DOI: 10.53469/jtpes.2024.04(01).05

Classification and Recognition of Ancient Glass Cultural Relics Based on K-Means Clustering

Jie Wang[1,*], Ting Yu[1], Qin Wang[2], Angxuan Gu[2]

[1]School of Science, Southwest Petroleum University Chengdu, 610500, China
[2]School of Petroleum and Natural Gas Engineering, Southwest Petroleum University, Chengdu 610500, China
Correspondence Author, 2443396572@qq.com

Abstract: ***The qualitative analysis based on mathematical statistics and the quantitative analysis based on chi square test explored the relationship between the weathering of ancient glass artifacts and their colors, decorations, and types, respectively. It incorporates Principal Component Analysis (PCA), K-means clustering, and fuzzy analysis methods into the research on composition analysis and categorization of ancient glass products. With the aid of analysis of variance and mathematical statistical theory, a quantification table for sub-category divisions is established. Fuzzy mathematical methods, including fuzzy recognition, are applied to develop a sub-category identification model for ancient glass based on the quantification table. Finally, the model's rationality and sensitivity are validated through the classification and identification of given sample data.***

Keywords: Classification of Ancient Glass; Principal Component Analysis; K-Means; Fuzzy Recognition; Chi-square Test.

1. INTRODUCTION

Glass has been recorded in Chinese historical materials for a long time, but because of the confusion of name and texture, and the research on ancient Chinese glass started late in modern times, the research on the composition of ancient silicate glass is relatively lacking. Previous works on ancient glassware mostly studied the cultural and artistic forms of glass and the laws of its own operation and development from the perspective of dynasty succession from the aspects of cultural exchange and chemical analysis[1], and few scholars systematically established mathematical models to conduct quantitative research on the classification methods of glass cultural relics.

At present, the methods used in glass classification research mainly include regression analysis in statistics[2], linear model[3], principal component analysis[4], artificial neural network algorithm in machine learning[4], random forest[5], feature selection method[6], etc. All kinds of methods have advantages and disadvantages, but for specific data samples, There is little research into the methods that can produce clear classification criteria.

In this paper, by analyzing the chemical composition of ancient glass samples, combined with the type of glass, the chemical composition of different glass varieties with significant differences was selected by variance analysis to quantitatively determine the classification limits of high-potassium glass and lead-barium glass. Then, principal component analysis[4]and KMEAN clustering were used to classify ancient glass into subclasses according to the selection index scores of the importance values of variables. According to the different types of glass, the quantitative standard table for the division of ancient glass subclasses is established. The subclass recognition model of ancient glass based on subclass division quantization table is established by fuzzy mathematics method, and its accuracy is verified by variance analysis.

2. THEORETICAL BACKGROUND

2.1 Principal Component Analysis

Principal component analysis (PCA) is a modern data analysis method, simplified calculated by the observations into a set of orthogonal variable, so as to highlight the differences and similarities between samples of a statistical method. Based on different analysis purposes, PCA can achieve functions such as dimensionality reduction, data compression, feature factor extraction and data visualization[7].

The general process is to select k principal components from m original chemical components, each principal component is independent of each other, and are some linear combination of the original components. The extracted principal components were sorted according to the size of their eigenvalues. The principal component with the largest eigenvalues had a strong explanatory ability to the category of glass. When the variance or cumulative contribution rate of chemical component content of k principal components was more than 70%, it indicated that the space composed by kk principal components could retain the information of the original m formed components to the maximum extent, so the results can be used to select a suitable chemical composition index for the clustered glass.

2.2 K-Means

Different from tasks such as classification and sequence labeling, clustering divides samples into several categories based on the internal relationship between data without knowing any sample labels in advance, so that the similarity between samples of the same category is high, and the similarity between samples of different categories is low (that is, the class cohesion is increased, and the class spacing is reduced)[6].

Clustering belongs to unsupervised learning, and K-means clustering is the most basic clustering algorithm. Its basic idea is to find a partition scheme of K clusters iteratively so that the loss function corresponding to the clustering result is minimized. The loss function can be defined as the sum of the squares of the error of the distance between each sample and the center point of the cluster:

$$J(c,\mu)=\sum_{i=1}^{M}x_i-\mu_{ci}^{\ 2} \tag{1}$$

Where x_i represents the i sample, c_i is the cluster x_i belongs to, μ_{ci} represents the central point corresponding to the cluster, and m is the total number of samples.

2.3 Fuzzy Recognition Model

Fuzzy recognition model is an effective method to solve the multi-index comprehensive evaluation. According to the index as the fuzzy subset of potential level, it is denoted $Ai\ (i=1,2,\cdots,N)$, Let $x_j\ (j=1,2,\cdots,M)$ represent the index of subclass division respectively[9]. domain $U=\left\{u\middle|u=(x_1,x_2,\cdots,x_N)\right\}$, so establish a membership function relative to a fuzzy subset of Ai[9]:

$$u_{Ai}\left(x_j\right)=\begin{cases}0 & ,\left|x_j-\bar{X}_j\right|>2S_j\\ 1-\left(\dfrac{x_j-\bar{X}_j}{2S_j}\right)^{\alpha} & ,\left|x_j-\bar{X}_j\right|<=2S_j\end{cases} \tag{2}$$

$$u_{Ai}(u)=\frac{1}{M}\sum_{j=1}^{M}u_{Ai}\left(x_j\right) \tag{3}$$

In the formula, $i=1,2,\cdots N$ represents the number of the fuzzy subset, $j=1,2,\cdots M$ represents the number of $M-th$ component indicators, x_j represents the measured value of the component indicator of each sample, $\overline{X}j$ and S_j represent the mean value and standard deviation of the $j-th$ component indicator of the high-potassium glass and barium lead samples, respectively. $u_{Ai}(x_j)$ represents the membership degree of the $j-th$ index of a glass sample relative to the fuzzy subset Ai, and $u_{Ai}(u)$ represents the degree of a glass sample relative to the fuzzy subset Ai.

According to the maximum membership principle, if $u_{Ai}(u)=\max\limits_{1\le i\le M}\left\{u_{Ai}(u)\right\}$, this u can be considered to belong to the fuzzy subset Ai.

3. EXPERIMENT AND RESULT ANALYSIS

3.1 Data Collection and Preprocessing

The data in this paper comes from a group of ancient glass relics provided by the official website of the 2022 National College Students Mathematical Contest in Modeling. First, the data is processed as follows:

1) Set the value of missing color of sample data to grayish yellow;

2) The chemical composition ratio is accumulated and the invalid data less than 85% or more than 105% are eliminated;

3) Set the proportion of undetected components to 0.

4) The data of two different parts of the same cultural relic is regarded as two different sample data, and the sample data is numbered by cultural relic sampling points.

3.2 Experimental Process and Results

3.2.1 Weathering in Relation to Type, Ornamentation, and Color

3.2.1.1 Qualitative Analysis based on Mathematical Statistics

It is found that lead-barium glass is easy to weather while high-potassium glass is not, indicating that the surface weathering of glass samples is closely related to the type of glass. Class B decorative glass is easily weathered, followed by Class C decorative glass and class A decorative glass, indicating that surface weathering is related to glass decoration; In the samples unearthed, the black glass is only weathered, and the green and dark blue glass is only unweathered. Due to the small number of samples, qualitative analysis shows that blue-green, light blue, black and dark green samples are more susceptible to weathering, and the relationship between surface weathering and color is more complex.

3.2.1.2 Quantitative Analysis based on Chi-square Test

Chi-square test is a very widely used hypothesis testing method proposed by Kar-pearson, the founder of statistics, in 1900. It is often used to test the significance analysis of the difference between two or more sample rates or composition ratios, and to explain whether there is a certain relationship between two types of attribute phenomena. The basic idea is to compare the degree of agreement between the actual value and the theoretical value (the theoretical value is also called the expected value) to explore whether different sample sizes come from the same whole[10].

First, the interaction table of glass surface weathering and decoration is shown in Table 1, and the expected frequency table of glass surface weathering and decoration is shown in Table 2. The chi-square value is calculated according to the following steps:

1) The hypothesis H_0 is established that glass ornamentation has no effect on weathering rate, that is, glass ornamentation is not related to weathering.

2) Determine the test level $\alpha=0.05$;

3) The test method was selected and the test statistic $\chi 2$ was calculated.

4) Determine the value of p and make an inference.

According to the chi-square statistical formula, the chi-square value is calculated by substituting the data in Table 1 and Table 2. The chi-square statistical formula is:

$$\chi^2 = \Sigma \frac{\left(f_0 - f_e\right)^2}{f_e} \tag{4}$$

In the formula, f_0 is the original statistic and f_e is the expected statistic. And plugging in the data gives:

$\chi = 4.9517$

The specified significance level α=0.05, if the obtained p-value < 0.05, then we consider the result represented by the sample is a small probability event, then we have reason to reject the null hypothesis.

Chaka square critical value table, distributed in degrees of freedom k=1, p=0.0 value of 3.84. We get the Chi-square value of 4.9517, which is greater than 3.84, so we can reject the null hypothesis, indicating that the surface weathering is related to the decoration.

Table 1: Interaction table of weathering and ornamentation on glass surface

Tattoo	Number		Total (pieces)	Weathering rate (%)	Unweathered rate (%)
	weathering	non-weathering			
A	11	11	22	50.00	50.00
B	6	0	6	100.00	0.00
C	17	13	30	56.67	43.33
Total	34	24	58	58.62	41.38

Table 2: Table of expected frequency of weathering and ornamentation on glass surface

Tattoo	Frequency		Total (pieces)
	weathering	non-weathering	
A	12.90	9.10	22
B	3.52	2.48	6
C	17.59	12.41	30
Total	34	24	58

Similarly, the relationship between glass surface weathering and color, and the relationship between glass surface weathering and glass type were respectively chi-square tests. In the test results, the chi-square values of glass surface weathering and glass type were 6.8749, and the chi-square values of glass surface weathering and glass color were 9.4619, and the null hypothesis could be rejected. It shows that the surface weathering is related to the type and color of glass.

3.2.2 Component Prediction

First define the variables as shown in Table 3:

Table 3: Define variable

Variable symbol	Implication
N	The number of unweathered samples of a type of glass
M	The number of samples of weathering of a particular type of glass
C	The amount of a chemical composition before weathering

The average content of chemical composition before weathering can be obtained $\overline{C}_0 = \frac{\sum_{n=1}^{n=N} C_n}{N}$, the average content of its chemical composition after weathering can be obtained $\overline{C} = \frac{\sum_{m=1}^{m=M} C_m}{M}$,The absolute value of the mean value of the change of the chemical composition before and after weathering is $\overline{b} = \overline{C}_0 - \overline{C}$, and the change rate of the chemical composition before and after weathering is $T = \frac{\overline{b}}{C_0} \times 100\%$.

1) Prediction of composition content of high potassium glass before weathering

Because there are few high-potassium glass samples in the samples, and the content of most chemical components in the weathering process fluctuates too much, it is not conducive to model establishment. Therefore, the predicted value Y_i can be obtained by reasonably subtracting or adding the absolute value $\bar{b}$ of the mean value of the chemical composition x_i of the sample before and after weathering:

$$Y = x \pm \bar{b} \tag{5}$$

2) Prediction of composition of lead-barium glass before weathering

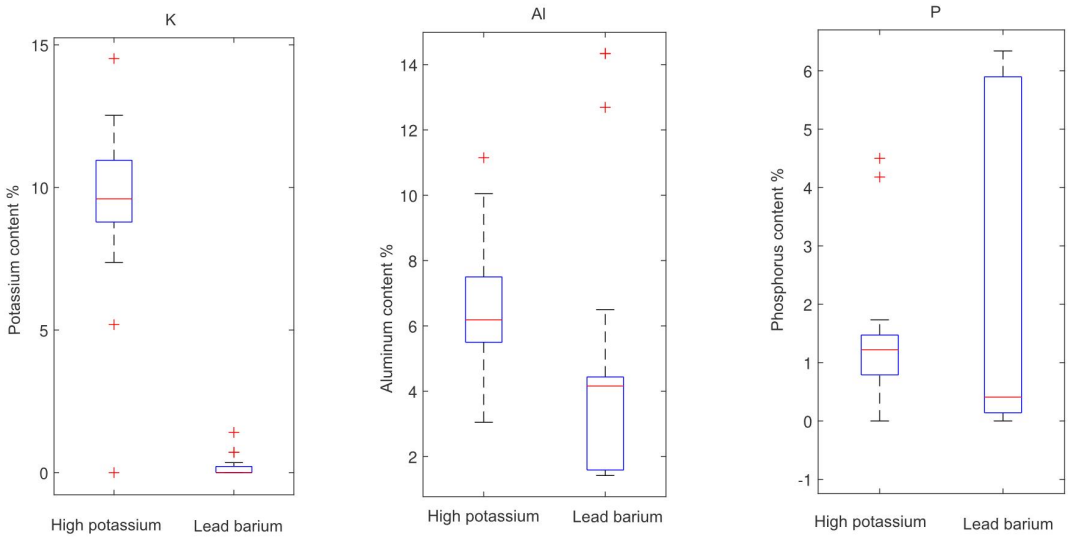
There are a large number of Pb barium glass samples, and most of the chemical components do not fluctuate much in the weathering process, so a correlation pre-weathering component prediction model is established. For the data processing after weathering of lead-barium glass, the predicted value Y_i is calculated by using the corresponding chemical composition change rate T_i before and after weathering of lead-barium glass, namely:

$$Y = \frac{x}{|T-1|} \tag{6}$$

Based on the validity of the data, the calculated component proportion will be accumulated and the data that is not between 85% and 100%, and the component proportion will be accumulated and expanded or reduced to 100% according to the corresponding proportion and equal proportion, so as to meet the requirements of data validity and make the prediction result reasonable. The prediction result will be taken as the prediction result.

3.2.3 Analysis of Variance

nce[11] tests whether the mean values of multiple normal populations with the same variance are equal by analyzing experimental data, and is used to judge the influence of various factors on experimental indicators. It is an effective method to identify the influence of relevant factors on experimental results. According to the chemical composition comparison of weathering or not of high-potassium glass and lead-barium glass respectively, Figure 1 retains the indicators with significant differences between the average content of the variance analysis in the two kinds of glass, which includes the chemical composition for K_2O, Al_2O_3, P_2O_5, BaO, SiO_2, PbO.



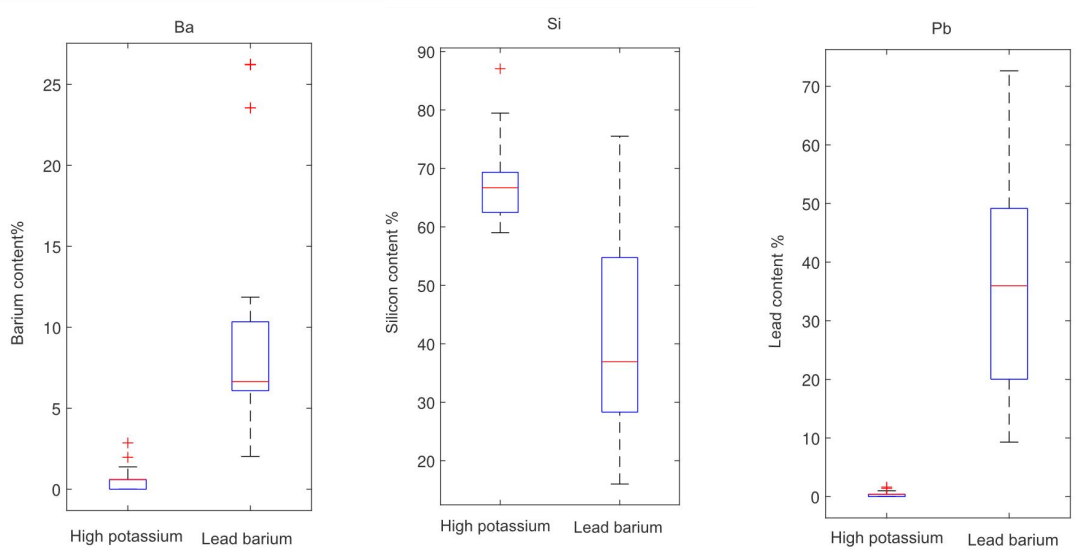


Figure 1: Analysis of variance

The maximum and minimum values of the differential component indexes of the two types of glass are used to describe the classification rules of high-potassium glass and lead-barium glass, as shown in Table 4:

Table 4: Classification standard for high potassium glass and lead barium glass

Category	K_2O	Al_2O_3	BaO	SiO_2	PbO
High potassium glass	7.37-12.53	3.05-10.05	0.0-1.38	59.01-79.46	0.0-1.41
Lead barium glass	0.0-0.71	1.42-6.50	2.03-11.86	16.03-75.51	9.30-72.63

3.2.4 Classification table of glass subclasses

Table 5: Eigenvalues and eigenvectors of the two principal components extracted after principal component analysis, as well as corresponding contribution rates and cumulative contribution rates. In this paper, two principal components $F1$ and $F2$ re extracted, namely:

$$\begin{aligned} F1 &= 0.37X_1 + 0.27X_2 + 0.37X_3 + 0.32X_4 + 0.32X_5 + 0.10X_6 \\ &+0.21X_7 - 0.08X_8 - 0.12X_9 - 0.14X_{10} - 0.06X_{11} \\ &-0.03X_{12} + 0.10X_{13} + 0.02X_{14} \end{aligned} \tag{7}$$

$$\begin{aligned} F2 &= 0.29X_1 + 0.19X_2 + 0.16X_3 - 0.05X_4 - 0.04X_5 - 0.04X_6 \\ &+0.07X_7 + 0.27X_8 - 0.33X_9 + 0.35X_{10} - 0.32X_{11} \\ &+0.32X_{12} - 0.01X_{13} + 0.15X_{14} \end{aligned} \tag{8}$$

Their cumulative contribution rate reached 72%. The first principal component $F1$ had moderate positive loads on SiO_2, K_2O, CaO, MgO and Na_2O, and light negative loads on P_2O_5 and BaO. Therefore, the first principal component can be called the high-potassium glass component.

The second principal component $F2$ has a moderate positive load on SiO_2, BaO, P_2O_5, SrO, CuO and PbO, so it can be called the second principal component of lead barium glass.

3.2.5 Clustering Result

According to the conclusion of principal component analysis, the first principal component is called high-potassium glass component, and the second principal component is called lead-barium glass component. Therefore, the components with large loads on the first principal component are selected: SiO_2, K_2O, CaO, MgO, Na_2O. The components of SiO_2, BaO, P_2O_5, SrO, CuO and PbO, which have a large load on the second principal component, are used as cluster indexes to divide the subclasses of high potassium and lead barium

glasses.

According to the cluster pedigree diagram, the samples of high-potassium glass can be divided into two categories, and the samples of lead barium glass can be divided into four categories. The results are shown in Table 5：

Table 5: Classification table of glass subclasses

High potassium glass	a	High copper potassium glass	1, 3, 7, 9, 10, 12, 13, 21
	b	High magnesium calcium potassium glass	4, 5, 6, 14, 16, 18, 22, 27
Lead barium glass	c	High silicon and high sodium lead barium glass	20, 22, 23, 25, 30, 35, 37, 42, 45, 46, 47, 53, 55
	d	High copper high strontium lead barium glass	8, 24, 26
	e	Low barium lead barium glass	28, 29, 30, 31, 32, 33, 44, 48, 49
	f	High phosphorus high lead barium glass	2, 10, 11, 19, 26, 34, 36, 38, 39, 40, 43, 49, 50, 51, 52, 54, 56, 57, 58

According to the clustering results of high-potassium glass and lead-barium glass, the mean value and standard deviation of each chemical component index of each subclass were calculated. The mean value was taken as the midpoint of the class, and the standard deviation of 2 times was taken as the deviation, and the component interval table of each subclass glass compound in Table 7 was formed. The sample data of high potassium glass and Pb barium glass are identified by fuzzy recognition algorithm, and compared with known classification labels, it is found that the accuracy of this model is higher.

Table 6: Subclass glass compounds composition interval table (XSD±2)

Category	High copper potassium glass	High magnesium calcium potassium glass	High silicon and high sodium lead barium glass	High copper high strontium lead barium glass	Low barium lead barium glass	High phosphorus high lead barium glass
CuO	3.18±2.69	1.73±2.00	1.60±2.81	8.46±0.00	0.41±0.47	0.84±1.75
PbO	0.48±0.86	0.34±0.96	22.34±15.73	36.38±11.32	19.65±15.10	50.15±18.52
BaO	0.80±1.85	0.39±0.98	10.54±8.25	26.23±0.00	4.41±4.65	7.41±7.28
P_2O_5	1.17±0.64	1.64±3.05	0.77±3.04	0.14±0.00	0.90±2.62	3.93±5.63
SrO	0.03±0.06	0.05±0.08	0.27±0.47	0.91±0.00	0.20±0.25	0.16±0.26
SiO_2	69.8±15.50	66.13±10.65	52.92±16.98	25.55±10.63	60.99±22.47	30.35±15.93
Na_2O	0.63±1.71	0.76±2.27	2.67±5.14	0.00±0.00	0.54±1.93	0.69±2.33
K_2O	8.64±7.32	10.02±4.27	0.17±0.39	0.00±0.00	0.35±0.77	0.04±0.31
CaO	5.34±3.31	5.32±6.03	0.85±2.09	0.47±0.00	2.08±2.40	2.21±2.43
MgO	0.75±0.83	1.41±0.93	0.59±1.09	0.00±0.00	0.89±1.04	0.36±0.63

3.2.6 Fuzzy Recognition Application and Model Verification

The fuzzy identification model was used to identify the chemical composition data of unknown glass cultural relics, and the variance analysis results were verified in Table 5. Meanwhile, the variance analysis results in Table 4 were used to classify these glass cultural relics. Among them, glass A1, A6 and A7 were high potassium glass, and glass A2, A3, A4, A5 and A8 were lead barium glass, which was consistent with the classification results in Table 6 .The accuracy of this model can be considered.

Table 7: Classification of unknown glass samples

A1	A2	A3	A4
High copper potassium glass	High phosphorus high lead barium glass	High copper high strontium lead barium glass	High copper high strontium lead barium glass
A5	A6	A7	A8

Low barium lead barium glass	High copper potassium glass	High copper potassium glass	High copper high strontium lead barium glass

In addition, 5 unknown samples were randomly selected, and their chemical composition content ±1% was analyzed for sensitivity. The results, as shown in Table 8 showed that the subclass division results remained unchanged.

Table 8: Results of chemical composition content +1% of samples from 5 unknown categories

SERIAL NUMBER	TYPE	SiO_2	Na_2O	K_2O	CaO	MgO	CuO	PbO	BaO	P_2O_5	SrO
1	a	77.67	0.00	0.00	6.02	1.84	2.09	0.00	0.00	1.05	0.03
2	f	39.31	0.00	0.00	4.20	0.50	0.00	35.90	10.25	1.40	0.48
3	d	31.63	0.00	1.35	7.12	0.80	0.21	39.18	4.64	2.65	0.51
4	d	35.12	0.00	0.78	2.86	1.04	0.95	24.04	8.23	8.37	0.28
5	e	68.60	0.00	0.26	1.33	0.99	0.33	17.06	4.00	1.03	0.12
1	a	79.23	0.00	0.00	6.14	1.88	2.13	0.00	0.00	1.07	0.03
2	f	40.11	0.00	0.00	4.28	0.52	0.00	36.62	10.45	1.42	0.48
3	d	32.27	0.00	1.37	7.26	0.82	0.21	39.98	4.74	2.71	0.53
4	d	35.82	0.00	0.80	2.92	1.06	0.97	24.52	8.39	8.53	0.28
5	e	69.99	0.00	0.26	1.35	1.01	0.33	17.41	4.08	1.05	0.12

4. CONCLUSION

The classification of cultural relics provides a key basis for archaeologists to implement protection and restoration measures. This paper makes full use of computer software Matlab and statistical software SPSS to explore the relationship between weathering of ancient glass relics and their colors, patterns and types from both qualitative and quantitative aspects, and effectively combines the analysis of variance and principal component analysis in mathematical statistics. K-MEANS clustering and fuzzy recognition methods are used to establish the subclass recognition model of ancient glass based on the subclass division quantization table. The results show that the classification recognition model has high stability and sensitivity. This study can be applied to the actual situation of the classification of unknown substances such as chemical experiments and archaeological research, and can also provide analytical methods for further improving the accuracy of ancient cultural relics classification.

ACKNOWLEDGMENT

This paper is supported by the fund: Training Program (S202310615243) and Open Experiment at Southwest Petroleum University (2022KSZ09006)

REFERENCES

[1] Wang Zhihao; Zhao Xiangwei; Li Zhiqun; Guo Ming; Xiao Wanyue; Liu Zhijian; .Prediction and subclassification method of weathering silicate glass based on machine learning [J]. Journal of the Chinese Academy of Ceramics, 2012, 23, 2:416-426.

[2] Tang Haijun, Zhu Fang, Yang Yunkai et al. Based on regression analysis of short YanZhi factors affecting research [J]. Journal of anhui agriculture tong, 2023, 29 (11) : 123-126. The DOI: 10.16377 / j.carol carroll nki issn1007-7731.2023.11.023.

[3] Wang Yichun, DENG Yuefang, Huang Wei et al. Research and countermeasure analysis on the practice status of municipal waste classification based on generalized linear model -- A case study of Nanjing City [J]. China Collective Economy, 2022(22):89-92.

[4] Cao Jiajia, Niu Bo, ZHANG Mingjin. Research on bulletproof performance prediction model of glass fiber reinforced composite based on improved neural network [J]. Journal of Ordnance Engineering, 2019, 44(07):163-169.

[5] Zhang Jiawei, Guo Linming, Yang Xiaomei. Oversampling and Random forest improvement Algorithm for Unbalanced Data [J]. Computer Engineering and Applications, 2019, 56(11):39-45. (in Chinese)

[6] Li Qing-Lin, Xu Cheng-tai, Wang Hai-hai, et al. Scientific analysis of ancient glass unearthed from Yangzhai Site in Yu County, Henan Province [J]. Archaeology and Cultural Relics, 2011(4):105-110.

[7] Tian Rui, Wang Fengyan, Sun Shangde et al. Evaluation of frying quality of vegetable oil based on Principal component analysis [J/OL]. China Oils and Fats :1-14[2023-08-16].https://doi.org/10.19902/j.cnki.zgyz.1003-7969.230251.
[8] Li Qingchang, Lin Beisen, LIU Liping, Luo Anna, Yang Guotao, CAI Bin, GAO Huajun, YE Changwen, LV Hongkun. Cluster analysis of chemical components of cigar tobacco leaves at different harvesting periods and its application in grading [J]. Tobacco Science and Technology, 2022, 55(06):35-41. (in Chinese)
[9] Chen Bei-Fei, ZHU Peng-li, Chen Chao. Classification and fuzzy recognition of physical fitness Level of college students [J]. Journal of Fujian Agriculture and Forestry University, 1992(02):237-240.
[10] Zhang Huayu, Zhu Huiping. Study on the technical characteristics of the first four boards of Chinese excellent table tennis player Ma Long: A case study of Chi square test [J]. Sports Science and Technology Literature Bulletin, 2019, 31(05):19-24.DOI:10.19379/j.cnki.issn.1005-0256.2023.05.006.
[11] Wang Haoya, Zhang Qiang, Dong Gaofeng, WANG Limin, He Zhijun, Xiang Ming. Application of MATLAB in chemical composition difference analysis of different flue-cured tobacco varieties [J]. Hubei Agricultural Sciences, 2011, 50(1):165-168.

Author Profile

Wang Jie, an undergraduate student at Southwest Petroleum University. Research direction: Computational Mathematics

Yu Ting, Master's degree, research direction: Fuzzy Mathematics and Its Applications.

Wang Qin, an undergraduate student at Southwest Petroleum University. Research direction: Theory and Technology of Oil and Gas Transportation and Storage.

Gu Angxuan is an undergraduate student at Southwest Petroleum University. Research direction: Theory and Technology of Oil and Gas Transportation and Storage.