Component Analysis and Identification of Glass Products Based on Hierarchical Clustering and Naive Bayes

Linna Tang *,†, Mingjun Tang †, Lingji Zhang †

Information Engineering University, Zhengzhou, China

* Corresponding Author Email: 874890724@qq.com

† These authors contributed equally.

Abstract. In the process of weathering of ancient glass products, the interaction of internal elements and environmental elements will directly lead to the change of their composition proportion, thus affecting the judgment of the composition. In this paper, cultural relics were analyzed by chi-square test and spearman coefficient, and descriptive data analysis was used to process and predict the composition content of chemical data after weathering, and to preliminarily classify them with nuclear naive Bayesian method. The contour coefficient S was used to find the best K value, and the K-Means cluster analysis was used to select the chemical indicators of whether the different types of glass were weathered or not in 4 categories. Therefore, the best classification results were obtained respectively, and the relationship and difference between chemical components were obtained by Pearson correlation coefficient and principal component analysis. From the perspective of data analysis, this paper conducts statistical analysis for the data distribution law, uses K-Means and naive Bayes to divide the subclasses, and establishes the intelligent divisions of three different machine learning models after voting on the data. By combining the correlation analysis and the second reference number, the largest difference is the correlation of the chemical composition between the weathered high potassium according to the different European distance of the weathered high potassium, and the smallest difference is the correlation between the weathered lead barium and the unweathered lead barium.

Keywords: Kappa test machine learning, K-means decision tree classification, naive Bayesian path analysis.

1. Introduction

The main composition of pure quartz sand is (sio2), which is the main raw material of refining glass. Due to its high melting point, it needs to add auxiliary solvent when refining to reduce the melting temperature [1-2]. Different cosolvents added in the refining process will bring about differences in the chemical composition to form different categories of glass products. It is known that in the process of weathering, the ancient glass will change its composition proportion due to the large number of interactions between the internal elements and the environmental elements, affecting the judgment of the category [3-5]. So need to through the analysis of the chemical composition data statistical relationship, significance, correlation, differences between data establish classification standard to achieve the identification of weathered cultural relics, and to identify the results of the feasibility and sensitivity analysis, in order to achieve intelligent identification of weathered cultural relics, improve the efficiency of cultural relics detection important value and practical significance.

This paper needs to establish a model to solve the following problems: (1) for the basic information of glass cultural relics given in attached table 1, discuss the relevance of the other three elements for surface weathering. Combined with Schedules 1 and Table 2, the proportion of chemical composition before and after weathering of two different glass products analyzed statistical rules, and the content of each chemical composition before weathering was predicted. (2) Further analysis of the chemical composition content before and after weathering for subclassification, and the reliability analysis of the classification results based on this [6-7]. (3) Establish classification standards based on the division of the second question subcategories to intelligently divide the data of attached Table 3, and analyze the rationality and sensitivity of the division. (4) Based on the previous analysis, different types of chemical components were analyzed.

2. Problem analysis

Problem 1 requires that the weathering of the glass surface should be discussed as the relationship between ornamentation, color and glass type, and the statistical changes of the chemical composition elements of the classified glass before and after weathering are given. We need to convert nominal variables into the actual data variable, the need to note that the sample of serious weathering data, due to the difference greater influence law statistical analysis, the problem also emphasizes the scope of the total, therefore, using the asymptotic significance of Pearson as a standard, analysis to weathering as the target variable decoration, color, type, here need to note that the square distribution because of the data contingency error makes the value unreliable, so need to use Kappa consistency test for further analysis [8].

Question 2 let us analyze the law of different categories of glass and the classification of categories, combined with the first question CV value processing abnormal data, with a box plot filter out a data, through the kernel naive Bayesian classification method for machine data in table 2 learning and divided into two categories, when learning only consider two glass categories under the division of the data. The contour coefficient (silhouette coefficient) is used to select the k worthy of the best classification cluster, based on the cluster analysis (K-Means) combined with the chemical composition data in Schedule 2 to give the classification clusters before and after weathering of different glass species, combined with the F test results and the significance of significance. Add a perturbation data size to each chemical element and an additional perturbation (±1%, ±2%, ±5%, ±10%, ±20%, ±30%, ±40%) to obtain the new classification data, and the sensitivity of the subclass division is obtained by the comparison of before and after the change types [9].

Based on the classification clusters of different glass types given in question 2, using the decision tree classification, naive Bayes classification and gradient descent method, take most of the above machine learning model as the final classification result by voting method, add the data size increase and decrease perturbation (±1%, ±2%, ±5%, ±5%, 10%, ±10%, ± 20%, ±30%, ±40%) to each chemical element and analyze the sensitivity [10].

Question 4 requires us to add 5 experiments to the existing experiment, and the model obtained from the first three questions predicts the possible 5 possible optimal cases. We will again use the particle swarm algorithm through the regression model of the second question, and find out the experimental conditions of the optimal 5 sets of parameters for the additional 5 experiments in the current search space, and give the corresponding explanation. Problem four requires us to analyze the correlation and difference of different types of chemical components. We need to use normal analysis to obtain the distribution law of the data first[11]. If it conforms, Pearson correlation coefficient analysis is used to analyze the correlation. The general number error matrix is established based on the correlation coefficient, and the calculated distance is used to analyze the difference of the data correlation relationship of various glass products.

3. Model assumptions

Suppose that the chemical composition content in the blank in Annex 2 is 0. 2. Suppose that the sampling points of severe weathering have a strong impact on the statistical laws. 3. 4 Suppose that the variables did not influence the outcome analysis except the data recorded in the experiment.5. Suppose that the significance of the cultural relics given in Schedule II is generally representative of the prediction.

4. Model construction and solution

4.1. Problem 1 model building and solution

The questions are divided into three questions. By converting the variables in Schedule 1 and analyzing the relationship between weathering and other variables, combining the glass types and weathering in Schedule I and Table 2 and describing the data, and using statistical laws to predict the data before weathering. Before analyzing the relationship between the data in Table 1, it can be known that the four type variables are all discrete type variables, so the correlation analysis can first describe the correlation between weathering, type, pattern and color to observe the general correlation trend.

According to the mathematical statistics and analytical knowledge, because the discrete type variables are difficult to have a good linear relationship, the Pearson coefficient cannot measure the correlation. However, the Spearman correlation coefficient measures the correlation between the two noncontinuous variables that do not obey the normal distribution, and only judges its monotonicity.

The deeper the red color of the heat map, the greater the relationship. If the significance is presented, after the relationship, the closeness of the relationship can directly depend on the size of the correlation coefficient. Generally, above 0.7, the relationship is very close; between 0.4~0.7; and between 0.2~0.4, the relationship is general. If the correlation coefficient value is less than 0.2, but it is still significant, indicating that the relationship is weak, but there is still a correlation. From the correlation table and heat map, we can intuitively see the color and pattern, and the type has a strong correlation with whether weathering is not. Therefore, glass type has a positive correlation with weathering, that is, lead-barium type glass is easy to weathering, and high potassium type is not easy to weathering.

Consistent with the observations in the above statistical figure, the influence of pattern type on weathering is not reflected. The guess is the accidental error caused by the high overlap between type B and lead barium type and the weathering of all type B type. The influence of decorative type on weathering is not reflected. It is speculated that the accidental error is caused by the heavy height of B type and lead barium sample, and the weathering of lead barium type leads to the weathering of all type B.

Using the non-parametric tests in the difference analysis, the chi-square test was used for classification variables, consistent with the data in this question. The comparison between type, decoration, color and weathering is tested respectively: Kappa consistency test is usually used for the consistency test of classification data, and weathering, type, decoration and color are all classification data, so the correlation between the four types of data can be verified through this test.

The Kappa value is visualized by heat map, and the value is represented mainly represented by the color depth in Figure 1.

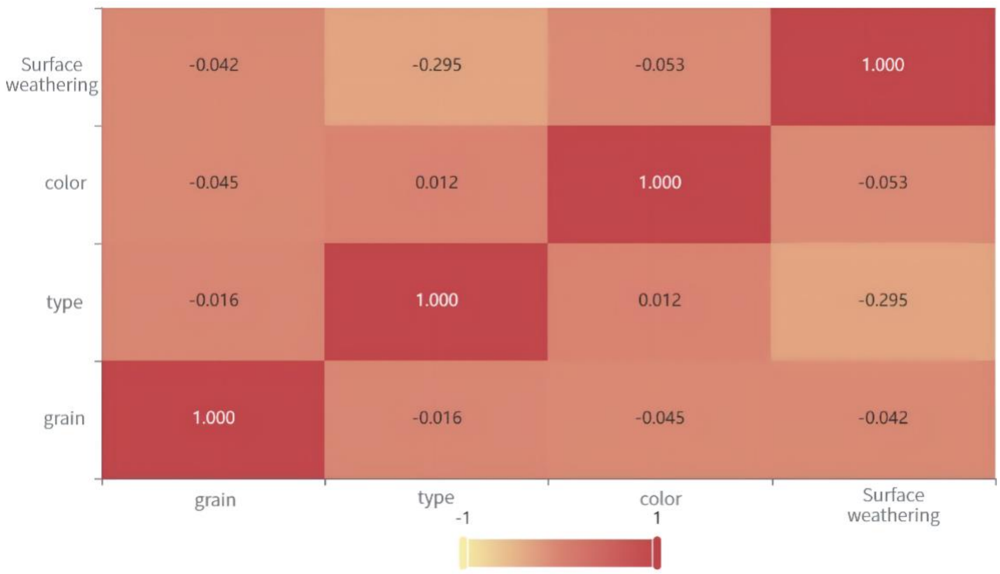

Figure 1. Thermal map of the Kappa correlation coefficient matrix

The relationship between type and surface weathering was analyzed by Kappa test with good significance. The Kappa values fall between 0.2 and 0.4, at a low consistency, so it can be determined that the type is positively correlated with the surface weathering presentation. Other elements have had little effect on the surface weathering.

Anomalous data was identified by further boxplots of each chemical data component before and after weathering for different glass types, showing silicon dioxide (SiO_2), sodium oxide (Na_2O), potassium oxide (K_2O), oxide Calcium (CaO), magnesium oxide (MgO), aluminum oxide (Al_2O_3), iron oxide (Fe_2O_3), copper oxide (CuO), lead oxide (PbO), barium oxide (BaO), phosphorus pentoxide (P_2O_5), oxide Strontium (SrO), tin oxide (SnO_2), sulfur dioxide (SO_2) frequency analysis The results of discrete trend analysis, the discrete trend uses the maximum value, minimum value, 25% quantile, median, 75% quantile, etc. Statistical metrics measure the variance (stability) of data distributions. The maximum value and minimum value are not the maximum and minimum values of the data, but the inner limit of the boxplot, that is, the points larger than the maximum value or smaller than the minimum value are regarded as abnormal points and eliminated.

It can be seen from the table that the content variance of lead-barium silica, lead oxide and barium oxide is large, and the data is volatile, while the content content of silica and potassium oxide is large. Explain through the average of weathering before and after 9 analysis data, to screen some abnormal data, including the influence of the greatest degree of weathering, severe wind, silica content is very little, the content of lead and barium will also change, in order to prevent the fluctuation of data to research, learn the influence of weathering, using partial correlation analysis, only to find the correlation trend between the two, so as to reduce the influence of data fluctuations.

Partial correlation analysis is based on the correlation analysis of the interference factors, and the control variables are generally related to the correlation analysis. According to the above statistics, it is found that the chemical composition relationship of high potassium and barium lead is very different, so when analyzing the data, the high potassium and barium lead are separated to find the relationship between the chemical composition and weathering respectively. Since there may be a mutual influence between each chemical component, a partial correlation analysis was used, using each chemical component as a control variable, to study the correlation between the pairwise, that is, the relationship between a chemical component and weathering.

It can be seen: in the high potassium type, silica content and weathering show positive correlation, other components are negative correlation, due to the significance of <0.05 can be considered to have obvious correlation, by reading the figure, silica, magnesium oxide, potassium oxide, calcium oxide, alumina, iron oxide and whether weathering show different size correlation. In the case of lead-barium, the content of silica is negatively correlated with whether weathering is not, and the lead oxide is positively correlated. By reading the map, we can determine the correlation between silica, nano oxide, lead oxide, calcium oxide, phosphorus pentoxide, strontium oxide and whether weathering shows different sizes.

Can be found that in addition to the composition of lead oxide, the rest of the component prediction is more accurate, analysis because there may be some correlation between composition and composition, the results of a single component prediction in some associated components will be inaccurate, secondly, for phosphorus pentoxide, although the variance is not big, but the amount of composition difference is very big, here because the volatility of phosphorus pentoxide is larger, the variance, the use of change rate can not well depict its relationship before and after weathering. In addition, we also make the thermodynamic statistical maps, see Appendix 4.

4.2. Problem 2 model building and solution

According to the requirements of this topic, the high potassium glass, lead barium glass chemical composition for the influence of classification law, need to "category" column character data scalar transformation, to ensure the consistency of the operation model, ensure that the prediction results are not affected by different dimensions and dimensional units of strange sample interference, respectively using "0", "1" floating point type on behalf of two glass types, and the cultural relics sampling point data part formatting processing.

Beyond the threshold data was removed according to the threshold value dividing the proportion of effective data of chemical composition. The ation of each line of data calculated the ratio of components, excluding 100% due to detection means and other reasons.

Different chemical composition indicators tend to have different dimensions and unit of dimension benchmarks, which can affect the results of data analysis. For example, during the machine training gradient descent, the objective function becomes "flat". For this reason, the direction of the gradient deviates from the minimum value, and the path length increases, that is, the training time is too long, and the fitting is poor. Therefore, the data accuracy is improved by the normalization of the chemical composition in the following equation.

Since it is known to classify the data into high potassium glass and lead barium glass, it is more suitable for Bayesian classification scenarios. At the same time, combined with the comprehensive accuracy of the whole 16 machine learning classification models, the Bayesian classification accuracy is as high as 100%, so the Gaussian nuclear Bayesian distribution is selected as the target model to summarize the classification rules of high-potassium glass and lead-barium glass.

The ROC curve is a line graph established with specificity and false positives rate as the X axis and specificity (sensitivity) as the Y axis. Area prediction is the accurate value, called the AUC value. The larger the AUC means that the higher the prediction accuracy, and the lower the prediction accuracy. As shown Figure 2.

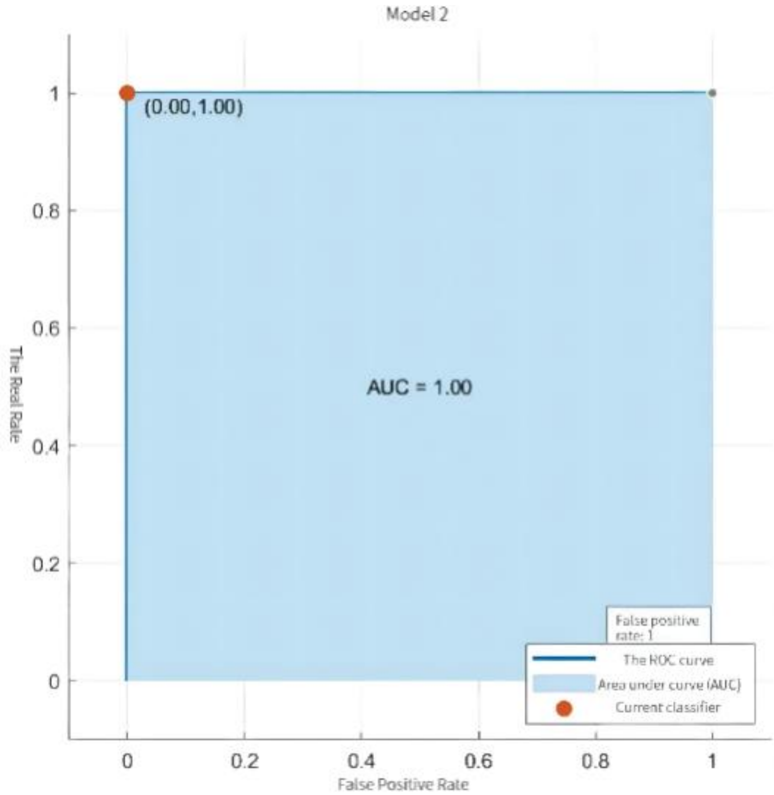

Figure 2. ROC Curves

ANOVA is a variance analysis method, calculate the importance score of each chemical component such as list judge 14 break the importance of chemical elements, from the importance score, the classification law of high potassium glass and lead barium glass, priority of importance is more than 10 (marked red), these different input categories have significant difference in output variables, so the distribution pattern of high potassium glass and lead barium glass is strongly correlated with K2O, PbO, CaO, BaO.

K-mean clustering is a data classification algorithm that divides the data into K groups set in advance, and then K objects are randomly selected as the initial clustering center, and finally obtains the best results. The data entropy was reduced by comparing the content of chemical components in the table as "weathered high potassium", "unweathered high potassium", "weathered lead barium" and "unweathered lead barium", to improve the credibility and rationality of the subsequent subcategories. The 14 chemical components should be considered as 14-dimensional vectors, with the Euclidean distance calculated from the Euclidean algorithm.

For example, in weathered high potassium glass, the content of magnesium oxide and alumina is significantly different, so it is selected for subclass division, which is the same way.

Matlab programming was used to build the perturbation matrix with the same size as the original data, and add the two to bring the results back to the K-Means classification model, and calculate the Euclidean distance from the coordinates of the cluster center point. The interference samples were summarized into the subclass of short center points, and compared with the original subclass to calculate the accuracy of the model, setting the disturbance size change of 1%, 2%, 5%, 10%, 20%, respectively, so that when the disturbance range is 20%, the accuracy of the model is reduced to 93%.

According to the table data, small perturbation are added to the original data. If the judgment result does not change, it shows good anti-interference, the sensitivity is low and the model is robust. When the disturbance increases to 20% of the original data, the accuracy of the judgment drops to 96% and 93% after the disturbance, indicating that the model cannot accept the noise interference of 20%, and can accept the noise interference within 10%, reflecting the sensitivity of the model.

5. Problem 3 model building and solution

Problem 3 needs to analyze the unknown relics data in sensitivity Table 3, combine the data classification model of hierarchical clustering in problem 2, predict the category of unknown relics, and conduct the sensitivity test based on the disturbance.

The model of subclass discrimination is based on Euclidean distance, to play the advantages of automatic fitting of machine learning models, and to avoid overfitting, underfitting and noise influence due to less data. Using decision tree classification, gradient descent, naive Bayesian classification three machine learning classifier model practical training, respectively in table 3, data classification after the sensitivity of table 2 data as a training set, finally the three models of sensitivity table three data evaluation results voting method, reduce the negative impact of machine learning, make the results more credible.

Decision tree is a predictive model, classifying unidentified instances, and also a descriptive model. Decision tree classifier is based on information entropy learning. The essence of decision tree learning is to summarize a set of classification rules from the training set, judge whether or not according to the data threshold, and then give the answer to the question. Therefore, the interpretable nature of the decision tree classifiers is relatively good.

In the heat map of the data confusion matrix, the abscissa is the predicted value and the ordinate is the true value, the darker the color of the thermal block located in the diagonal line, the more correct the test data is. In the decision tree classification model, the rest of the test set is located in the diagonal line, which improves the credibility of the decision tree classification model.

The Naive Bayesian algorithm is an algorithm that determines the category of the new sample according to the conditional probability (posterior probability) of the existing characteristics of the new sample. It determines the sample as the class with the greatest posterior probability. Relevant model building parameters have been given in Question 2, where the model is called again to detect the dataset.

Combined with the naive Bayesian model in the second question, the basic types of unknown cultural relics (high potassium or lead and barium) are pre-divided first, which is conducive to the corresponding four categories of classification models for reliable prediction.

6. Problem 4 model building and solution

In the second question, the two major categories are subclassified. After the validated data, some subcategories have only one sample, which obviously can not find relationships between their components, so large categories are still used to classify categories. The data were divided into four categories: weathered lead barium, unweathered lead barium, weathered high potassium and unweathered high potassium. Large changes will have an impact on the correlation between the components.

Because the Pearson correlation analysis needs to ensure that the input data conforms to the normal relationship and is a random continuous variable, the content and frequency of the number of chemical components. The figure below shows several of these categories.

Based on the analysis of the third question, different glass products have different internal correlation before and after weathering, so the correlation is divided before and after differentiation of flavor lead-barium glass and high potassium glass.

According to the correlation analysis, can respectively for different categories of chemical composition between the correlation coefficient matrix and significant, sex coefficient, combining with both can analyze the correlation of different chemical components, the following is different categories of chemical composition correlation matrix, through the data visualization with thermodynamic graph intuitive can see, the deeper the color represents the higher the correlation.

According to the heat map, Under the category of weathered and high potassium, Calcium oxide, magnesium oxide, alumina and the two pairs have a strong correlation, Potassium oxide and iron oxide, Copper oxide and phosphorus pentoxide also have a strong correlation; in the weathered lead-barium category, Calcium oxide, magnesium oxide, alumina and iron oxide, Calcium oxide and magnesium oxide, Magnesium oxide and alumina have a general correlation, Copper oxide and barium oxide, Calcium oxide and phosphorus pentoxide, Aluminum oxide and strontium oxide, There is a strong correlation between barium oxide and sulfur dioxide; In the unweathered HK category, Potassium oxide and calcium oxide, Lead oxide and barium oxide, Aluminum oxide, magnesium oxide, and phosphorus pentoxide, Magnesium oxide and strontium oxide have a general correlation, There is a strong correlation between phosphorus pentoxide and strontium oxide and iron oxide; In the unweathered lead-barium type, Copper oxide and barium oxide, The strong association between calcium oxide and tin oxide, Lead oxide and tin oxide have a general correlation.

The previous part has yielded associations between the chemical components in four different cases, namely the correlation coefficient matrix. This question obtains the difference in association between pairwise species by solving the Euclidean distance of the correlation coefficient matrix.

The 2-norm, also known as the Euclidean norm, is usually used to find the spatial linear distance between vectors, to obtain the vector length of the difference matrix of different categories of correlation degrees, that is, finds the linear distance between the correlation degree, and also describes the difference between the correlation degree. After calculation, the difference between the pairs is as follows (matlab calculation code in the appendix).

It can be analyzed that the biggest difference is the correlation between the weathered high potassium and unweathered high potassium, and the smallest difference is the correlation between the weathered lead barium and unweathered lead barium.

7. Conclusion

The advantages of the model were analyzed by chi-square distribution for the four classified variables in problem 1, and then the relationship between the color significance test value and the P-value was found in contradiction, so the correlation was further found with the help of Spearman again. Make a descriptive analysis of the data before using the statistical rules, using the box chart.28 Nuclear naive Bayesian analysis method was used to learn the two chemicals, and the accuracy reached 100%, and the contour coefficient method was used to find the best K value, to ensure the accuracy and rationality of the model. The naive Bayes model classification by 100% by cross-validation. The sensitivity is better after changing the data and increasing the perturbation coefficient of the data. The disadvantages of the model determine the posterior probability through the prior and the data to determine the classification, and there may be some correlation between the variables. The improvement of the model further collects more sample data to ensure the training degree and accuracy of machine learning. The promotion of the model: 1. Linear regression or weighted average is used to predict the content of each chemical composition before weathering. 2. Gray correlation degree or path analysis method can be used to calculate the relationship between chemical formation, which can be applied to the fields of cultural relic protection, identification, chemical preparation, chemical weathering analysis and other fields.

References

[1] Yu Hongmei, Luo Yanhong, Sa Jian, et al. Within-group correlation coefficients and their software implementation [J]. China Health Statistics, 2011, 28 (5): 497-500.

[2] IvanIdris Idris. Python data analysis [M]. People's Posts and Telecommunications Press, 2016.

[3] Liang Beichen, Dai Jingmin. The Application of Partial Least-squares Method in System Fault Diagnosis [J]. Journal of Harbin Institute of Technology, 2020, 52 (3): 156-164. DOI:10.11918/201805149.

[4] Li Hang. Statistical Learning Methods [M]. Tsinghua University Press, 2012.

[5] He Xiaoqun. multivariate statistical analysis. Beijing: China Renmin University Press, 2012.

[6] Saroj, Kavita. Review: study on simple k mean and modified K mean clustering technique [J]. International Journal of Computer Science Engineering and Technology, 2016, 6 (7): 279-281.

[7] Xu Weichao. Review of the correlation coefficient studies [J]. Journal of Guangdong University of Technology, 2012, 29 (3): 12-17.

[8] Zhou Zhihua. machine learning [M]. Tsinghua University Press, 2016.

[9] Xia Bangshi, Wu Jinhua. Application of the Kappa concordance test in laboratory medicine research [J]. The Chinese Journal of Laboratory Medicine, 2006, 29(1): 83-84.

[10] NARAYAN S, KRISHNA C S, MISHRA V, et al. Ultron-AutoML:an opensource, distributed, scalable framework for efficient hyper-parameter optimization [C] // Proceedings of IEEE International Conference on Big Data (Big Data), 2020: 1584-1593.

[11] Xie Juanying, Ding Lijuan, Wang Mingzhao. Unsupervised feature selection algorithm based on spectral clustering [J]. Software Journal, 2020, 31 (4): 1009-1024.