A Comprehensive Study on the Use of Multiple Classification Models for the Type Classification Problem of Unknown Artifacts

Jiajun Chen, Zhenhui Ou, Nuo Chen, Shulong Lv *

College of Maynooth International Engineering, Fuzhou University, Fuzhou, 350108, China

* Corresponding author: wujispace@fzu.edu.cn

Abstract. The early glass was often made into bead-shaped ornaments in the West Asian and Egyptian regions and introduced to China. The ancient glass in China was made locally after absorbing its technology, and the appearance of glass products is similar to that of foreign ones, but the chemical composition is different. Therefore, it is a very meaningful task to analyze and identify the composition of ancient glass products. In this paper, based on the known chemical composition data of excavated artifacts, the statistical pattern of whether the artifacts are weathered or not is visualized by using frequency distribution charts, and the interference of random errors on the results is tested through a chi-square test. Subsequently, the two major categories of high potassium glass and lead-barium glass are classified, while the problem of classifying the types of unknown artifacts is solved. In this paper, a scientific, reasonable, and systematic system of glass-type classification is established, the prediction is carried out on this data set, and more accurate results are obtained with a correct rate of 90%, which proves that the model has a strong generalization and practical significance.

Keywords: K-means, Low variance filtering, High relevance filtering, multi-category model.

1. Introduction

For thousands of years, glass was introduced to China from West Asia and Egypt as an important trade evidence, and was made from local materials, resulting in a variety of ancient glass with similar appearance but different chemical compositions. Depending on the composition of the flux added during the production process, ancient glass can be divided into various types [1]. For example, lead ore is added as a flux in the glass firing process, and its content of lead oxide (PbO) and barium oxide (BaO) is high, so this kind of glass is called lead-barium glass. And high potassium glass is fired with a high content of potassium substances such as grass wood ash as a flux.

Ancient glass is highly susceptible to weathering by the burial environment, and glass weathering has a great impact on the category analysis of ancient glass artifacts, so it is a very challenging task to analyze and identify the composition of ancient glass products [2-3]. In this paper, based on the known chemical composition data of excavated artifacts, we first visualize the statistical law of three factors, namely, ornamentation, glass type and color of artifacts, and whether the artifacts are weathered or not, through frequency distribution charts. Then, we used the low variance filtering and high correlation filtering methods to filter the chemical composition of the two categories of high potassium glass and lead-barium glass and used a combination of multiple classification models to solve the problem of classifying the types of unknown artifacts [4].

The article evaluates and optimizes the models used in the whole set of glass artifact identification systems, and analyzes the advantages of each model comprehensively and objectively, while improving the deficient parts of them, with the following main contributions: (1) Qualitative analysis of glass weathering, followed by the summary of statistical laws according to glass types, and prediction of the proportion of chemical composition before weathering at weathering points. (2) Analyzing the distribution patterns of high potassium and lead-barium glasses in terms of chemical composition, and developing mathematical models in the two categories of glasses separately, performing subclass classification and giving specific criteria as well as classification results and

evaluating them. (3) For high potassium and lead-barium glasses, correlation analysis and difference analysis of correlations were performed for their chemical compositions, respectively.

2. Analysis of the relationship between weathering of glass artifacts

2.1. Data pre-processing

We first processed the outliers in the data, i.e., we summed each row of data and observed the total proportion of its chemical composition. The summation of the proportions of each component in the glass product should be 100%, but the summation of the proportions of its components may not be 100% due to testing methods and other reasons. Therefore, the data with the cumulative sum of the composition proportions between 85% and 105% were considered valid data, and the two sets of data values that were not in the interval of 85%-105%, i.e., the heritage sampling point 15 and the heritage sampling point 17, were removed.

For the given data, the large differences in the mean values of each column can have a significant impact on the variance results, causing them to not fully reflect the degree of fluctuation or dispersion of their data. Therefore, the data in each column were normalized using the min-max standardization method, i.e.

$$A_{ij}^{*} = \frac{A_{ij} - \min(A_{ij} \mid k = j)}{\max(A_{ik} \mid k = j) - \min(A_{ik} \mid k = j)} \tag{1}$$

Observational analysis of the data in both the tin oxide and sulfur dioxide columns was too low in sample size and too large in variance to provide valid data feedback for the overall modeling. Tin oxide and sulfur dioxide only account for 1/16 and 3/16 of the total number of high potassium glasses, respectively, and 5/49 of the total number of lead-barium glasses, respectively. the sample size is too small to be convincing, so they are directly excluded. For other numerical gaps in the table, the values were replaced with 0 values.

2.2. Model building and solving

The relationship between surface weathering and its glass type, decoration, and color is first visualized using frequency distribution plots, as shown in Figure 1 below.

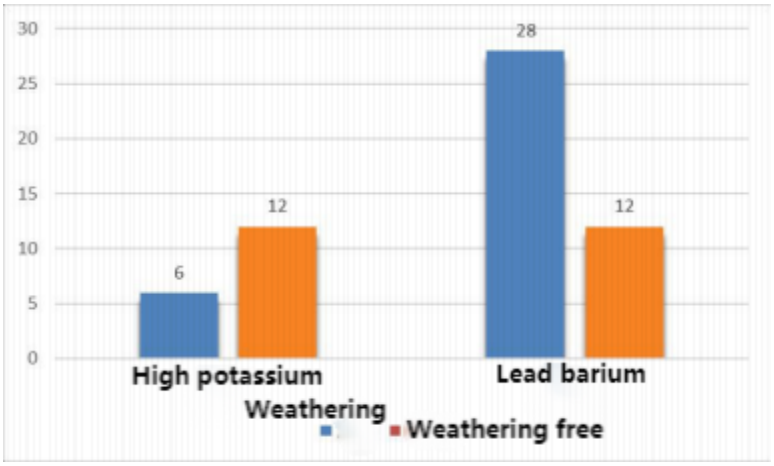

Figure 1. Perspective view of data on glass type and surface weathering

It can be seen from Figure 1 that for the high potassium type, 66.7% of the artifacts are unweathered and only 33.3% are weathered, while for the lead-barium type, 70.0% of the artifacts are weathered, which is much higher than the 30.0% that are unweathered. The percentage of glass artifacts with high potassium is not easy to weather.

Figure 2 shows that the glass objects with motif A are weathered and unweathered in equal proportions, the objects with motif B are all weathered, and the objects with motif Care weathered at

56.67%, which is more than half. It is tentatively concluded that the different decoration of the artifacts affects the weathering of the artifacts, with artifacts with B decoration being the most susceptible to weathering, followed by artifacts with C decoration, and finally artifacts with A decoration being the least affected.

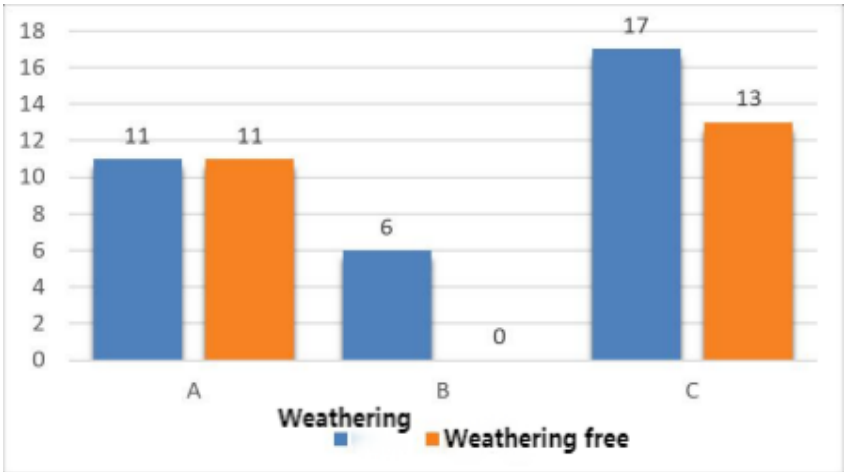

Figure 2. Perspective view of data on ornamentation type and surface weathering

According to the size of the weathering ratio, the cross-frequency table of artifacts color and whether weathering in descending order: black > blue-green = light blue > dark green > purple > light green > green = dark blue. As some of the color artifacts sample size is too small to produce accurate analysis, so select a sample size greater than 5 colors. Through the above analysis, to some extent, it can be considered that the degree of influence of artifact color on weathering satisfies: blue green ≈ light blue > dark green.

From these cross-tabulations alone, it is not possible to distinguish whether these differences are real or due to random errors arising from the small sample size of the extracted data. Therefore, we used a chi-square test for validation. The chi-square test uses the chi-square statistic (χ^2), which can be used to measure the degree of correlation between two categorical variables and measures the difference between the observed cell values and the values that would be expected if the row variables were uncorrelated.

$$\chi^2 = \sum \frac{(f_0 - f_e)^2}{f_e} \tag{2}$$

Where f_0 denotes the frequency of observations, f_e denotes the frequency of expectations, and the χ^2 statistic describes the closeness of the observations to the expectations.

The decision to reject the original hypothesis is made by comparing the calculated χ^2 statistic with the critical value in the chi-square distribution. This is shown in Table 1 below:

Table 1. Cardinality test for type

	Value	Degree of freedom	Progressive saliency (bilateral)
Pearson Cardinal	4.957[a]	2	.084
Continuity correction	5.452	.020	
Likelihood ratio (L)	7.120	2	.028
Number of active cases	58		

It can be seen that the p-value of the statistic of the two-sided test of the chi-square distribution is less than 0.1, which means that the difference between the weathering or not of the different glass types is real.

The cardinality test for ornamentation can be seen in Table 2, which shows that the difference between the weathering of different ornamentation is real, that is, the weathering ratios of

ornamentation with A, B, and C are different, and the weathering ratios can be obtained by combining the cross-linked table with the various ornamentation of the artifacts: B>A>C.

Table 2. Cardinality test on ornamentation

	Value	Degree of freedom	Progressive saliency (bilateral)
Pearson Cardinal	4.957[a]	2	.084
Likelihood ratio (L)	7.120	2	.028
Number of active cases	58		

For the chi-square distribution of color, we can see from Table 3 that the p-value of the statistic for the two-sided test is greater than 0. 1, so we do not believe that the color difference has a significant effect on the difference in weathering in this test. We could not obtain from the sample data of this test which colors are more likely to weather the surface of the glass artifacts.

Table 3. Cardinality test on color

	Value	Degree of freedom	Progressive saliency (bilateral)
Pearson Cardinal	9.432[a]	8	.307
Likelihood ratio (L)	12.636	8	.125
Number of active cases	58		

After the above methodological analysis, it can be concluded that glass artifacts of the lead-barium type are more susceptible to weathering, while glass artifacts of the high potassium type are relatively less susceptible to weathering. artifacts with the B motif are the most susceptible to weathering, followed by those with the C motif, and those with the A motif are the least affected. It was not possible to obtain from the sample data of this test which colors of glass artifacts were more susceptible to surface weathering.

3. Modeling the classification of high potassium glass and lead-barium glass

3.1. Application of multi-classification model classification methods

For each type of glass, there are more types of chemical composition, i.e., more types of eigenvalues, and also the labels of the glass types are known, so the problem is a supervised binary classification problem. The following multiple classification models are used to determine the classification laws of glass, and the models obtained from each are used as the classification laws of glass types.

The K-nearest neighbor (KNN) classifier usually considers similar samples to be more similar than dissimilar samples when using data with known labels to make predictions for samples with unknown labels [5-7]. Therefore, in the feature space, similar samples should be more similar in distance, while dissimilar samples are more distant. Based on this idea, given a sample with a label to be predicted, the distance between the sample to be classified and all the samples in the sample space is calculated to find the k points with the smallest distance value, called k nearest neighbors. The label with the highest number of occurrences in these k categories is taken as the final result of the sample to be predicted. In this paper, we find that the model with k=5 has the highest accuracy in the test set after several experiments, so we use the model with k=5 as the classification law for glass types.

A random forest classifier is a method for classifying data based on multiple decision trees [7-8]. It uses decision trees as the basic unit and integrates several decision trees through the idea of integrated learning and then trains based on multiple trees to construct several different training sets to expand the classification differences to obtain accurate classification results.

For the parameter settings made by the random forest model, the data were divided according to the ratio of the training set: test set = 7:3, and the classification model yielded the results shown in Table 4.

Table 4. Random forest parameter setting

Node split evaluation	Number of decision trees	With Playback Sampling	Maximum depth of the tree	Maximum number of leaf nodes
Gini	100	True	10	50

Naive Bayes is based on the Bayesian principle of learning the probability distribution from the input features to the output labels [9]. In practical applications, it is often assumed to be a probability distribution of some known type, and this step only requires learning the distribution parameters from sample data. Its assumption that the features between samples are independent of each other given the labels allows converting a high-dimensional problem into a problem with multiple one-dimensional probability distributions.

3.2. Application of K-means clustering model

To classify subclasses within each category, we used low variance filtering and high correlation filtering to downscale the data, then clustered them using the k-means clustering method, and finally showed the best classification combinations by visualization tools. After min-max standardization of the data, the variance of each column falls within the [0,1] interval, which facilitates further data processing and prevents the influence of the original feature value size on the variance performance.

For low variance filtering, if a data set, the values of a column are the same, that is, it has very low variance, we usually think that the low variance variables also carry little information, so it can be deleted directly. Putting it into practice means calculating the variance magnitude of all variables and then deleting the smallest ones according to the variance threshold.

$$s^2 = \frac{(M - x_1)^2 + (M - x_2)^2 + (M - x_3)^2 + \ldots + (M - x_n)^2}{n} \tag{3}$$

Where s 2 denotes the variance and M denotes the mean value.

For the high potassium glass, with a variance threshold of 0.03, after low variance filtering, the five chemical components of potassium oxide, calcium oxide, magnesium oxide, aluminum oxide, and iron oxide were screened with a variance greater than 0.03, indicating that these five components are more discrete and easier to be classified. For lead-barium glass, with a variance threshold of 0.03, after low variance filtering, the variance of nine chemical components, such as silica, sodium oxide, magnesium oxide, aluminum oxide, iron oxide, copper oxide, lead oxide, barium oxide, and strontium oxide, is greater than 0.03, indicating that these nine components are more discrete and easier to classify.

Considering the non-linearity and high coupling between the independent variables of this question, it is necessary to perform high correlation filtering to retain only the variables with the highest contribution values from those that are highly correlated with each other. The traditional Person correlation coefficient is often used to measure whether two data sets are on top of a line, and is only applicable to the case where the two variables are linearly related. In this question, the variables are non-linearly related, and Person correlation analysis is not applicable, so we choose the distance correlation coefficient as a measure of the correlation between variables.

$$R^{2}(x, y) = \frac{v^2(x, y)}{\sqrt{v^2(x, x)v^2(y, y)}} \tag{4}$$

$$v^2(x, y) = \frac{1}{n^2}\sum_{i,j}^{n} A_{i,j}B_{i,j} \tag{5}$$

$$A_{i,j} = \left\|x_i - x_j\right\|_2 - \frac{1}{n}\sum_{k=1}^{n}\left\|x_k - x_j\right\|_2 - \frac{1}{n}\sum_{l=1}^{n}\left\|x_i - x_l\right\|_2 + \frac{1}{n^2}\sum_{k,l=1}^{n}\left\|x_k - x_l\right\|_2 \tag{6}$$

$$B_{i,j} = \left\|y_i - y_j\right\|_2 - \frac{1}{n}\sum_{k=1}^{n}\left\|y_k - y_j\right\|_2 - \frac{1}{n}\sum_{l=1}^{n}\left\|y_i - y_l\right\|_2 + \frac{1}{n^2}\sum_{k,l=1}^{n}\left\|y_k - y_l\right\|_2 \tag{7}$$

$$v^2(x,x) = \frac{1}{n}\sum_{i,j}^{n} A_{i,j}^2 \tag{8}$$

$$v^2(y,y) = \frac{1}{n}\sum_{i,j=1}^{n} B_{i,j}^2 \tag{9}$$

K-means clustering is the most basic and commonly used clustering algorithm, and its basic idea is to iteratively find K clusters such that the loss function corresponding to the clustering result is minimized [10]. The loss function can be defined as the sum of squares of the errors of the individual samples from the cluster centroids to which they belong.

$$J(c,\mu) = \sum_{i=1}^{M}\left\|x_i - \mu_{c_i}\right\|^2 \tag{10}$$

Where x_i represents the i-th sample, c_i is the cluster to which x_i belongs, μ_{ci} represents the centroid corresponding to the cluster, and M is the total number of samples.

Radviz visualization is a visualization technique that maps a series of points in multidimensional space to two-dimensional space by a nonlinear method, and it is a multidimensional visualization method based on the design idea of a circular parallel coordinate system. The m radius of the circle represents the m-dimensional space, and a point in the coordinate system is used to represent multiple information objects, and the principle of its implementation refers to the equilibrium theorem of force on objects in physics.

For high potassium glasses, the suitable chemical composition screened was subclassified. The classification results for k = 2 and k = 3 were visualized in the form of RadViz radar plots as shown in Figure 3 and Figure 4.

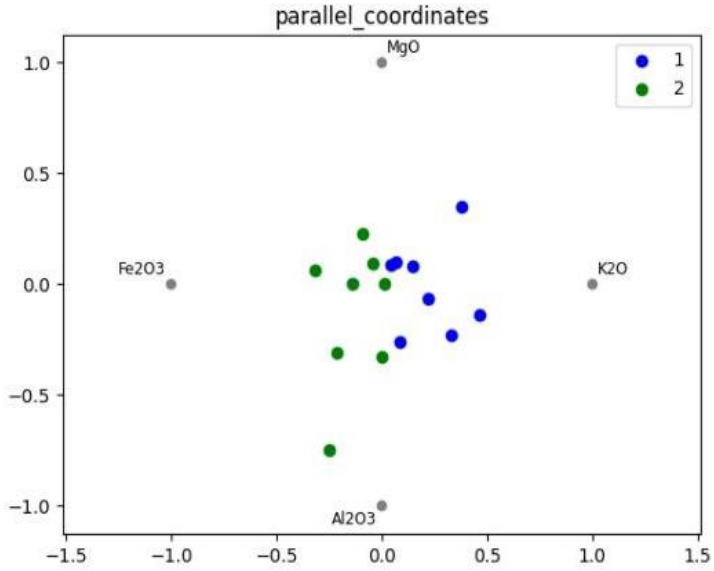

Figure 3. Visualization of high potassium 2 class results

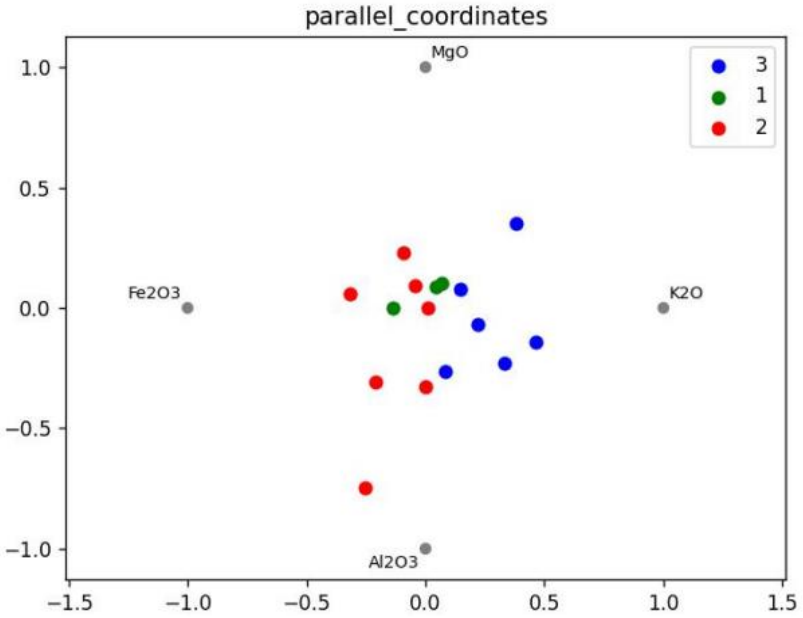

Figure 4. Visualization of high potassium 3 class results

From the comparison, we can see that the clustering classification result is clearest when the k value is 2, so we divide the high potassium into 2 subclasses.

For lead-barium glass, the appropriate chemical composition was screened for subdivision. k = 3 and k = 4 classification results were visualized in the form of RadViz radar plots, as shown in Figures 5 and 6.

So, we can get the clearest clustering classification result when the k value is 3, so we divide the lead barium into 3 subclasses.

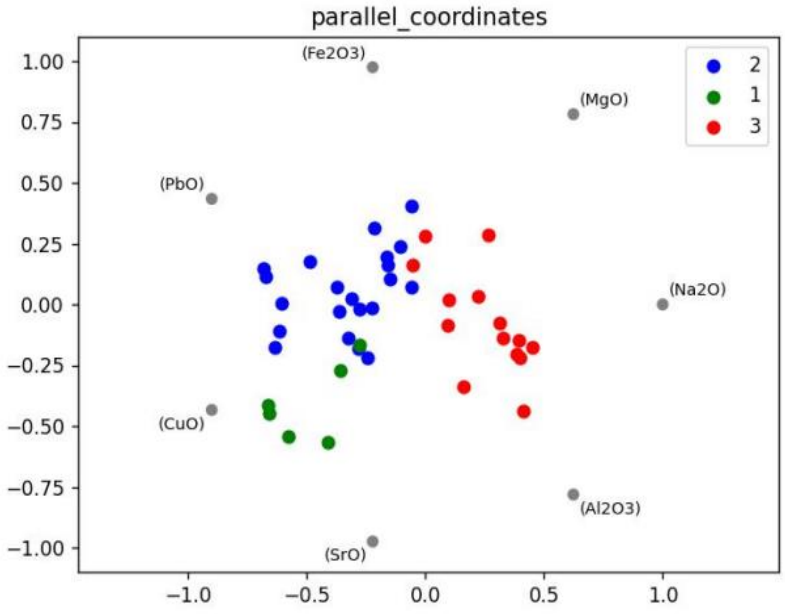

Figure 5. Visualization of the classification results of lead and barium 3 categories

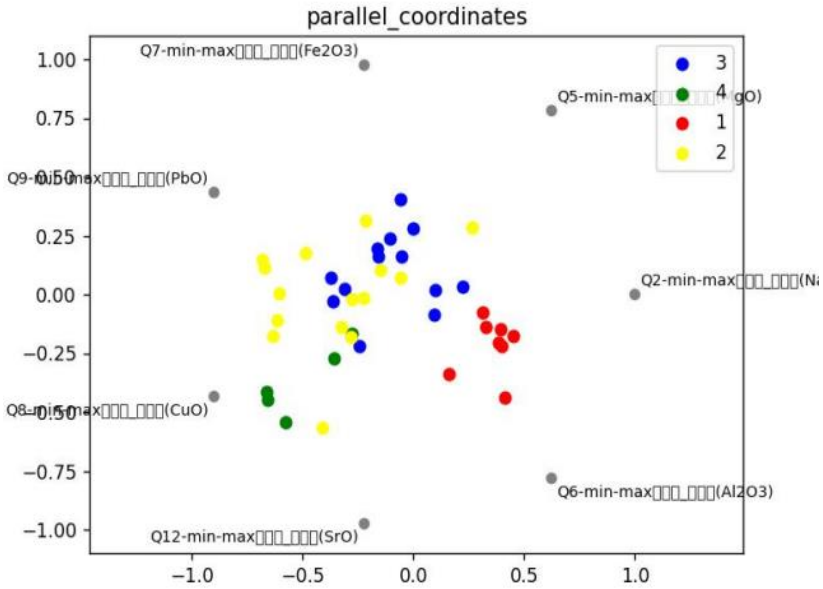

Figure 6. Visualization of the classification results of lead and barium 4 categories

4. Conclusion

In this paper, a scientific, rational, and systematic system of glass-type classification is established based on the known chemical composition data of excavated artifacts. The use of frequency distribution charts and the cardinality of the proposed method to derive whether the artifact's weathering and decoration, glass type has a more obvious pattern. And by using the idea of polynomial fitting and mean value, the chemical composition content data of unweathered artifacts to predict the chemical composition content of weathered artifacts before weathering. The article also completes the subdivision of high potassium glass into two major subclasses and lead-barium glass into three major subclasses, thus achieving the purpose of subclassifying glass types.

References

[1] Ma Shuangyu, Piao Fengxian, Li Junting, Wang Yikun, YUAN Yuhang, Zang Taoliang. Research on composition analysis and identification of ancient glass products [J]. Modeling and Simulation, 2022, (6th issue).

[2] Yin Yulong. Composition analysis of ancient glass products by Association Prediction [J]. Contemporary Chemical Research, 2023, (1st issue).

[3] Wang Zifan. Characteristics of Glass in Early Ancient China [J]. Tiangong, 2020, (3rd issue).

[4] Guo Jiaxin. Chemical composition analysis and identification of glass relics [J]. Kehai Stories Expo, 2022, (32nd issue).

[5] Chen Zhong, Wang Jiegui, Tang Xiwen, Yang Hang. Incoming Wave Direction Estimation Method Based on K-Nearest Neighbor Algorithm [J]. Journal of Detection and Control, 2022, 44 (1): 24 - 28.

[6] Chen Yuming, Li Wei. Particle Vector and K-Nearest Neighbor Particle Classifier [J]. Journal of Computer Research and Development, 2019, 56 (12): 2600 - 2611.

[7] Li Hhengkai, Wang Lijuan, Xiao Songsong. Land use random forest classification in southern hilly region based on multi-source data. Transactions of the Chinese Society of Agricultural Engineering, 2021, 37 (7): 244 - 251.

[8] Dong Hongyao, Wang Yidan, Li Lihong. Overview of stochastic forest optimization algorithms [J]. Information and Computer (Theoretical Edition), 2021, (17th issue).

[9] Wang Huayu. Review of Naive Bayes Algorithm [J]. Mathematics World, 2019, (4th issue).

[10] Yang Junchuang, Zhao Chao. A review of K-Means clustering algorithm [J]. Computer Engineering and Applications, 2019, 55 (23): 7 - 14.