Classification and Identification of Ancient Glass Objects based on K-means++ and Fisher Discriminant

Yuchao Du [1,#], Jie Zhou [2,#,*], Jingxuan Li [2,#]

[1] School of School of Computer Science, China West Normal University, Nanchong, China, 637009

[2.]School of School of Mathematics & Information, China West Normal University, Nanchong China, 637009

* Corresponding Author Email: wllhdsdzj@163.com

[#]These authors contributed equally.

Abstract. The significance of this study is to classify and identify the types of ancient glass products according to their chemical composition. The data were selected from the proportion of chemical compositions that have been analyzed for ancient glass, and the best number of clusters (k) for the division was roughly determined using the elbow rule for the five chemical compositions of high potassium and lead barium, respectively, and brought into the k-means++ algorithm for cluster analysis, and then the final determination of k and the evaluation of the rationality of clustering were performed using the contour coefficient, and finally the Fisher discriminant analysis method based on variable meritocracy combined with eigenvalues, Wilke Lambda, and classification function coefficients to identify unknown categories of glass artifacts. The model used was analyzed and evaluated with good results, and the model is applicable to the classification of ancient glass artifacts and identification of the type to which they belong.

Keywords: K-means++, Fisher Discrimination, Chemical Composition of Ancient Glass.

1. Introduction

The main chemical composition of ancient glass, as a valuable physical evidence of early trade transactions, is silica [1], which is subject to change due to the weathering process. At present, the classification and identification of ancient glass mainly use isotope labeling method [2], however, the drawbacks such as the high instrumentation requirement for stable isotope determination and the small variety of stable isotopes that can be used as tracers, as well as the high price, make this method not very efficient. In this paper, based on the weathering of glass artifacts, a mathematical model is developed to classify and identify the chemical composition of different categories of glass artifacts.

2. Materials and Methods

2.1. Data acquisition and pre-processing

The data selected for this paper were obtained from the 2022 National Student Mathematical ModelingCproblem(http://www.mcm.edu.cn/html_cn/node/5267fe3e6a512bec793d71f2b2061497.html).

The preprocessing is divided into two steps in total in order to improve the accuracy of the experiment.

Step 1: The data whose component ratios may be cumulative and between 85% and 105% due to testing methods are considered as valid data. We performed an initial screening of the data in Form 2 and removed two sets of useless data, leaving 67 sets of valid data.

Step 2: The glass artifacts in the high potassium and lead-barium categories were classified by the presence or absence of weathering on the surface, and the four categories into which they were divided were counted using the entropy weight method combined with the optimization method of the topsis [3-5] model for the chemical composition content. After analyzing the results, it was found

that the difference between the entropy weights of the unweathered points of the weathered lead-barium class and the unweathered lead-barium class was very small, and it was decided to classify the unweathered points of the lead-barium class as the unweathered class.

2.2. Introduction to the method (advantages, principles)

2.2.1 k-means++ algorithm clustering analysis steps and advantages

K-means++ algorithm [6-7] is a division-based clustering algorithm, using the elbow rule to determine the optimal number of clusters (k) for the division of the five chemical components of high potassium and lead barium, respectively. k value is established, and then brought into the k-means++ algorithm for cluster analysis to obtain the results of the division of high potassium and lead barium glass artifacts. k-means++ clustering algorithm steps (for high potassium glass as an example, and the same for lead-barium glass) is as follows.

1) Enter the data set of the 5 chemical components of high potassium glass in the attached Form II.

$$W = \{x_1, x_2, x_3, \dots x_n\} \tag{1}$$

2) Randomly select a point from the dataset$W$ as the first cluster center, e.g. $\varphi_1 = x_1$.

3) Calculate the distance between the other sample points and the cluster centerφ_k in turn, noted as$L(x_i)$; the probability that the other sample points will be selected as the next cluster center is$\frac{L(x_i)^2}{\sum_{i=1}^{n} L(x_i)^2}$, using the roulette wheel method to select the next cluster center.

4) Repeat step 3 until$k$ clustering centroids are selected.

5) Calculate the Euclidean distance between each samplex_i and$k$ clustering centers, and assign them to the class corresponding to the clustering center with the smallest distance φ_k.

$$d(x, y) = \sqrt{(x_1 - \varphi_i)^2 + (x_2 - \varphi_i)^2 + \cdots (x_n - \varphi_i)^2} \quad (i = 1,2, \cdots k) \tag{2}$$

6) Recalculate the cluster centers for each category φ_k.

$$\varphi_k = \frac{1}{|\varphi_k|} \sum_{x_i \in \varphi_k} x_i \tag{3}$$

Repeat steps (5) and (6) until the clustering centerφ_k no longer changes, then stop the cycle.

Pros.

1) K-means clustering algorithm is widely used in the field of cluster analysis, and has good classification classification effect, which can meet most of the classification requirements.

2) The elbow rule is used to determine the K-value of K-means++, which reflects the optimization and improvement of the K-means++ algorithm and improves the performance of the algorithm.

3) The fast nature of the K-means++ algorithm can be used to handle big data, making the model available for big data analysis.

2.2.2 Introduction to the principle and advantages of Fisher's discriminant method

Fisher's discriminant method [8-10] is based on the classification criteria of existing data, first classify the research object, then further select a number of variables that can provide a comprehensive description of the research object, then according to certain discriminant criteria, combined with eigenvalues, Wilke Lambda, classification function coefficients, identify and establish one or more discriminant functions, use the discriminant function to go to the unknown category of samples The discriminant function is then used to classify samples of unknown categories.

Pros.

1) Fisher's discriminant method reduces the dimensionality of the data, removes redundant variables to improve the accuracy of the algorithm and saves the training computation time of the model.

2) Fisher's discriminant analysis method can be widely used in medical, chemical, food and production fields because of its simplicity and high accuracy of ideas.

2.2.3 Flow chart of Fisher's discriminant method based on variable meritocracy

The flow chart of Fisher's discriminant method is shown in Figure 1.

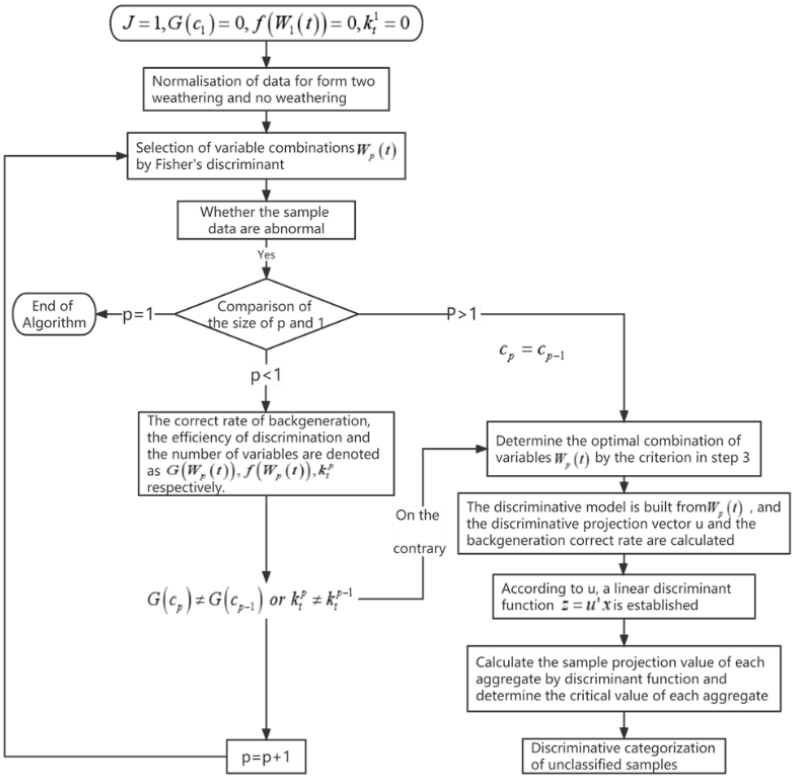

Figure 1. Flow chart of Fisher's discriminant method

2.3. Model evaluation metrics

2.3.1 Sensitivity

Sensitivity is used to test the stability of the model by floating all data up and down continuously by 5% at the same time, and then substituting the data into the model. If the obtained results do not change much in the floating of -5% to 5%, it means that the model is more sensitive and more stable; on the contrary, it means that the model is less sensitive and unstable.

2.3.2 Contour coefficient

The profile coefficient combines the cohesiveness and separation of clusters and is used to evaluate the clustering effectiveness. It is an evaluation of the effectiveness of the division results for the K-means++ clustering algorithm. The value of the contour coefficient is between [-1,1], and the larger the value of the contour coefficient, the better the clustering effect; conversely, the smaller the value of the contour coefficient, the worse the clustering effect.

2.3.3 Wilke Lambda

The Wilke Lambda value is used to test the influence of Fisher's discriminant function. The smaller the Wilke Lambda value, the greater the influence; conversely, the larger the Wilke Lambda value, the smaller the influence.

2.3.4 Significance

When the significance of this discriminant function is lower than 0.05, it means that the significance level of this discriminant function is high and the corresponding discriminant ability is stronger; on the contrary, when the significance of this discriminant function is higher than 0.05, it means that the significance level of this discriminant function is low1 and the corresponding discriminant ability is weaker.

3. Results and Analysis

3.1. Calculating the number of clusters using the elbow rule

All the data of high potassium and lead-barium from Form II were imported into python separately and two distortion curves were generated as shown in Figure 2.

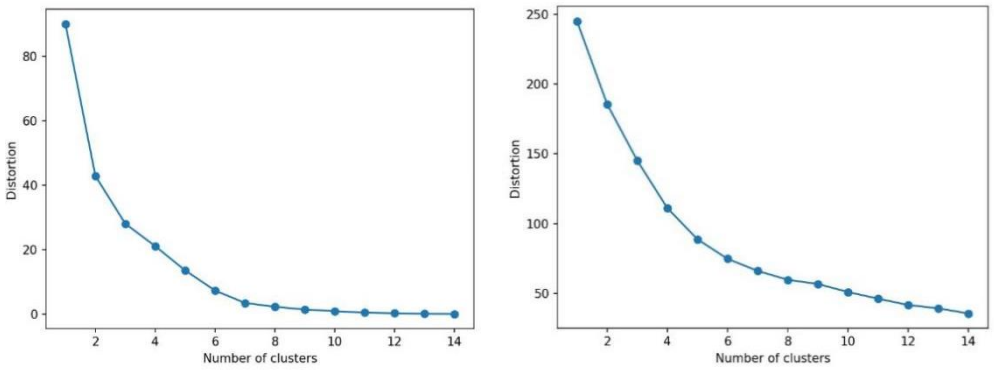

Figure 2. distortion curve graph

The left graph shows the aberration curve for the high potassium class, and the right graph shows the aberration curve for the lead-barium class. Since the value of k changes abruptly when it is close to the true number of clusters, it can be seen from the graph that k=2 or k=3 is the "elbow" of the total aberration, so 2 or 3 is the best number of clusters to divide the subclasses of high potassium. From the right figure, we can see that k=3 or k=4 is the "elbow" of the total aberration, so 3 or 4 is the best number of clusters for the subclass of lead and barium.

3.2. Clustering classification results

1) K-means++ clustering algorithm clustering division.

The K-means++ clustering algorithm was used to calculate the initial cluster centers, the final cluster centers, and the distances between the final cluster centers of the five chemical components by spss. Only the final clustering classification results are discussed. For the high potassium class artifacts, the first class contains all weathered and two non-weathered class artifacts for k=2. For k=3, the first class contains all weathered and one unweathered artifacts, the second class contains all weathered and one unweathered artifact, and the third class contains all unweathered artifacts.

According to Table 1, the analysis knows that the categories classified with the K-means++ clustering algorithm consider the presence or absence of weathering as the basis for classification.

Table 1. Specific division of high potassium class relics

	k=2		k=3		
	1	2	1	2	3
Artifact Number	03	01	01	07	18
	07	03	03	09	21
	09	04	04	10	
	10	05	05	12	
	12	06	06	22	
	18	06	06	27	
	21	13	13	03	
	22	14	14		
	27	16	16		

For the lead-barium class artifacts, according to the above table, the analysis knows that the categories classified by K-means++ clustering algorithm also consider unweathered, normal weathering, and severe weathering as the basis for classification.

Table 2. Specific division of lead and barium artifacts

	k=3			k=4			
	1	2	3	1	2	3	4
Artifact Number	02	11	08	02	11	39	08
	19	20	08	19	20	40	08
	30	23	24	25	23	43	24
	30	25	26	30	28	51	26
	34	28	26	30	29	54	26
	36	29		34	31	54	
	38	31		36	32		
	39	32		38	33		
	40	33		41	35		
	41	35		43	37		
	43	37		49	42		
	43	42		50	42		
	49	42		50	44		
	50	44		51	45		
	51	45		52	46		
	51	46		55	47		
	52	47		56	48		
	54	48		57	49		
	54	49		58	53		
	56	50					
	57	53					
	58	55					

2) Analysis of the reasonableness of the contour coefficient on the results.

There are 18 artifact numbers in the high potassium category, and the distribution of high potassium glass contour coefficients is shown in Figure 3, using the number of classifications as the horizontal coordinate and the contour coefficient corresponding to that classification as the vertical coordinate.

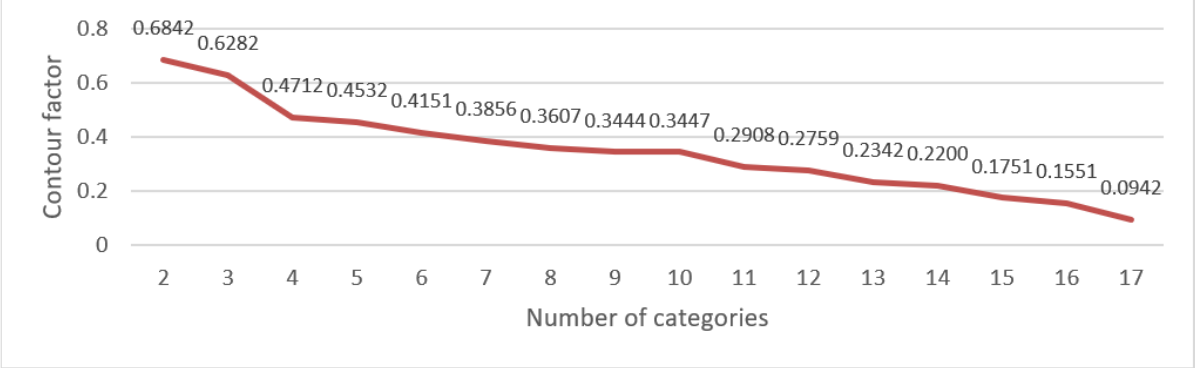

Figure 3. Distribution of contour coefficients of high potassium glass

Since the value of the contour coefficient is between [-1,1], the larger the value of the contour coefficient, the better the clustering effect is indicated. From the above figure, we can see that the value of the contour coefficient is the largest when the number of classifications k=2, so 2 is the best

number of clusters for the high potassium class artifacts. The reasonableness of the model is better reflected.

There are 41 artifact numbers of lead-barium class artifacts, with the number of classification as the horizontal coordinate, the corresponding contour coefficient of that classification number as the vertical coordinate, to make the distribution of lead-barium glass contour coefficient as shown in Figure 4.

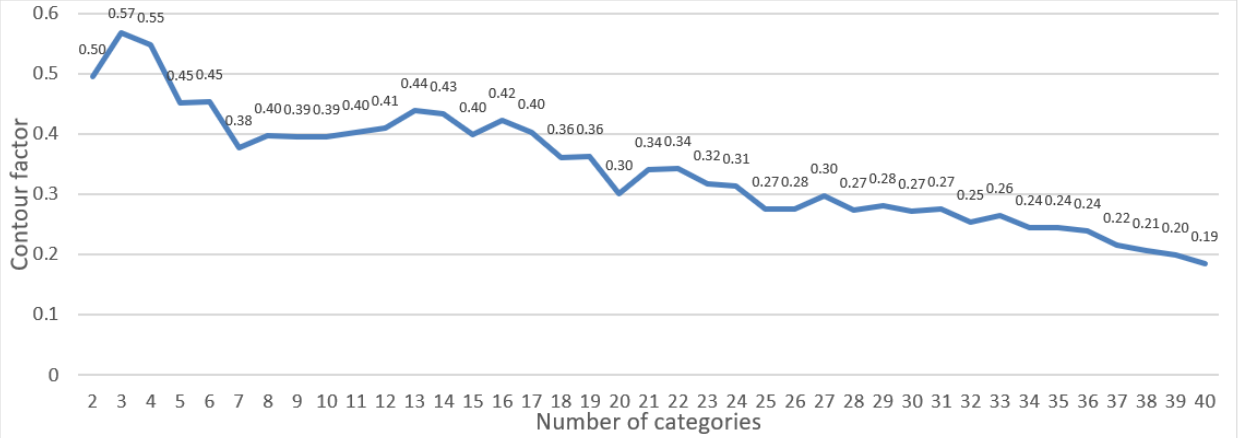

Figure 4. Profile coefficient distribution of lead-barium glass

According to the same criterion for judging the contour coefficient, we can see from the above figure that the value of the contour coefficient is the largest when the number of classifications k=3, so 3 is the best number of clusters for the lead barium class artifacts. The reasonableness of the model is better reflected.

3.3. Fisher discriminant analysis to establish the discriminant function

The data from Form II were classified as weathered or unweathered, and Form III was classified according to the same criteria. All 14 chemical constituents of weathered and unweathered Form II data were substituted into spss for Fisher's discriminant analysis based on variable merit, assuming high potassium type 1 and lead-barium type 2.

1) Fisher's discriminatory process for form II weathering of this category.

The summary table of the discriminant function of the canonical rule was obtained as shown in Table 3. From Table 3, the eigenvalue of the discriminant function is 58.394 It can be seen that the discriminant function has a good representation of the mean squared deviation. Wilke Lambda is 0.017 is smaller, the smaller the value, the greater the influence. The significance is 0, which is significantly smaller than 0.05, indicating that the discriminant function has a high level of significance and the corresponding discriminatory power is stronger.

Table 3. Weathering Fisher's discriminant characteristic values

Function	Eigenvalue	Wilke Lambda	Significance
1	58.394[a]	0.017	0.000

Note: a. The first 1 canonical discriminant function is used in the analysis.

Since the program only gets a judgment function, the coefficients of the classification function are used to determine the classification. The discriminant functions on class 1 (high potassium) and class 2 (lead barium) can be obtained at y_1, y_2 respectively.

$$y_1 = -355.835 + 812.817x_1 - 3918.238x_2 - 4730.314x_3 + 8283.505x_4 + 1030.623x_5 + 4312.281x_6 - 1878.046x_7$$

$$y_2 = -33.933 + 238.325x_1 - 999.689x_2 - 1487.26x_3 + 3180.071x_4 + 316.288x_5 + 1331.784x_6 - 506.163x_7$$

x_1 *is* SiO_2; x_2 *is* Na_2O; x_3 *is* K_2O; x_4 *is* SrO; x_5 *is* SO_2; x_6 *is* MgO; x_7 *is* Al_2O_3

2) Fisher's discriminations for the category of non-weathering.

The summary table of the canonical discriminant function is obtained as shown in Table 4. From Table 4, the eigenvalue of the discriminant function is 6.371 It can be seen that the discriminant function has a good representation of the mean squared deviation. Wilke Lambda is 0.136 is smaller, the smaller the value, the greater the influence. The significance is 0, which is significantly smaller than 0.05, indicating that the discriminant function has a high level of significance and the corresponding discriminant ability is stronger.

Table 4. Non-weathering Fisher's discriminant characteristic values

Function	Eigenvalue	Wilke Lambda	Significance
1	6.317a	0.136	0.000

Note: a. The first 1 canonical discriminant function is used in the analysis.

Since the program only gets one judgment function, the coefficients of the classification function are used to determine the classification. The discriminant functionsw_1 ,w_2 can be obtained for class 1 (high potassium) and class 2 (lead barium), respectively.

$$w_1 = -9.086 + 179.879a_1 + 0.345a_2$$
$$w_2 = -6.102 + 2.874a_1 + 48.951a_2$$

Judgment equation ina_1 is $K_2O; a_2$ is PbO .

3.4. Classification results from discriminant function

1) Fisher's discriminatory results for this category of form three weathering

The data corresponding to the four groups of weathered artifact numbers A2, A5, A6, A7 in Form 3 were substituted into the determination functionsy_1 ,y_2 as shown in Table 5.

Table 5. Judgment results of weathering

Number	y1	y2	Classification (1/2)
A2	-92.7551	44.2411	2
A5	-18.9441	75.0931	2
A6	318.1169	163.1394	1
A7	242.1942	142.6985	1

Comparing the larger value ofy_1 ,y_2 , the artifact belongs to the classification where the larger value is located. It is known that A2 and A5 belong to the lead-barium category, and A6 and A7 belong to the high-potassium category. (1 is high potassium, 2 is lead-barium).

2) For the form three no weathering this type of Fisher discriminant results.

The data corresponding to the four groups of weathered artifact numbers A1,A3,A4,A8 in Form 3 were substituted into the determination functionsw_1 ,w_2 as shown in Table 6.

Table 6. Judgment results of no weathering

Number	y1	y2	Classification (1/2)
A1	-9.086	-6.102	2
A3	-6.5031	13.3119	2
A4	-7.5812	5.806	2
A8	-8.599	4.3018	2

If you compare the larger value ofw_1 andw_2 , the artifact belongs to the classification where the larger value is located. It is known that A1, A3, A4, and A8 are all lead-barium classes. (1 is high potassium, 2 is lead-barium).

3.5. Sensitivity analysis of classification results

By floating all the data in Form III up and down continuously by 5% at the same time, and substituting the data classification into the discriminant function of weathering and no weathering, the classification results obtained are basically unchanged, so the sensitivity of the model is good.

4. Conclusions

In this paper, we classify ancient glass products and identify the types to which they belong according to their chemical compositions. In terms of classification, the elbow rule is used to make distortion graphs for the five chemical compositions of high potassium and lead barium respectively, and the best clustering number k values for the division of high potassium and lead barium subclasses are roughly derived, which is optimized and improved for the subsequent algorithm. k values are brought into the k-means++ algorithm for clustering analysis, which can quickly and accurately solve the specific classification of the artifacts, and finally the contour coefficient is used to finally determine the k values and conduct The reasonableness analysis of the clustering concluded that the model is better. Fisher's discriminant analysis can be used to identify the type of glass by combining the eigenvalues, Wilke's Lambda, and the coefficients of the classification function, and can be widely used in medical, chemical, food, and production fields because of its simplicity and accuracy. In summary, this thesis makes a reasonable answer in subclassifying and identifying glass types by establishing a mathematical model and using intelligent algorithms starting from the chemical composition of glass to make some contributions to contemporary research on glass weathering prevention, application and digitization.

References

[1] Weizhao Wang,a Qishan Huangb, Fei Lic, Donghong Gu c and Fuxi Gan c a Museum of Hepu, Hepu, 536100, P.R. China. b Museum of Guangxi, Nanning, 530022, P.R. China. China. c Shanghai Institute of Optics and Fine Mechanics, the Chinese Academy of Sciences, Shanghai, 201800, P.R. China. A Study Of The Ancient Glass Of Han Dynasty Unearthed In Hepu County, Guangxi Province[C].

[2] Dussubieux Laure, Fenn Thomas R., Abraham Shinu Anna, Kanungo Alok Kumar. Tracking ancient glass production in India: elemental and isotopic analysis of raw materials[J]. Archaeological and Anthropological Sciences,2022,14(12).

[3] Zelin Su,Huiqi Zhang,Ziqing Zhao. The Evaluation Model of Suppliers Based on TOPSIS and Entropy Method[J]. Information Systems and Economics,2021,2(2).

[4] Wu Hua-Wen,Li En-qun,Sun Yuan-yun,Dong Bao-tian.Research on the operation safety evaluation of urban rail stations based on the improved TOPSIS method and entropy weight method[J]. Journal of Rail Transport Planning & Management,2021,20.

[5] XIANGYU LI,Jie SHEN. Evaluation of the Development of National Higher Education System Based on TOPSIS and Entropy Weight Method[C]//. Proceedings of the 2021 International Conference on Applied Mathematics, Modeling and Computer Simulation (AMMCS 2021). Proceedings of the 2021 International Conference on Applied Mathematics, Modeling and Computer Simulation (AMMCS 2021). ,2021:643-652. doi: 10.26914/c.cnkihy.2021.069165.

[6] Shruti Aggarwal,Paramvir Singh. Cuckoo, Bat and Krill Herd based k-means++ clustering algorithms[J]. Cluster Computing,2019,22(6).

[7] Fernando Arce,Erik Zamora,Carolina Fócil-Arias,Humberto Sossa. Dendrite ellipsoidal neurons based on k-means optimization[J]. Evolving Systems,2019,10(3).

[8] Merlin Murilo,Pinto Allan,de Almeida Alexandre Gomes,Moura Felipe A,Da Silva Torres Ricardo,Cunha Sergio Augusto. classification and determinants of passing difficulty in soccer: a multivariate approach[J]. Science and Medicine in Football,2022,6(4).

[9] Ping Jian. Rough Set-Fisher Discriminant Analysis Method for Prediction of Classification of Rock Burst Risk [J]. Journal of Physics: Conference Series,2021,1995(1).

[10] P. N. Senthil Prakash,N. Rajkumar. improved local fisher discriminant analysis based dimensionality reduction for cancer disease prediction [J].. Journal of Ambient Intelligence and Humanized Computing,2020,12(prepublish).