aluminum oxide, iron oxide, copper oxide, lead oxide, barium oxide, phosphorus penoxide, strontium oxide, tin oxide and sulfur dioxide were mapped to the range of (0, 1) to weaken the "unfair" phenomenon.

(2) Variance calculation: After mapping, the variance of the data is calculated to obtain the square of the 14 chemicals of the high potassium glass. The differences were 0.1468, 0.1038, 0.1336, 0.1446, 0.1292, 0.0885, 0.0672, 0.0859, 0.1007, 0.0866, 0.0810, 0.1336, 0.0556, 0.1118; The variances of the 14 chemicals in lead-barium glass are: 0.0697, 0.0514, 0.0375, 0.0652, 0.0528, 0.0467, 0.0420, 0.0538, 0.0612, 0.0553, 0.0755, 0.0555, 0.0259, 0.0379.

(3) Index determination: select the largest variance of several indicators as the division index, the final high potassium glass selected silica, potassium oxide, calcium oxide, magnesium oxide, strontium oxide, sulfur dioxide a total of 6 division index, lead-barium glass selected silica, calcium oxide, lead oxide, barium oxide, phosphorus pentoxide, strontium oxide a total of 6 indicators.

3.2. Determination of the value of K

The K-Means algorithm is a very common and simple clustering algorithm that can be lightened briefly by setting a K value.

There are many ways to determine the K value, commonly used are the head pat method, elbow method, interval statistics method, contour coefficient method.

Canopy method et al [5], since the sample data is only 56 cases, there are not many subclasses under each main class, and the K value should be relatively small, so this paper uses the elbow method to determine [6].

When the value of K is less than the actual value, every time K increases by 1, the cost will be reduced by a large margin, and when the value of K is close to the real value, every time K increases by 1, the cost will not change significantly, so the turning point is the actual k value, similar to the elbow.

Elbow Method calculation formula:

$$D_k = \sum_{i=1}^K \sum dist(x, c_i)^2 \quad (2)$$

Using sample data, the elbow figure 3 is as follows:

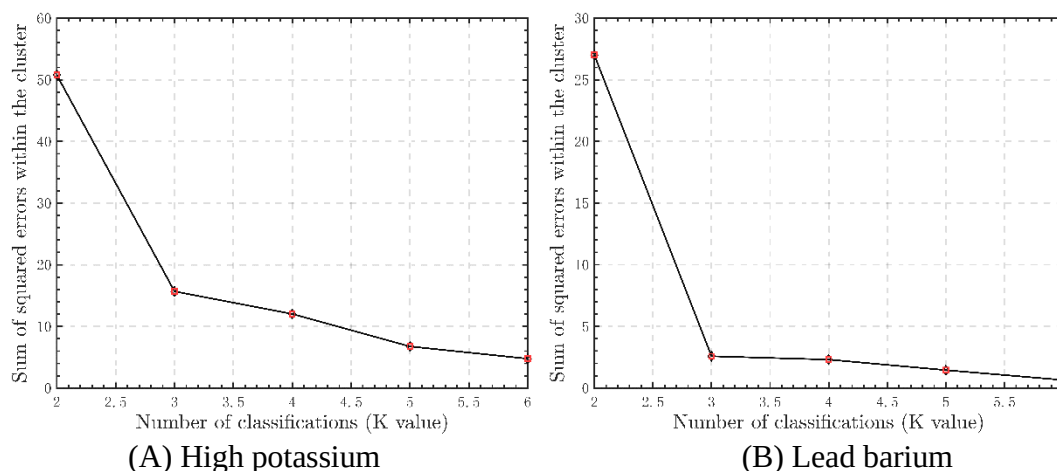


Figure 3. Elbow Method

It is not difficult to see that whether it is high-potassium glass or lead-barium glass, the inflection point is K=3, so high-potassium glass and lead-barium glass.

There are 3 subclasses each.

3.3. Robust K-Means classification

Algorithm steps [7]:

(1) K value is selected as the initial cluster center. Elbow method is adopted in this paper, no matter it is high potassium glass or lead barium glass Glaze, its K is 3;

(2) According to the mean of the object set in the cluster, let each object be assigned to the most similar cluster, due to the high potassium glass and the number of lead-barium glass samples is different, so the number of objects in the object set is different.

(3) The mean of cluster was updated and the mean of each object in cluster was calculated.

$$\bar{x}_i = \sum_{x \in D_i} \frac{x}{|D_i|} \quad (3)$$

Here D_i is the data set in the cluster.

(4) Loop continuously until the target function, namely: There is no significant change.

Traditional K-Means algorithm, Max-Min initialization often selects the boundary or outlier (outlier) value as. In the beginning, the cluster center and density cluster center can not be well reflected, so it is improved [7]. The principle is as follows:

For the data set G , which contains n objects, take the mean of all the objects and the closest object as the initial is the cluster center m_1 .

For other objects, the minimum distance between them and the cluster center can be calculated as follows:

$$\min dist(n_p) = \min \{\|n_p - m_s\|^2\} \quad (4)$$

In the formula, n_p represents other objects, that is, the object that is not selected as the center of the cluster, while m_s represents the selected cluster heart.

Other objects with the maximum and minimum distance will be used as the next batch of centers,

$$m_i = \operatorname{argmax}\{\min dist(n_p)\} \quad (5)$$

Where m_i represents the temporary cluster center.

Make $\varepsilon = \frac{1}{2} \times \min dist(m_i)$, and then look for all objects whose distance from m_i is less than ε ,

$$N_\varepsilon(m_i) = \{n_p: \|n_p - m_i\|^2 \leq \varepsilon\} \quad (6)$$

Select the object closest to the mean of $N_\varepsilon(m_i)$, update the cluster center of the algorithm,

$$m_i = \operatorname{argmin}\{\|n_p - \overline{N_\varepsilon(m_i)}\|^2\} \quad (7)$$

If m_i does not change at this time, proceed to the next step; otherwise, return to (4) until it does not change.

Continue to cycle (3) - (5) until $k = 3$ clustering centers are selected.

The final subclass division results are shown in Table 4.

Table 4. Comparison of the range of each indicator

	subclass	Heritage Number	Characteristics of cultural relics
High potassium glass	G1	11, 3, 5, 6, 18	High potassium glass: magnesium oxide content is the highest, the larger may contain strontium oxide, calcium oxide and silicon dioxide content changes greatly
	G2	12, 1, 4, 13, 15, 16	High potassium glass: calcium oxide content is the highest, silica content is stable
	G3	7, 8, 10, 12, 22, 27 etc	High potassium glass: silica content is the highest, potassium oxide, calcium oxide, magnesium oxide content is little, no strontium oxide
Lead barium glass	B1	34, 36, 38, 48, 56 etc	In lead-barium glass, the contents of phosphorus pentoxide and strontium oxide are the lowest, while the contents of lead oxide and barium oxide fluctuate greatly
	B2	2, 11, 19, 39, 40 etc	Lead barium glass: lead oxide content is the highest, silica content is stable
	B3	8, 26, 24	The contents of barium oxide and strontium oxide are the highest, while the contents of silica and calcium oxide are the lowest

3.4. Robustness evaluation

This model uses the normalized mutual information [9] to evaluate the robustness, which is formulated as

$$NMI(R, Q) = \frac{\sum_{i=1}^I \sum_{j=1}^J P(i, j) \log \frac{P(i, j)}{P(i)P(j)}}{\sqrt{H(R)H(Q)}} \quad (8)$$

Where R and Q denote two intervals, two partitions of the input data set, including the probability that cluster I and cluster J. $P(i)$ are assigned to cluster R_i from a randomly selected sample in the data set, and $P(i, j)$ is the probability that the sample belongs to both cluster R_i and cluster Q_j . $H(R)$ is the entropy associated with all probabilities $P(i)$ ($1 \leq i \leq I$) in partition R.

The larger the normalized mutual information $NMI(R, Q)$ value is, the more accurate and stable the division between clusters is, which indicates that the clustering effect of the algorithm is better and the robustness of the model is better.

The robustness is evaluated by adding uniformly distributed noise to different proportions of the data, where the abscissa is the proportion of added noise and the ordinate is MNI, as shown in figure 4.

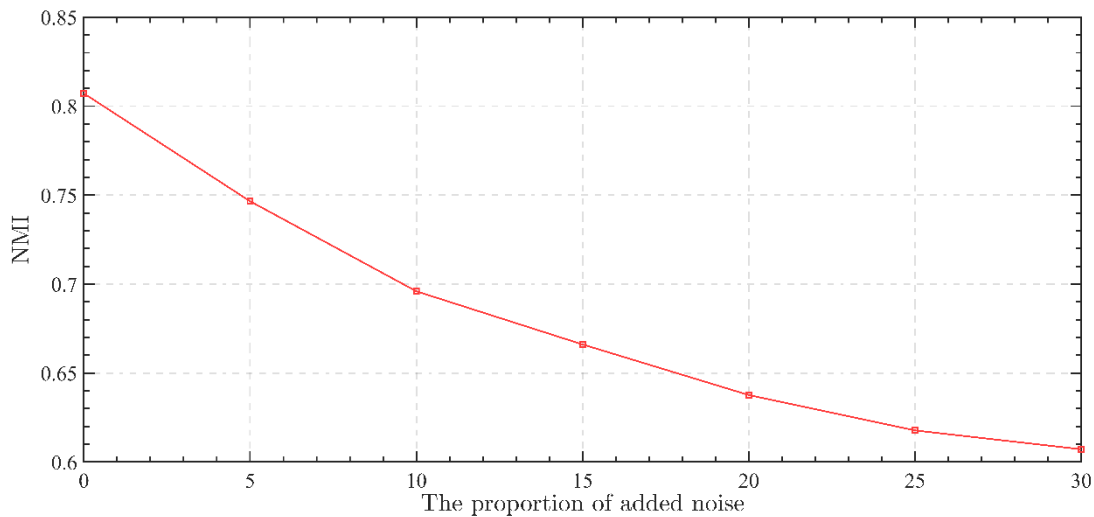


Figure 4. Evaluation of robustness

The NMI value is always above 0.6, and the overall robustness of the model is good and stable.

3.5. Assessment of reasonableness

(1) Mathematical analysis:

In the aspect of index selection, the index is backward based on the purpose. Team members conducted thorough analysis, mining, and analysis of the data.

Backward inference, assuming there is already a subclass, the influence of the subclass on the main component is analyzed, and the resulting influence can be reflected by backward inference. The "unfair" phenomenon of the initial data was found, and the data was mapped. Reduce the "unfair" phenomenon, and finally select the appropriate indicator, the indicator selection is reasonable and logical.

(2) In terms of algorithm selection, this paper does not directly use ordinary K-Means algorithm, traditional K-Means algorithm it is too "arbitrary", either one or the other, the sample size of this problem is small, it is difficult to protect if the ordinary K-Means algorithm is used directly. The stability, accuracy and robustness of the classification of subclasses are poor, so we refer to the existing scientific research results and use robust K-Means. The elbow method is used to determine the K value, which improves the accuracy of subclass classification to a certain extent.

4. Conclusion

To classify glass on the Silk Roads [10], in this paper, a robust K-Means subclass division model is established. The height is first determined by a random forest algorithm. The characteristic importance of potassium glass and lead barium glass is combined with form data for classification, statistics, comparison, and oxygen. Lead chemical, potassium oxide and other important indicators with high feature importance are carefully analyzed to obtain statistical laws, and then the backward method is used.

Reasonable indicators are divided by the definition and characteristics of subclasses, and then the elbow method is used to determine the high potassium glass and lead barium glass. The number of subclasses (k values) is all 3, and a robust K-Means algorithm is established to divide the subclasses and analyze the characteristics of 6 subclasses. Finally, the robustness assessment is used to analyze the sensitivity, and the rationality is proved by logical theory and division result analysis.

References

- [1] Cao Shiyuan, Feng Bailing, Wen Rui. Preliminary study on the influence of the Sheri Buried in the rise and Fall of Domestic glassware in Song and Yuan Dynasties [J]. Luoyang Archaeology, 2019, 0(1): 54-61.
- [2] Qiu Yang, Li Sheng, Jin Liang, Zhang Mimi, Wang Jie. Bridge Anomaly Monitoring Data Recognition Method Based on Statistical feature Mixing and Random forest importance Ranking [J]. Chinese Journal of Sensors and Actuators, 202, 35(06): 756-762.
- [3] Huang Mei, Zhu Yan. K-anonymous Feature Optimization Based on the Importance of Random Forest Features [J]. Computer Applications and Software, 20, 37(03): 266-270.
- [4] Yang Youjian, Guo Lin, Lin Leman, Bai Zhixu. Study on Spatio-temporal Evolution characteristics of Typhoon Rainstorm in Wenzhou City based on K-means rain classification method [J]. Water Resources and Hydropower Technology (English and Chinese), 2002, 53(04): 55-70. (In Chinese)
- [5] ZHANG Dongmei, Li Min, Xu Dachhuan, ZHANG Zhenning. Theory and algorithm review of K-means problem [J]. Science in China: Mathematics, 2020, 50(09): 1387-1404.
- [6] Yang Hanyu, Zhao Xiaoyong, Wang Lei. Review of Data normalization methods [J/OL]. Computer Engineering and Applications: 1-11 [2022-11-13].
- [7] Wang Jianren, Ma Xin, Duan Ganglong. Improved K-means Clustering K-value Selection Algorithm [J]. Computer Engineering and Applications, 2019, 55(08): 27-33. (In Chinese)
- [8] Li Qian, Wang Enwei, Lei Jingsheng, Song Shuo. A robust k-means Algorithm for Power User Segmentation [J]. Journal of Guilin University of Electronic Technology, 2017, 37(05): 406-410.
- [9] Leon Danon, Albert Diaz-Guilera, Alex Arenas. The effect of size heterogeneity on community identification in complex networks [J]. Journal of Statistical Mechanics: Theory and Experiment, 2006, 2006(11).
- [10] Gan Fuxi. Ancient Silk Road and ancient Chinese glass [J]. Chinese Journal of Nature, 2006(05): 253-260+246.)