A classification model for ancient glass artefacts based on Fisher discriminant analysis and K-means clustering

Peng Shi [1,#,*], Zongzi Long [2,#]

[1] School of Anesthesia Xuzhou Medical University, Xuzhou China, 221004

[2.]School of Dentistry, Xuzhou Medical, University, Xuzhou, China, 221004

* Corresponding Author Email: pengshi619@gmail.com

[#]These authors contributed equally.

Abstract. Classification is important for the study of ancient glass objects. Accurate identification and classification is important in presenting the cultural origins, heritage, and changes of glass artefacts. In order to accurately classify ancient glass artefacts, this paper first classifies ancient glass artefacts into two categories based on Fisher discriminant analysis, establishes a K-mean clustering model to subdivide the two types of artefacts, and further conducts sensitivity analysis to establish a classification model for ancient glass artefacts.

Keywords: Fisher discriminant analysis, K-means clustering, Sensitivity analysis.

1. Introduction

China's glass culture is vast and profound, and ancient glass is highly susceptible to weathering by the environment in which it is buried. During the weathering process, the internal elements of the glass were exchanged in large quantities with those of the environment, resulting in changes in the proportions of the chemical composition it contained and in its apparent appearance, thus affecting the correct judgement of its category [1-3].

The accurate identification and classification of ancient glass artefacts is therefore of great use in the exploration of the category of glass artefacts, the date of excavation and even the place of attribution [4]. However, existing methods of classifying glass artefacts are relatively few and far between, and accuracy and efficiency cannot be guaranteed. To address these issues and given that the fundamental difference between each glass artefact depends on the chemical composition, this paper uses the chemical composition of the artefact as the independent variable and the type of glass artefact as the dependent variable to show the classification pattern of glass artefacts using Fisher's discriminant analysis. The number of K-means clustering centroids is adjusted and the values of the clustering evaluation criteria SC, DB and CH are calculated separately to analyses the sensitivity of the classification model by comparing their values.

2. Modeling

2.1. Fisher discriminant analysis

Fisher's discriminant function is to find several directions in the original sample space, and after projecting the samples into that direction, establish a discriminant criterion to separate the original samples using the distance discriminant method. The distance between the projections of samples in each class in that direction is as far as possible, while the projections of samples within each class are as close as possible, so that the classification effect is optimal, namely maximizing the inter-class distance while minimizing the intra-class distance [5]. At this point, the samples have the closest separability in that space [6].

Let the overall $G_i(i=1,2,\cdots,k)$ be a sample of glass artefacts, and from these k samples construct the Fisher discriminant function with m different chemical compositions X_i as independent variables.

$$T = u_1 X_1 + u_2 X_2 + \cdots + u_m X_m = \boldsymbol{u}' X \tag{1}$$

Among them, u_i is the discriminant function coefficient of m indicators; X is the set of m chemical composition indicators; $\boldsymbol{u}$ is the discriminant function feature vector.

In order to maximize the interclass dispersion and minimize the intraclass dispersion of the samples whose discriminant function eigenvectors can be projected in this projection space, the algorithm is as follows.

Let the G_i mean and covariance matrices be **μ**(i) and **C**(i), respectively. Under the condition that $X \in G_i$, calculate the expected difference $E(\boldsymbol{u}'X)$ and the variance $D(\boldsymbol{u}'X)$, respectively.

$$\begin{gathered} E(\boldsymbol{u}'X \mid G_i) = \boldsymbol{u}' E(X \mid G_i) = \boldsymbol{u}'\boldsymbol{\mu}(i) \\ D(\boldsymbol{u}'X \mid G_i) = \boldsymbol{u}' D(X \mid G_i)\boldsymbol{u} = \boldsymbol{u}'\boldsymbol{C}(i)\boldsymbol{u} \\ D_b = \sum_{i=1}^{k} \left(\boldsymbol{u}'\boldsymbol{\mu}(i) - \boldsymbol{u}'\overline{\boldsymbol{\mu}}\right)^2 \\ D_e = \sum_{i=1}^{k} \boldsymbol{u}'\boldsymbol{C}(i)\boldsymbol{u} = \boldsymbol{u}'\left(\sum_{i=1}^{k} \boldsymbol{C}(i)\right)\boldsymbol{u} = \boldsymbol{u}' E \boldsymbol{u} \end{gathered} \tag{2}$$

Where, D_b is the between-group variance in the one-way ANOVA, D_e is the within-group variance and $\overline{\boldsymbol{\mu}}$ is the mean vector. $\overline{\boldsymbol{\mu}} = \frac{1}{k}\sum_{i=1}^{k} \boldsymbol{\mu}(i)$

By applying the idea of ANOVA, Fisher's discriminant function is obtained by determining $\boldsymbol{u}$ so that $\frac{D_b}{D_e}$ reaches a maximum. When there is only one discriminant function, the m chemical composition indicators of the sample Gi are brought into the discriminant function to determine whether the glass artefact belongs to high potassium glass or lead-barium glass.

To examine the goodness of the discriminant criterion, a backgeneration estimation method based on the training samples was used to calculate the misclassification rate. All the training set samples are brought into the discriminant equation in turn, and the discriminant criterion is used to back judge. The misclassification rate is denoted as η.

$$\eta = \frac{N}{n_1 + n_2 + \cdots + n_p} \tag{3}$$

The number of misclassified samples is N, where $n_1, n_2, \cdots, n_p$ is the capacity of the p training set of glass artefact samples.

2.2. K-mean clustering model building and solution

The K-mean clustering is a division-based clustering method, the basic idea is to cluster K points in space as the center, and group the objects closest to them according to the distance between the samples and the center point, while updating each clustering center one by one through an iterative method until the best clustering result is obtained [7].

Assuming that the sample set is to be divided into K categories, the algorithm proceeds as follows:

STEP1: The initial centers of the K categories are selected appropriately, generally at random initially.

STEP2: In each iteration, for any sample of glass artefacts, find the Euclidean distance to each of the K centers and assign that sample to the class in which the center with the shortest distance is located.

STEP3: Update the centers of these K categories using the mean value method.

STEP4: Repeat steps 2 and 3 for K clustering centers. If the distance travelled by the centers of a category satisfies certain conditions, the iteration ends and the classification is completed.

2.3. Sensitivity analysis

It is significant to evaluate the clustering results of different clustering numbers to confirm whether the clustering numbers in this paper are a fair choice among different clustering numbers. Also, as the initial clustering centers are chosen randomly, the calculation process of the K-means clustering algorithm can be somewhat random, which may affect the final clustering results. If the clustering results show considerable fluctuations when the calculation is repeated under the same conditions, this means that the dataset is not suitable for the application of the K-means clustering algorithm. Therefore, this paper uses three metrics Calinski Harabasz (CH), Silhouette coefficient (SC) and Davies Bouldin (DB) to evaluate the performance of the K-means clustering algorithm [8].

The SC metric calculates dissimilarity within a class by calculating the average distance from sample i to other samples in the same class, and dissimilarity between sample i and out-of-class by calculating the minimum of the average distance from sample i to all other samples in the class. The SC of a sample i is obtained by subtracting the out-of-class dissimilarity from the ratio of the out-of-class dissimilarity to the greater of the in-class dissimilarity and the in-class dissimilarity. The contour coefficient of the clustering result takes a value between [-1,1], the larger the value, the closer the similar samples are to each other and the further the different samples are from each other, the better the classification result is.

$$\mathrm{S}=\frac{(b-a)}{\max(a,b)} \tag{4}$$

In this case, a is the average of the dissimilarity of sample i to other points within the same class and b is the minimum of the average dissimilarity of sample i to other classes.

DB is then the sum of the average intra-category sample distances to the class centres of any two categories, divided by the distance between the centroids of the two categories, taking the maximum value. A smaller DB thus means a smaller distance within a category, while a larger distance between categories gives a better classification result [9].

$$\mathrm{DB}=\frac{1}{k}\sum_{i=1}^{k}\max_{i\neq j,i,j\in[1,K]}\frac{s_i+s_j}{M_{ij}} \tag{5}$$

In this case, si denotes the dispersion of sample points in the class and M_{ij} is the distance between class i and the centre of class j.

The CH metric measures the tightness within a category by calculating the sum of the squares of the distances between each glass artefact sample in the class and the class centre, in addition to measuring the separation of the dataset by calculating the sum of the squares of the distances between the centroids of each category and the centroids of the whole dataset [10]. Further, the CH metric is obtained from the ratio of separateness to tightness. Thereby, a larger CH represents a tighter class in itself and a more discrete class from class to class.CH is specified by the following formula.

$$\mathrm{CH}=\frac{SSB}{SSW}\frac{m-k}{k-1} \tag{6}$$

Here, $SSW=\sum_{i=1}^{m}|x_i-C_{pi}|^2$,$SSB=\sum_{j=1}^{k}n_j\ |C_j-\overline{X}|^2$,m is the number of training set samples and k is the number of categories. Cpi and C_j are the class centroids for each category and are the centroids of the entire heritage sample dataset.

2.4. Analysis of experimental results

By selecting a total of 67 lead-barium glass and high-potassium glass artefacts, Fisher's discriminant function was established as follows.

$$\begin{aligned}T=&0.050X1+0.485X2-0.0643X3-0.475X4+0.709X5+0.363X6+0.355X7\\&-0.115X8+0.141X9+0.195X10+0.267X11+0.00147X12+2.286X13+0.663X14\end{aligned} \tag{7}$$

Table 1. Results of the test set sample backband discriminant function

	high potassium glass artefacts	lead-barium glass artefacts
Identified as high potassium glass (1)	11	1
Identified as barium lead glass (2)	0	28
misclassifications	0	1

From the results of the discriminant function in Table 1, it can be concluded that the discriminant accuracy of the 40 high potassium glass artefacts was 100%, while the discriminant accuracy of the lead-barium glass artefacts was 96.6%. Bringing all 27 training samples back into the discriminant function yielded a false positive rate of $\eta = 0$, indicating that the Fisher discriminant function is able to respond well to the classification pattern of high potassium glass artefacts and lead-barium glass artefacts based on the chemical composition content of the test points of the artefact samples.

The results of the k-means clustering analysis are as follows.

Table 2. Analysis of K-mean clustering field sex differences for high potassium glass artefacts

	Clustering categories (mean ± standard deviation)		F	P
	Category 2 (n=9)	Category 1 (n=9)		
SiO_2	63.624±3.558	89.663±7.119	96.333	0.000***
Na_2O	0.927±1.427	0.0±0.0	3.796	0.069*
SrO	0.048±0.051	0.008±0.023	4.516	0.050**
O_2	0.136±0.205	0.0±0.0	3.923	0.065*
SnO_2	0.0±0.0	0.262±0.787	1	0.332
BaO	0.579±1.001	0.219±0.657	0.814	0.380
K_2O	10.818±2.37	1.986±3.22	43.917	0.000***
P_2O_5	1.523±1.652	0.533±0.451	3.007	0.102
CaO	6.363±2.64	1.327±1.419	25.408	0.000***
MgO	1.133±0.672	0.437±0.593	5.435	0.033**
PbO	0.41±0.64	0.139±0.333	1.271	0.276
Fe_2O_3	2.312±1.643	0.44±0.735	9.737	0.007***
CuO	2.819±1.565	1.492±1.136	4.233	0.056*
Al_2O_3	7.349±2.346	2.764±1.67	22.809	0.000***

Note: ***, **, * represent 1%, 5%, 10% level of significance respectively

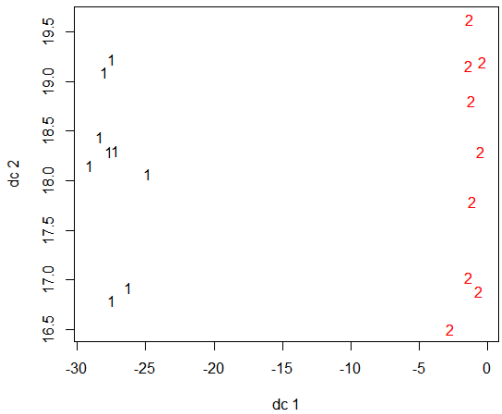

Figure 1. K-mean clustering of high potassium glass artefacts

The results of the K-mean clustering of high potassium glasses are shown in Figure 1 and Table 2. It can be seen that silica, cesium oxide, potassium oxide, calcium oxide, magnesium oxide, iron oxide and aluminum oxide are significantly different between the categories classified by the K-mean clustering analysis. The clustering results for the high potassium glass artefacts are mainly influenced by the level of silica, potassium oxide, calcium oxide and aluminum oxide content. There are nine high potassium glasses in Class 1, which are glass artefacts with a high silica content and a relatively low content of other chemical components, and nine high potassium glasses in Class 2, which are glass artefacts with a low silica content and a high content of other chemical components, and these glasses can be named polymetallic high potassium glasses.

Table 3. Analysis of K-mean clustering field ability differences for lead barium glass artefacts

	Clustering categories (mean ± standard deviation)		**F**	**P**
	Category 2 (n=28)	Category 1 (n=21)		
SiO_2	24.915±9.204	57.49±9.133	151.314	0.000***
SrO	0.442±0.269	0.223±0.2	9.803	0.003***
Na_2O	0.172±0.527	1.881±2.4	13.423	0.001***
SO_2	1.269±4.063	0.174±0.799	1.473	0.231
SnO_2	0.047±0.138	0.073±0.288	0.183	0.671
BaO	12.377±9.949	7.975±4.621	3.526	0.067*
K_2O	0.163±0.336	0.188±0.172	0.095	0.759
P_2O_5	4.916±4.253	1.128±1.909	14.425	0.000***
PbO	43.448±10.93	19.884±6.462	77.121	0.000***
Fe_2O_3	0.68±0.87	0.624±1.065	0.04	0.842
CuO	2.415±2.992	1.165±1.273	3.215	0.079*
CaO	2.731±1.715	1.142±0.964	14.522	0.000***
Al_2O_3	2.62±1.498	5.064±3.888	9.278	0.004***
MgO	0.602±0.671	0.704±0.581	0.31	0.580

Note: ***, **, * represent 1%, 5%, 10% level of significance respectively

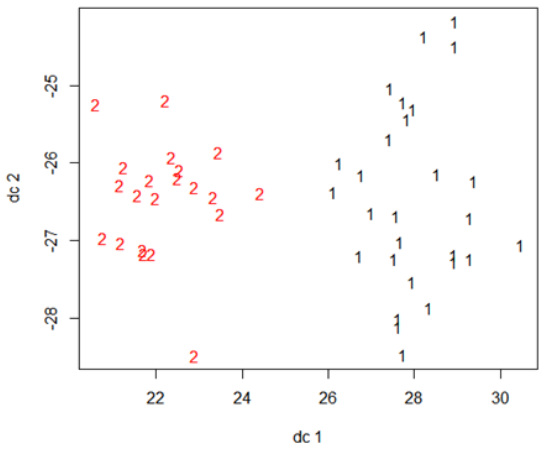

Figure 2. K-mean clustering of lead barium glass artefacts

The results of the K-mean clustering of lead and barium glass are shown in Figure 2 and Table 3. It can be seen that silica, cesium oxide, sodium oxide, phosphorus pentoxide, lead oxide, calcium oxide and aluminum oxide are significantly different between the categories classified by the K-mean clustering analysis. The clustering results for the lead-barium glass artefacts are mainly influenced by the level of chemical composition such as silica, potassium oxide, phosphorus pentoxide and lead oxide. Among them, there are 21 Class 1 lead-barium glasses, which belong to glass relics with high silica and sodium oxide content and low lead oxide and other chemical components, and this type of lead-barium glass can be named standard lead-barium glass; there are 28 Class 2 lead-barium glasses,

which belong to glass relics with particularly low silica content, particularly high lead oxide content and high other chemical components, and this type of lead-barium glass can be named high lead-barium glass.

In summary, combined with the K-means clustering analysis, this paper classifies high potassium glasses into two subclasses: high silica high potassium glasses and polymetallic high potassium glasses by combining the content of silica, potassium oxide, calcium oxide and aluminum oxide, and lead barium glasses into two subclasses: standard lead barium glasses and high lead barium glasses by combining the content of silica, potassium oxide, phosphorus pentoxide and lead oxide.

Based on the above theory, this paper varies the number of classification results for K-mean clustering and analyses the sensitivity to changes in system parameters or surrounding conditions i.e. changes in the three evaluation criteria as follows.

Table 4. Sensitivity analysis of classification results for high potassium glass

K	SC	DB	CH
2	0.651	0.444	378.170
3	0.572	0.639	177.267
4	0.374	1.010	110.300
5	0.357	0.942	76.816

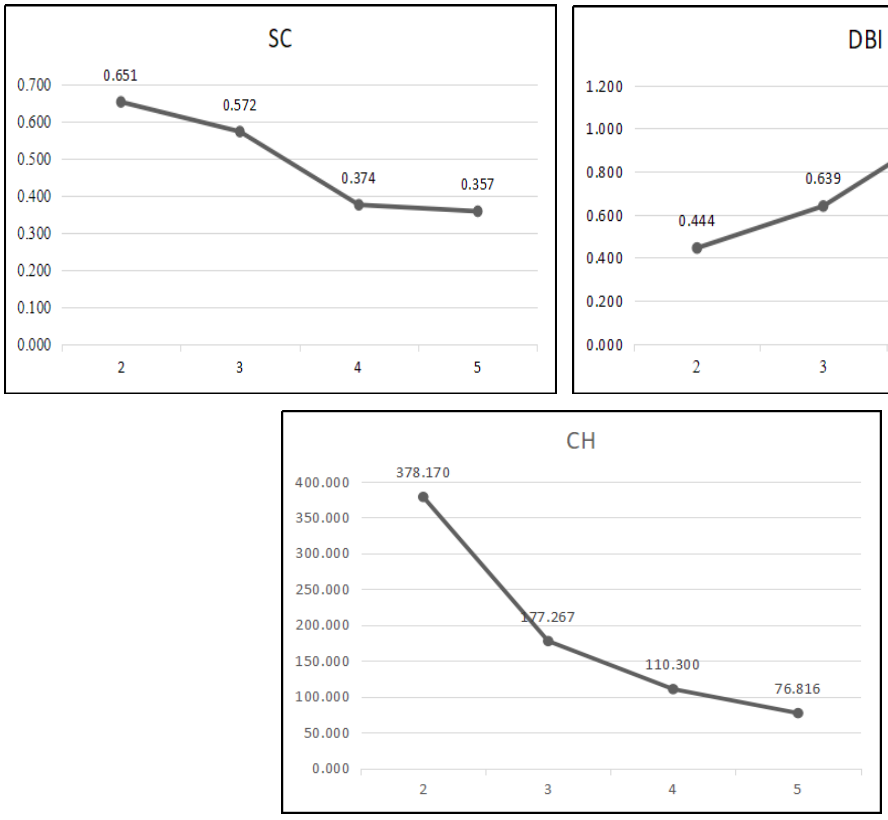

Figure 3. Line graph for analysis of classification results for high potassium glass

The sensitivity analysis of the classification results of high potassium glass is shown in Figure 3 and Table 4. When K equals to 2, SC is the largest, the samples of the same category are closest to each other, and the samples of different categories are the farthest apart; DB is the smallest, the distance within the category is the smallest, while the distance between the categories is the largest; CH is the largest, the samples within the category are the closest, and the samples between the categories are the most dispersed, i.e., the best classification effect is achieved when K=2 which is the number of sub-classification results of high potassium glass is 2 selected in this paper .

Table 5. Sensitivity analysis of classification results for high potassium glass

K	SC	DB	CH
2	0.512	0.697	193.052
3	0.343	0.937	94.472
4	0.388	0.921	61.612
5	0.329	0.998	45.182

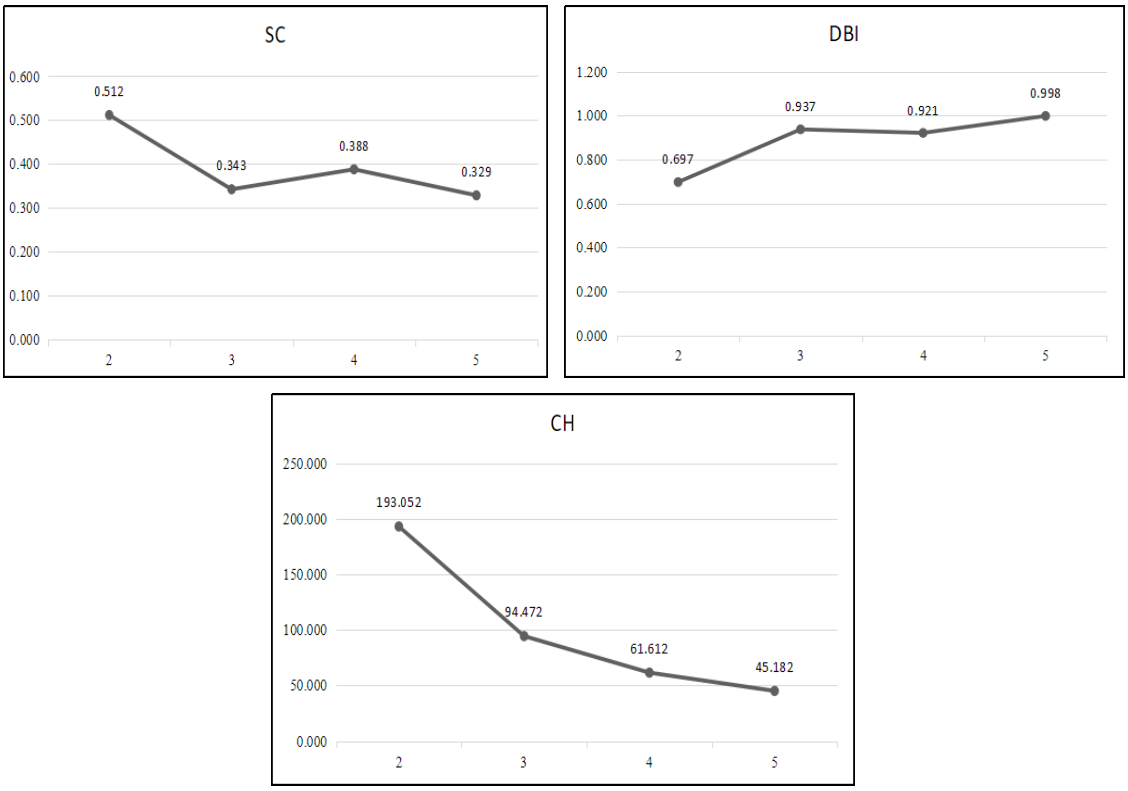

Figure 4. Line graph for analysis of classification results for lead barium glass

The sensitivity analysis of the classification results of lead-barium glass is shown in Figure 4 and Table 5. When K is equal to 2, SC is the largest, with samples of the same category closest to each other and samples of different categories farthest apart; DB is the smallest, with the smallest distance within a category while the largest distance between categories; CH is the largest, with the closest samples within a category and the most scattered samples between categories, verifying that the best classification effect is achieved when K=2, i.e. the number of subclassification results of high potassium glass is 2, selected in this paper.

3. Conclusions

Given that the fundamental difference between each artefact depends on the chemical composition, this paper uses the chemical composition of the artefact as the independent variable and the type of glass artefact as the dependent variable and uses discriminant analysis to establish Fisher's discriminant function to show the classification pattern of glass artefacts. This paper then uses K-means clustering to cluster glass artefacts based on their chemical composition. The number of K-mean clustering centroids was adjusted and the values of the clustering evaluation criteria SC, DBI and CH were calculated for K equal to 2 and K not equal to 2. The sensitivity of the classification results was further analyzed by comparing the different effects of the values on the clustering analysis. The experimental results show that the classification model of ancient glass artefacts based on Fisher's discriminant analysis and K-means clustering has certain simplicity and accuracy and has some practical application value.

References

[1] Surface Weathering of Tuffs: Compositional and Microstructural Changes in the Building Stones of the Medieval Castles of Hungary[J]. Luigi Germinario;Ákos Török. Minerals. 2020(4)

[2] The use of Brazilian Test as a Quantitative Measure of Rock Weathering [J] . A. Aydin,A. Basu. Rock Mechanics and Rock Engineering . 2006 (1)

[3] Weathering features of a five-story stone pagoda compared to its quarrying site in Geumgolsan Mountain, Korea[J]. Jo Young Hoon;Lee Chan Hee. Environmental Earth Sciences. 2022(6)

[4] Ó Foghlú Billy,Wesley Daryl,Brockwell Sally,Cooke Helen. Implications for culture contact history from a glass artefact on a Diingwulung earth mound in Weipa[J]. Queensland Archaeological Research,2016,19.

[5] Bian Wei,Tao Dacheng. Asymptotic Generalization Bound of Fisher's Linear Discriminant Analysis.[J]. IEEE transactions on pattern analysis and machine intelligence,2014,36(12).

[6] Rig Das,Ratnakar Dash,Banshidhar Majhi. Hyperspectral Image Classification Based on Quadratic Fisher's Discriminant Analysis and Multi-class Support Vector Machine[J]. IETE Journal of Research,2014,60(6).

[7] Zhe Liu,Jianmin Bao,Fei Ding. An Improved K-Means Clustering Algorithm Based on Semantic Model[P]. Information Technology and Electrical Engineering 2018,2018.

[8] Liu Wei,Zou Peng,Jiang Dingguo,Quan Xiufeng,Dai Huichao. Zoning of reservoir water temperature field based on K-means clustering algorithm[J]. Journal of Hydrology: Regional Studies,2022,44.

[9] Jumadi Dehotman Sitompul Bernad,Salim Sitompul Opim,Sihombing Poltak. Enhancement Clustering Evaluation Result of Davies-Bouldin Index with Determining Initial Centroid of K-Means Algorithm[J]. Journal of Physics: Conference Series,2019,1235.

[10] Xu Wang,Yusheng Xu. An improved index for clustering validation based on Silhouette index and Calinski-Harabasz index[J]. IOP Conference Series: Materials Science and Engineering,2019,569(5).