Classification of ancient glass artifacts based on logistic regression and k-means clustering

Jiahe Gao *

School of Economics, Southwestern University of Finance and Economics, Sichuan, Chengdu, 611130

* Corresponding author: jackhuntergao@gmail.com

Abstract. Chinese glass products have a long history. In this paper, for the weathering process and heritage components of ancient glass, we firstly established a logistic regression model to analyze the role of each class of chemical components on glass classification; then, in this paper, we identified the key elements with the greatest influence on glass classification: potassium oxide and lead oxide by gradually adding regression elements; subsequently, for each class of high potassium and lead barium glass, we Then, for each type of glass of high potassium and lead-barium, we selected some chemical components to establish a K-means clustering model, and after calculation, the high potassium glass was divided into four subclasses of GJ-1, GJ-2, GJ-3, and GJ-4, and the lead-barium glass was divided into four subclasses of QB-1, QB-2, QB-3, and QB-4, and the characteristics of each subclass were described. Finally, reasonableness and sensitivity tests were conducted to prove the reasonableness and robustness of the classification results.

Keywords: K-means clustering, logistic regression, glass artifacts.

1. Introduction

Glass is one of the oldest man-made materials invented with the birth of civilization, born in 2500 B.C. in the two river valleys of Central Asia [1-2]. At the end of Spring and Autumn Period, glass-making technology was winding to the Central Plains through the Western region, and was locally digested and recreated according to the aesthetics and needs of the East, resulting in a series of glass with local characteristics such as lead-barium glass in the Warring States period and potassium glass in the Song and Yuan dynasties, which gradually became an important example of the exchange and integration of Eastern and Western civilizations [3-4].

Ancient glass is constrained by the production process and is susceptible to weathering by the surrounding environment during long periods of burial, leading to changes in the internal chemical composition of cultural relics, which is a great obstacle to the identification and restoration process [5]. In order to divide the glass by category and for each classification according to the relative content of each component, the samples in the two major categories of high potassium and lead-barium were further divided into different subcategories [6]. For this dichotomous selection problem, in this paper, the major categories were transformed into 0-1 dependent variables, and then a logistic regression model was established to determine the effect of the high or low content of each chemical element on the classification results by regression. And accordingly, the importance and direction of the influence of each chemical element composition on the glass classification status were determined. To find the key elements influencing the classification, we used the chi-square test to determine the correlation between the classification results and each element, and accordingly obtained the key elements influencing the glass classification. For the classification of large subclasses, some characteristic elements were first selected to obtain the elbow relationship between the number of subclasses and the degree of distortion, and the appropriate number of classifications was selected, after which the subclass classification results were obtained by substituting into the k-means clustering algorithm [7-8].

2. Glass classification model building and solving

2.1. Glass classification law analysis

The artifact samples were divided into two categories: high-potassium glass and lead-barium glass. In order to find the classification pattern of glass, this paper attempts to establish a logistic regression model to solve the influence of the content of each chemical composition on the determination of glass type, using the chemical composition in glass as the independent variable and the glass type as the dependent variable [9-10].

Since the weathering process has an effect on the chemical composition of glass artifacts, in order to eliminate this interference, we used the predicted pre-weathering chemical composition content in Problem 1 as the true chemical composition of the weathered sampling sites.

(1) Full-component logistic regression model

Since the artifact samples were divided into two categories, high potassium glass and lead-barium glass, we set the artifact category variables to 0-1 variables as follows:

$$T = \begin{cases} 1, Lead\ barium\ glass \\ 0, High\ Potassium\ Glass \end{cases} \tag{1}$$

Set as the probability that the sample belongs to high potassium glass.

We consider that:

$$P = \begin{cases} \geq 0.5, Lead\ barium\ glass \\ < 0.5, High\ Potassium\ Glass \end{cases} \tag{2}$$

To map P to the real number domain, we transformed P using the Sigmoid functional form to obtain Ln p/(1-p). To measure the effect of the content of different chemical components on the glass classification, we used Ln p/(1-p) as the dependent variable and the chemical component content of the sample as the independent variable to establish the logistic regression equation.

$$Ln\,{}^{P_m}/_{1-P_m} = \beta_0 + \beta_1 x_{m1} + \cdots + \beta_{14} x_{m14} \tag{3}$$

The above equation is transformed to obtain. where the subscript m denotes the mth glass artifact; β0 is a constant term, β1, β2.....β14 are regression coefficients indicating the effect of a change in the content of an element on the classification result, and xm, n denotes the nth element content of the mth glass artifact. After substituting the data and solving the regression equation, we obtain the regression coefficients for each chemical component as follows.

The above equation is transformed to obtain. where the subscript m denotes the mth glass artifact; β0 is a constant term, β1, β2.....β14 are regression coefficients indicating the effect of a change in the content of an element on the classification result, and xm, n denotes the nth element content of the mth glass artifact. After substituting the data and solving the regression equation, we obtain the regression coefficients for each chemical component as follow figure 1.

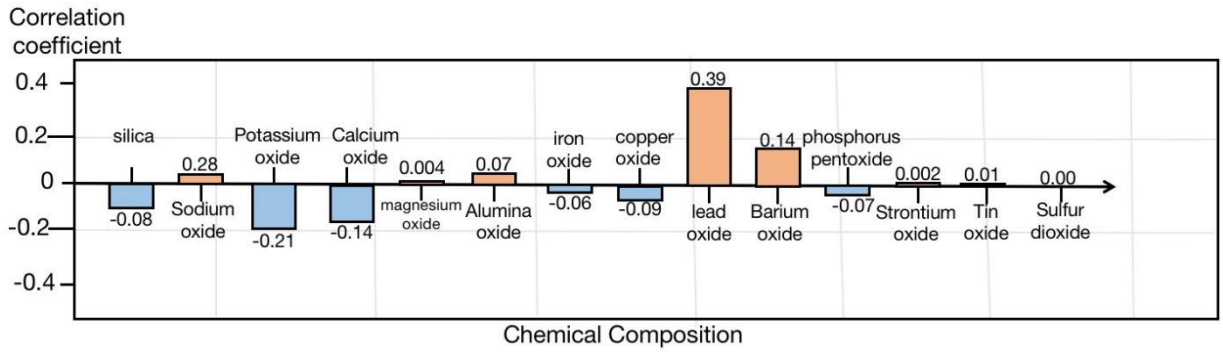

Figure 1. Logistic regression coefficients of 14 categories of components

If a certain regression coefficient is positive, it may make P larger and closer to 1 when the proportion of corresponding components is larger, that is, the more likely to judge this artifact glass as lead-barium glass, and conversely if a certain regression coefficient is negative, it may make P smaller and more likely to judge this artifact as high-potassium glass when the proportion of corresponding components is larger and closer to 0.

It can be tentatively judged that the factors that significantly affect the classification judgment include lead oxide, barium oxide, potassium oxide, and calcium oxide.

(2) Key Feature Selection

In the previous analysis, we initially obtained the influence of each chemical composition on the glass classification judgment, and now we want to find the key influencing components for the classification judgment, and the feature selection steps are as follows:

Determine the correlation between the classification results and the components using the chi-square test 2. Rank the chemical components in descending order of correlation 3. Build a logistic regression classification model in Python 4.

If the accuracy of the model reaches a high level, the training is finished; if the accuracy of the model does not reach a high level, the training is finished.

Return to 4 and add the most correlated component among the remaining components to the classification basis.

The accuracy of the classification model with different number of classification bases is as follows figure 2:

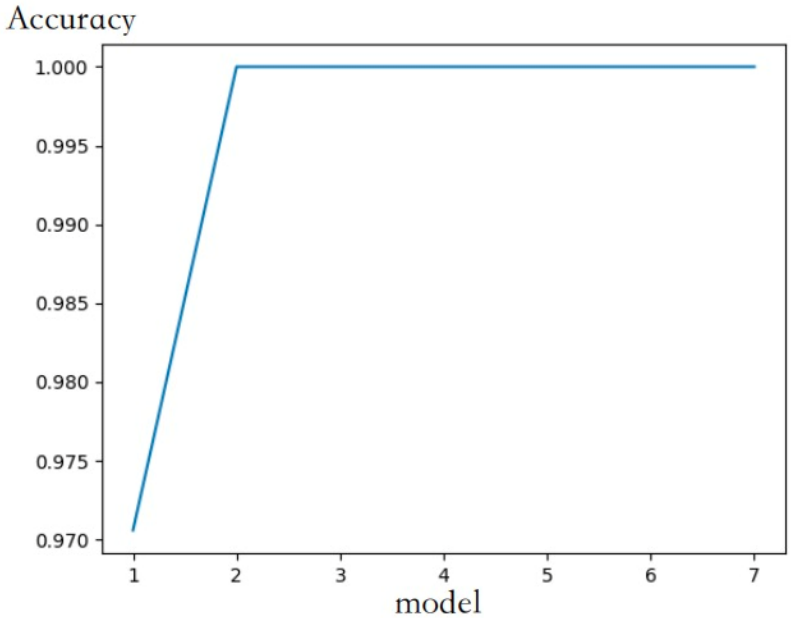

Figure 2. Logistic regression coefficients of 14 categories of components

We find that the model accuracy has reached 100% when the two chemical components with the highest correlation are used as the basis for classification, so we refer to the two chemical components with the highest correlation as key features, which are potassium oxide and lead oxide.

2.2. Glass subclass classification model

In this subsection, we need to subdivide the two categories of lead barium glass and high potassium glass into different subcategories, due to the limitation of space, only lead barium glass is used as an example here, and high potassium glass is classified in the same way as lead barium glass.

(1) Selection of feature elements based on filtering method

Considering that not all elements will have a large impact on the segmentation subcategories, we first identify the feature elements that need to be focused on, which are characterized by large differences among the data of the element that contribute to the classification. In the specific processing, we first normalize the data of each component to eliminate the influence caused by the difference in content between components, and then realize the selection of feature elements based on the variance of the data of each component; the larger the variance, the greater the influence of the element on the classification. The normalized variance of each chemical component in lead barium glass is as follows figure 3.

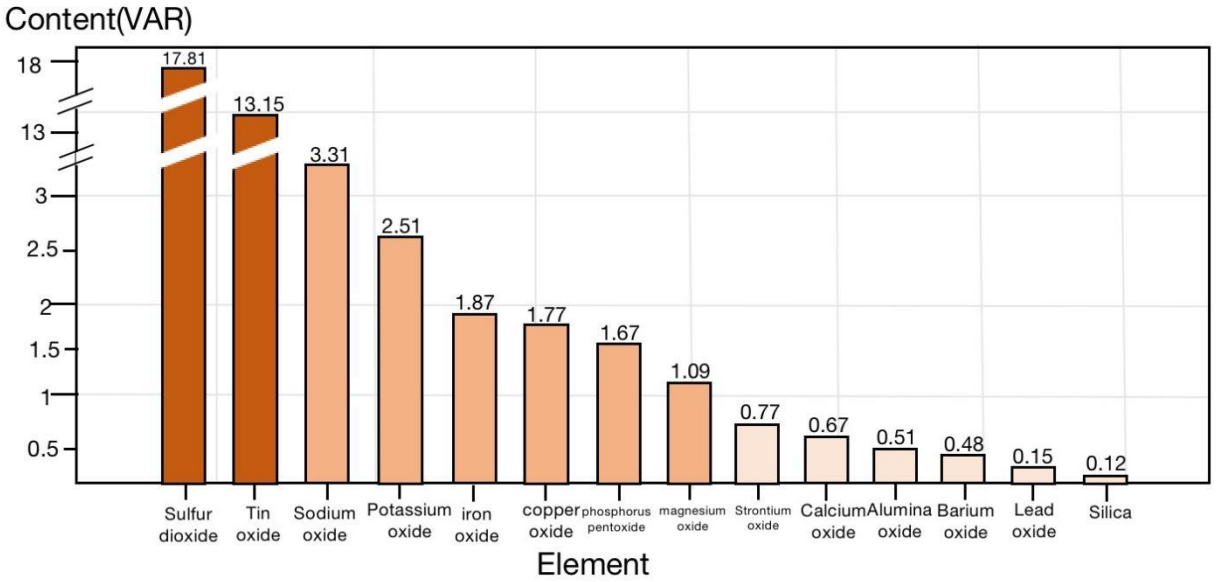

Figure 3. Logistic regression coefficients of 14 categories of components

Here, we set a variance of 1 as the cut-off, and the chemical components with a variance greater than 1 are regarded as characteristic elements. It should be noted in particular that the variance of the two components, sulfur dioxide and tin oxide, is extremely large, but through analysis of the data, we believe that the large variance of both is due to the failure to detect these two elements in many artifacts, resulting in extremely small data means and thus extremely large variance. The data for these two components were not universal, so we discarded them as well.

Ultimately, the characteristic elements of lead-barium glass are: nitrogen oxide, potassium oxide, iron oxide, copper oxide, phosphorus pentoxide, and magnesium oxide.

In the same way, we found the characteristic elements of high potassium glass: NaO, PbO, BaO, P5, and SrO

(2) K-means clustering algorithm

After we get the feature elements, we adopt the unsupervised K-means clustering algorithm to aggregate the samples in lead-barium glass into several classes according to the data features, and the principle of the K-means algorithm is:

Step.1 Determine the number of clusters to be classified (cluster) k

Step.2 Randomly select k data as initial cluster centers (cluster centers)

Step.3 The Euclidean distances of the remaining sample points to these k initial clustering centers are calculated, and the sample points are assigned to the class with the closest distance to the class where the center is located

Step.4 Adjust the new class and calculate the cluster center

Step.5 Loop Step.3 and Step.4, and stop the loop if the cluster center converges or reaches the number of iterations

where we define the sum of squares of the differences in the corresponding feature element content of the Euclidean distance samples between sample points.

(3) Determining the number of subclasses using the elbow chart method

Since K-means is an unsupervised algorithm, we need to determine the number of classifications subjectively. Our clustering goal is that the distance from all sample points to their cluster centers is as small as possible, and the smaller the distance, the more compact the clusters are and the higher the intra-cluster similarity. Accordingly, we consider the square of the distance from all samples to its cluster center as the degree of distortion, which will be different when the value of k is different. For data with a certain degree of differentiation, the distortion decreases with the increase of the number of classifications. When the number of classifications reaches a certain threshold, the distortion will be greatly improved, and then decreases slowly, which means that increasing the number of classifications will have relatively less effect on the improvement of the distortion. This critical point can be considered as the point with better clustering performance, as shown in the following figure 4:

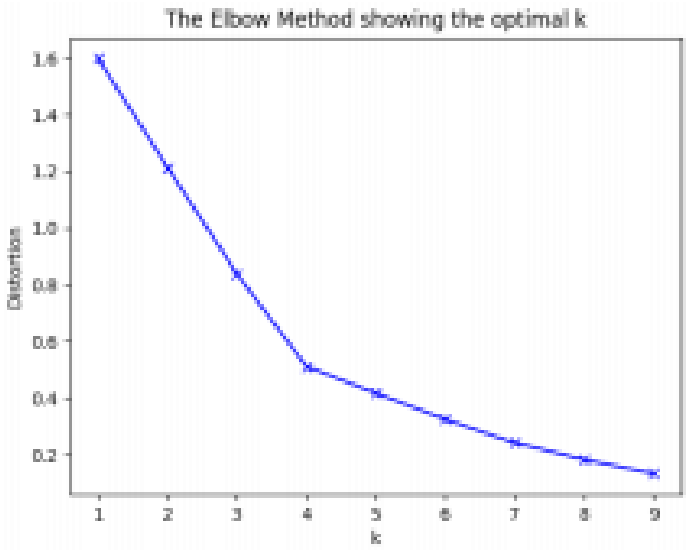

Figure 4. Variation in the degree of distortion of lead-barium glass data with the number of subclasses

Clearly, for lead-barium glass, the optimal number of subclasses to divide is 4.

We applied the same procedure to the data for high potassium glass and obtained the same optimal number of subclasses as 4.

(4) Subclass classification results

The data of high potassium glass and lead-barium glass were substituted into the K-means model respectively, and the clustering results were obtained as follows table1:

Table 1. Subclass classification results

High potassium subclass	Artifacts	Lead and barium subclass	Artifacts
GJ-1	1,3,4,5,7,9,10,12,18,22,27	QB-1	2,19,31,41,43,49
GJ-2	6	QB-2	8,25,26,29,37,44
GJ-3	13,14,16	QB-3	23,36,38
GJ-4	3,21	QB-4	11,20,43

It can be seen that the high potassium glass is divided into four subclasses GJ-1, GJ-2, GJ-3 and GJ-4, while the lead-barium glass is divided into four subclasses QB-1, QB-2, QB-3 and QB-4.

The mean values of characteristic elements for each subclass of high potassium glass and lead-barium glass are shown in the following figure 5:

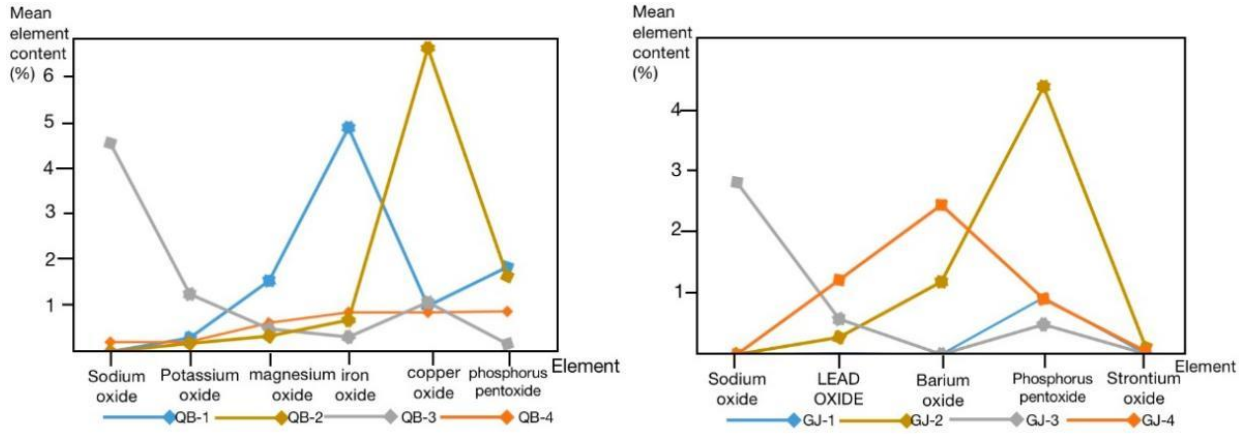

Figure 5. Mean value of characteristic elements for each subclass

We can clearly see that there are significant differences in the chemical composition content between samples of different subcategories of the same category. Specifically, the four subcategories of lead-barium glass are characterized as

QB-1 From the overall content of characteristic elements, the content of iron oxide is significantly higher than other characteristic elements in this subcategory, with a mean value as high as 4.90%; the content of sodium oxide and potassium oxide is lower, with mean values of 0% and 0.3%, respectively; the content of magnesium oxide, copper oxide, and phosphorus pentoxide is at an ordinary level in this subcategory, about 12%; after observing the filtered elements, we found that the content of silica and lead oxide is higher, with mean values of 42.80% and 30.16%, while tin oxide and sulfur dioxide were hardly detected in the samples; the color of the artifacts contained light blue, light green, purple, and black.

QB-2 From the overall content of characteristic elements, the contents of sodium oxide and potassium oxide are relatively low in this subcategory, 0.21% and 0.22%, respectively, while the mean values of magnesium oxide, iron oxide, copper oxide, and phosphorus pentoxide are similar, about 0.6-0.9%; after observing the filtered elements, we found samples with sulfur dioxide content as high as 30% in this category, which we think may be due to the special environment they are in. We think that it may be caused by the special environment in which it is located, and does not constitute the main judgment of the classification; the colors of the artifacts include purple, light blue, light green, dark green, dark blue, black, and blue-green, among which, light blue and dark green are slightly more.

QB-3 From the overall content of characteristic elements, sodium oxide is significantly higher than other characteristic elements in this subclass, with a mean value of 4.56%, potassium oxide and copper oxide are slightly lower, with a mean value of about 1.2%, and magnesium oxide, iron oxide and phosphorus pentoxide have a mean value of less than 0.5%; after observing the filtered elements, it is found that tin oxide and sulfur dioxide are not detected in the samples; the colors of artifacts include dark blue, light blue, dark green, and green, of which light blue is more.

QB-4 From the overall content of characteristic elements, copper oxide content is significantly higher than other characteristic elements in this subclass, with a mean value as high as 6.63%, while sodium oxide, potassium oxide, magnesium oxide, and iron oxide have mean values between 00.7%, which is significantly lower in this subclass; after observing the filtered elements, it is found that strontium oxide, tin oxide, and sulfur dioxide are hardly detected in the samples; the colors of artifacts are half purple and half light blue, respectively. Each subclass of high potassium glass is equally characterized.

(5) Rational Analysis

Here we take lead barium glass as an example for our analysis.

In order to justify the subclass division, we need to show that there are significant differences among the four subclasses of lead-barium glass. We see that the mean value of sodium oxide content of QB-3 is 4.56 significantly higher than the remaining three categories of 0%, 0.21% and 0%, and the mean value of iron oxide content of GB-1 is 4.9 significantly higher than the remaining three categories of 0.86%, 0.32% and 0.69%, and the mean value of copper oxide content of GB-4 is 6.64 significantly higher than the remaining three categories of 1%, 0.86% and 1.08%.

Combined with the pictures of the mean values of subclass characteristics, we clearly see that each subclass is significantly different from the other subclasses, indicating that our way of classification is reasonable.

(6) Sensitivity Analysis

Here we still use lead-barium glass as an example for our analysis.

In order to test the sensitivity of the subclass classification, we use the method of changing the characteristic elements: for lead-barium glass, the characteristic elements we chose in the previous section are: NaO, KO, FeO, CuO, P5, and MgO.

We perform two sensitivity experiments.

Experiment 1 subtracts magnesium oxide from the original characteristic elements and performs K-means clustering analysis again

Experiment 2: Strontium oxide was added to the original characteristic elements, and K-means clustering analysis was performed again.

The elbow plots of the two experiments are as follow figure 6.

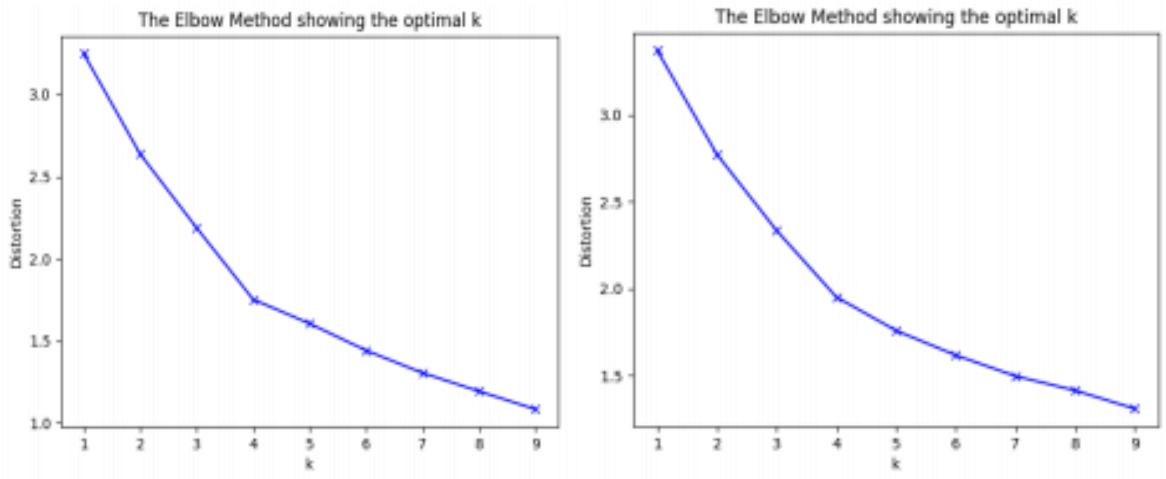

Figure 6. Elbow diagram of Experiment 1 and Experiment 2

Thus, we found that it is robust to choose the number of subclasses as 4.

In addition, the final partitioning results of Experiment 1 and Experiment 2 are identical to the original results, indicating that the K-means subclass partitioning method we used is also robust.

3. Conclusion

This paper gives the basis for the classification of glass categories based on the data of ancient glass artifacts. For this binary selection problem, we can transform the major categories into 0-1 dependent variables and then establish a logistic regression model to determine the influence of the high or low content of each chemical element on the classification results through regression. And accordingly determine the importance and direction of the influence of each chemical element composition on the glass classification status. To find the key elements influencing the classification, we used the chi-square test to determine the correlation between the classification results and each element, and accordingly obtained the key elements influencing the glass classification. For the

classification of large subclasses, some characteristic elements were firstly selected to get the elbow relationship between the number of subclasses and the degree of distortion, and the appropriate number of classifications was selected, after which the subclass classification results were obtained by substituting into the k-means clustering algorithm. To determine the robustness of the subclass classification results, vary the feature elements and compare the classification results before and after several times to determine the robustness of the model.

References

[1] Zhang Jiawen. Application of machine learning algorithms in webshell detection and its security research [D]. Nanjing University of Posts and Telecommunications, 2022.

[2] Li Wenmeng. QoS routing optimization algorithm based on traffic classification and reinforcement learning for SDN networks [D]. Nanjing University of Posts and Telecommunications, 2022.

[3] Wang Xianhao. Research on selective integration improvement algorithm based on diversity metric [D]. Nanjing University of Posts and Telecommunications, 2022.

[4] Cai Qingyuan. Research on the analysis method of customers' electricity consumption behavior based on load data preprocessing[D]. Nanjing University of Posts and Telecommunications, 2022.

[5] Tan Ruhan. Analysis and detection of user malicious behavior based on deep learning [D]. East China Jiaotong University, 2022.

[6] Liu Yang. Research on laryngoscopic image classification under improved convolutional neural network based on weighted entropy K-means clustering [D]. Changchun University of Technology, 2022.

[7] Yang Yanfei. Research on self-detonation detection of aerial glass insulators based on deep learning [D]. Chongqing University of Technology, 2021.

[8] Mao M. X. Ancient economic and social development of Nanliujiang River and the Maritime Silk Road [D]. Guangxi Normal University, 2020.

[9] Qi Haoyue. Exploration of foreign cultural elements of Liao Dynasty in the context of Silk Road [D]. Inner Mongolia Normal University, 2020.

[10] Jin Xinpei. Research on mixed reality fish tank system and its interaction technology [D]. Shandong University, 2020.