

# Research on Glass Classification and Identification based on Machine Learning and Monte Carlo Algorithm

Jiasheng Zhou <sup>1,\*</sup>, Zhihong Fan <sup>1</sup>, Wenxin Zhang <sup>2</sup>

<sup>1</sup> Fuzhou University, College of Computer and Data Science/College of Software, Fujian, Fuzhou, 350108, China

<sup>2</sup> Fuzhou University, College of Physics and Information Engineering, Fujian, Fuzhou, 350108, China

\* Corresponding author email: 1246511834@qq.com

**Abstract.** In order to classify and identify glass, we used logistic regression and various machine learning algorithms to classify glass of known composition and class, and used various metrics to evaluate the classification results comprehensively. To further improve the reliability and accuracy of the classification. In order to further improve the reliability and accuracy of the classification, we used grid search and Bayesian optimization to adjust the parameters of the SVM, and the optimized SVM was able to classify the glass with known components and classes with 99.25% accuracy. Finally, we applied the SVMs with grid search and Bayesian optimization to predict the type of glass with only the chemical composition identified and the category unknown.

**Keywords:** Machine Learning; Monte Carlo Algorithm; Glass Classification.

## 1. Introduction

Glass is an important physical evidence of early trade exchanges between China and the West in ancient times, and ancient and foreign glass objects in China are similar in appearance but comprise different chemical compositions [1]. For ancient glass objects in China, archaeologists have used a variety of tests to determine information about their decoration, type, colour and chemical composition, and this relevant information is closely linked to the differentiation and type of the artefacts; which one or more of these pieces of information played a key role in the type and nature of the glass remains to be further investigated. In recent years, the identification and classification of glass has been a hot topic among scholars worldwide. El-Khatib M J [2] proposed a neural network-based approach for glass identification, Kaspi O [3] proposed a structured workflow for glass fragment analysis based on a combination of elemental analysis using PIXE and machine learning tools, Bamer F [4] proposed a data-driven approach using rigorous statistical analysis and polynomial regression in a data-driven approach to identify variations in silica glass, Devi [5] performed glass type classification by using a dataset of glass types retrieved from the UCI machine learning repository, and Du Wenxiu et al [6] proposed their own classification method for fire-resistant glass. All previous studies have certain shortcomings, such as incomplete consideration and underutilization of data for artefact identification.

## 2. Classification Model Building and Solving

### 2.1 Logistic Regression-based Analysis of Heritage Types

First, we used logistic regression to analyse the chemical composition of two different types of glass artefacts in order to predict the type of unknown artefact. The lead-barium type of glass and the high-potassium type of glass were recorded as dependent variables 0 and 1, respectively. In addition, as differentiation is a qualitative variable, for this reason we introduced the dummy variable CONDITION to replace these two variables, with 0 and 1 indicating unweathered and weathered, respectively. The regression of the dependent variable using the indicators as independent variables is expressed as follows.

$$y = \beta_0 + \beta_1 \text{CONDITION} + \beta_2 x_2 + \dots + \beta_{14} x_{14} + \mu_i \quad (1)$$

The coefficients of the regression equation were obtained as shown in Table 1.

**Table 1.** Table of regression equation coefficients

| $\beta_0$             | $\beta_1$<br>(Weathering) | $\beta_2$<br>(SiO <sub>2</sub> )                 | $\beta_3$<br>(Na <sub>2</sub> O)               | $\beta_4$<br>(K <sub>2</sub> O)    |
|-----------------------|---------------------------|--------------------------------------------------|------------------------------------------------|------------------------------------|
| 29.117                | 0.333                     | -28.509                                          | -31.267                                        | -23.901                            |
| $\beta_5$<br>(CaO)    | $\beta_6$<br>(MgO)        | $\beta_7$<br>(Al <sub>2</sub> O <sub>3</sub> )   | $\beta_8$<br>(Fe <sub>2</sub> O <sub>3</sub> ) | $\beta_9$<br>(CuO)                 |
| -28.054               | -28.719                   | -31.027                                          | -26.361                                        | -26.877                            |
| $\beta_{10}$<br>(PbO) | $\beta_{11}$<br>(BaO)     | $\beta_{12}$<br>(P <sub>2</sub> O <sub>5</sub> ) | $\beta_{13}$<br>(SrO)                          | $\beta_{14}$<br>(SO <sub>2</sub> ) |
| -30.143               | -30.567                   | -29.997                                          | 0.000                                          | -27.766                            |

The predicted results for the unknown glass artefacts are shown in Table 2. For ease of presentation, we will denote high potassium glass as HP and lead-barium glass as LB.

**Table 2.** Logistic regression classification results

| Heritage number   | A1 | A2 | A3 | A4 | A5 | A6 | A7 | A8 |
|-------------------|----|----|----|----|----|----|----|----|
| Forecast Category | HP | LB | LB | LB | LB | LB | HP | LB |

In order to examine how good, the prediction results were, we randomly excluded a portion of the data by 30% and used it as a sample. For the remaining 70% of the sample, we again applied logistic regression to classify the 30% of the excluded sample and compared the classification results with the type of the original data.

The results obtained are shown in Table 3, where it can be seen that the overall percentage of correct classifications is only 82.6%, which shows that the use of logistic regression classification is not effective. Therefore, we introduced machine learning to classify the artefact types.

**Table 3.** Results of logistic regression classification

|           | Predicted HP | Predicted LB | Percentage correct |
|-----------|--------------|--------------|--------------------|
| Actual HP | 5            | 1            | 83.3               |
| Actual LB | 3            | 14           | 78.6               |

## 2.2 Machine Learning Related Evaluation Metrics

In order to evaluate how good, the classification is, a few machines learning related evaluation metrics are first introduced.

### (i) F1 score

First, let's take a binary classification problem as an example. In a binary classification problem, we have to determine the positive class first for the convenience of classification, and in machine learning we usually define the events that we are more concerned with as positive class events. The number of samples with high potassium is small, so we are more interested in whether the artefacts of type-high potassium are correctly classified, so we define high potassium as its positive class. The relevant indicators in Table 4 were then calculated separately.

**Table 4.** Indicators related to secondary classification

| Indicators          | Meaning                                                         |
|---------------------|-----------------------------------------------------------------|
| True Positive (TP)  | Number of predictions of positive classes into positive classes |
| False Negative (FN) | Number of positive classes predicted as negative classes        |
| False Positive (FP) | Number of negative classes predicted as positive classes        |
| True Negative (TN)  | Number of negative classes predicted as negative classes        |

In order to evaluate the effectiveness of machine learning classification models, the following four metrics are defined.

(1) Accuracy (ACC): The ratio of the number of correctly classified samples to the total number of samples.

$$ACC = \frac{TP + TN}{TP + FN + FP + TN} \quad (2)$$

(2) Recall (R): The proportion of samples that are actually positive classes that are predicted correctly.

$$R = \frac{TP}{TP + FN} \quad (3)$$

(3) Precision (P): The proportion of samples where the prediction was a positive class that were predicted correctly.

$$P = \frac{TP}{TP + FP} \quad (4)$$

In machine learning, the accuracy rate is not an accurate reflection of the effectiveness of a machine learning model, so we introduce a fourth indicator, the score.

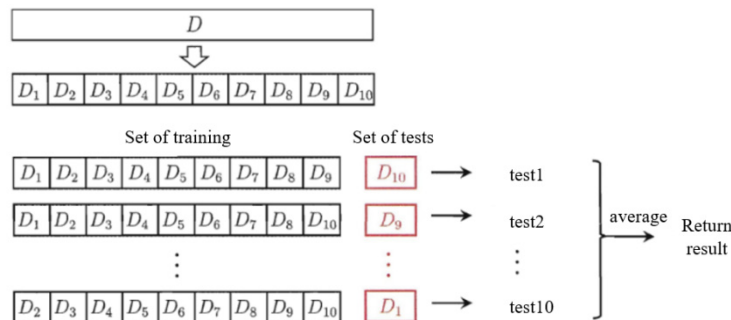
(4)  $F_1$  score: Reconciled average of completeness and accuracy rates.

$$F_1 = \frac{2TP}{2TP + FP + FN} \quad (5)$$

(ii) ROC curve and AUC

When performing classification, machine learning calculates the predicted probability of a sample. Depending on the set threshold, a positive class is classified when the predicted probability is greater than the threshold, and a negative class is classified when the opposite is true. The threshold setting directly affects the generalisation ability of the classifier, and the ROC curve is a curve comparing the generalisation ability of the same classifier based on different thresholds, with the true class rate (TPR) as the vertical coordinate and the false positive class rate (FPR) as the horizontal coordinate.

The AUC is defined as the area enclosed by the ROC curve and the axis below it. The range of the AUC is between [0,1] and the larger the AUC the better the model will classify. In general, if the AUC is less than or equal to 0.5, the model cannot be used, and models with an AUC greater than 0.85 perform better.



**Figure 1.** Tenfold cross validation diagram [7]

### (iii) Cross-validation

Cross-validation is used to prevent over-fitting problems caused by overly complex models. The data samples are first cut into K smaller subsets, each subset is made into a test set once and the rest is used as the validation set. The cross-validation is repeated K times. In cross-validation, ten-fold cross-validation is more commonly used, and Figure 1 shows the process of cross-validation graphically.

## 2.3 Theories Related to Machine Learning

### (i) Support Vector Machine (SVM)

Support Vector Machine (SVM) are a data mining method based on statistical learning theory that can handle regression problems and pattern recognition problems very successfully and can be extended to domain disciplines such as prediction and comprehensive evaluation. The mechanism of an SVM is to find an optimal classification hyperplane that satisfies the classification requirements such that the hyperplane maximises the blank area on either side of the hyperplane while ensuring classification accuracy [7]. Support vector machines can be broadly classified as linear or non-linear.

In a linear support vector machine, the hyperplane for the sample space can be described using the following linear equation.

$$w^T x + b = 0 \quad (6)$$

The hyperplane of the maximum margin needs to be solved for, with the constraint that the distance from the sample point to the decision boundary is greater than or equal to 1.

$$\begin{aligned} \max_{w,b} \frac{2}{\|w\|} &\iff \min_{w,b} \frac{1}{2} \|w\|^2 \\ \text{s.t. } y_i(w^T X_i + b) &\geq 1 \end{aligned} \quad (7)$$

A new optimisation problem is constructed by introducing a stochastic loss function, taking into account the classification error. The segmented values of the hinge loss function are handled using slack variables.

$$\begin{aligned} \min_{w,b} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^N \xi_i \\ \text{s.t. } y_i(w^T X_i + b) &\geq 1 - \xi_i, \xi_i \geq 0 \end{aligned} \quad (8)$$

Solving the above equation usually makes use of the duality of the optimisation problem to obtain the dual of the above equation and solve it.

$$\begin{aligned} \max_{\alpha} \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N [\alpha_i y_i (X_i)^T (X_j) y_j \alpha_j] \\ \text{s.t. } \sum_{i=1}^N \alpha_i y_i = 0, \quad 0 \leq \alpha_i \leq C \end{aligned} \quad (9)$$

For non-linear support vector machines, a kernel function  $K(x, z)$  can be used to map a low-dimensional data set to a higher-dimensional data set, where the data becomes linearly separable. In other words, given a kernel function  $K(x, z)$ , a non-linear classification problem can be solved in the same way as a linear classification problem. The commonly used kernel functions are polynomial kernel function and Gaussian kernel function. Based on the characteristics of the model, we use Gaussian kernel function to construct the non-linear support vector machine.

$$K(x, z) = \exp\left(-\frac{\|x - z\|^2}{2\sigma^2}\right) \quad (10)$$

The corresponding support vector machine is a Gaussian radial basis function classifier. In this case, the classification decision function is

$$f(x) = \text{sign}\left(\sum_{i=1}^{N_s} a_i^* y_i \exp\left(-\frac{\|x - x_i\|^2}{2\sigma^2}\right) + b^*\right) \quad (11)$$

(ii) Naive Bayes classifier (NBC)

Naive Bayes classifier [8] is a method based on Bayes' theorem and the assumption that the feature conditions are independent of each other. The joint probability distribution from input to output is first learnt from a given training set with the assumption that the feature terms are independent of each other, and then based on the learnt model, the input is found to maximize the posterior probability of the output.

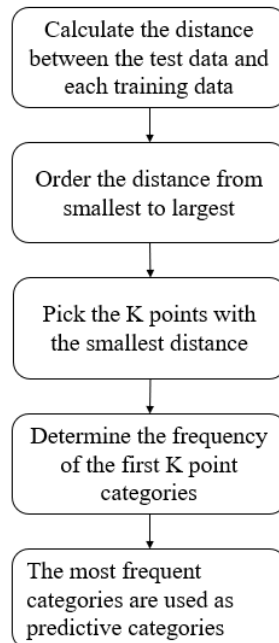
For a given set of elements  $y$  in a category and a subordinate feature vector  $x_1, x_2, \dots, x_n$ , the features are assumed to be independent of each other and are expressed in terms of Bayes' theorem as follows.

$$P(y|x_1, x_2, \dots, x_n) = \frac{P(y) \prod_{i=1}^n P(x_i|y)}{P(x_1, x_2, \dots, x_n)} \quad (12)$$

Since  $P(x_1, x_2, \dots, x_n)$  is a known constant [10], the classification rule is used.

$$\begin{aligned} P(y|x_1, x_2, \dots, x_n) &\propto P(y) \prod_{i=1}^n P(x_i|y) \\ \Rightarrow \hat{y} &= \underset{y}{\operatorname{argmax}} P(y) \prod_{i=1}^n P(x_i|y) \end{aligned} \quad (13)$$

The maximum a posteriori estimates  $P(y)$  and  $P(x_i|y)$  can be used, and different assumptions on  $P(x_i|y)$  can lead to different plain Bayesian classifiers.



**Figure 2.** Flow chart of the Naive Bayes algorithm

(iii) Random forests (RF)

Random Forest is an integrated algorithm. As individual classifiers often have problems such as poor classification accuracy and tendency to overfit, integrated learning is the aggregation of individual classifiers to integrate all classifiers. A random forest is a combined classification model consisting of many decision tree classification models. When a sample to be classified is input, the output of the random forest is determined by a simple vote on the classification result of each decision tree in the following steps.

1. randomly select  $N$  samples from the original dataset of  $N$  samples and build a decision tree model based on these  $N$  samples.
2. repeat the above process  $K$  times to construct  $K$  different decision trees.

3. input random variables and generate predictions for each decision tree, yielding K classification results.

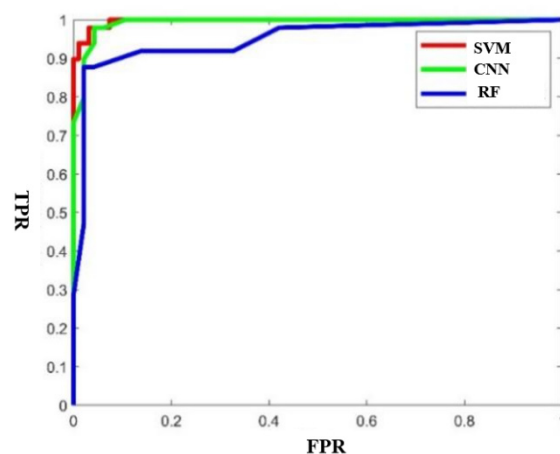
4. Vote on each record to determine its final classification based on the results of the K classifications.

## 2.4 Model Comparison

We used support vector machines, plain Bayes, and random forests to classify the food in Form 2 through supervised learning and cross-validation, respectively, plotting the ROC curves and confusion matrix, and calculating each parameter such as AUC, with the results shown in Table 5 and Figure 4.

**Table 5.** Comparison of evaluation indicators of glass heritage classification models based on machine learning methods

| Model | ACC    | $F_1$  | AUC    |
|-------|--------|--------|--------|
| SVM   | 0.9583 | 0.9583 | 0.9968 |
| NBC   | 0.9444 | 0.9425 | 0.9625 |
| RF    | 0.9514 | 0.9513 | 0.9484 |



**Figure 3.** Comparison of ROC curves of glass artefact classification models based on machine learning methods

Analysis of the results in Table 3 shows that the classification accuracy of SVM is as high as 95.83%, the highest than the other two, and the AUC is also better than both plain Bayes and random forest, and the AUC is greater than 0.85, the model performs better; in addition, according to the ROC curve in Figure 3, we can see that the ROC curve of SVM is always on the leftmost side. From this, we conclude the result that SVM has a better classification effect than CNN and random forest. In order to obtain a better model, we performed grid search optimization and Bayesian conditioning on the SVM.

## 2.5 SVM for Monte Carlo Simulation and Grid Search Optimization

The three types of SVM hyperparameters we have chosen are the kernel function, the box constraint level and the kernel size. They are explained as follows.

The kernel function breaks through the limitations of linear SVMs and allows SVMs to handle non-linear data distributions.

The box constraint level controls the number of support vector machines. Increasing the kernel constraint level reduces the number of support vectors, but also increases the training time. The box constraint level parameter is a soft boundary penalty in the original equations and is a hard "kernel" constraint on the original equations.

The size of the kernel scale allows for the partitioning of all elements of the prediction matrix. The Gram matrix is then calculated by applying the appropriate kernel parametrization.

The kernel functions can be of linear, polynomial and Gaussian types with positive box constraint levels and kernel sizes greater than 0.

In order to obtain the optimal interval for the hyperparameters while avoiding getting trapped in a local optimum solution, Monte Carlo simulation is used prior to the traversal of the hyperparameters using the grid search method. Monte Carlo simulations are statistical simulations or sampling implemented using an electronic computer to obtain an approximate solution to the problem by relating the problem being solved to a certain probabilistic model. The results of the Monte Carlo simulation show that the optimal interval for the box constraint level and kernel size is from 1 to 10.

To further optimise the SVM, a grid search method was applied. The grid search is a process of traversing every possible scenario within a specified range of parameters to see which set of parameters gives the best results for the model. The results of the grid search traversal of the three hyperparameters of the SVM are shown in Appendix 6. The optimal parameters were found to be a Gaussian kernel function, a box constraint level of 4 and a kernel size of 4. The accuracy of the optimised SVM was 99.25%.

## 2.6 SVM with Bayesian Optimization

Bayesian tuning is a heuristic tuning strategy. The basic idea is that given an optimized objective function, the posterior distribution of the objective function is updated by continuously adding sample points to better adjust the current parameters. Bayesian conditioning also has the advantages of fewer iterations, faster and more robust to various problems.

The Bayesian optimization framework consists of two main core components [8]: a probabilistic agent model and a collection function, which is implemented through an iterative process consisting of three main steps.

Step 1: select the next most "promising" assessment point based on the maximisation function  $x_t$ .

Step 2: evaluate the objective function value  $y_t = f(x_t) + \varepsilon_t$  based on the selected evaluation point  $x_t$ .

Step 3: the newly obtained input-observation pairs  $\{x_t, y_t\}$  are added to the historical observation set  $D_{1:t-1}$  and the probabilistic proxy model is updated in preparation for the next iteration.

Algorithm 1 is a Bayesian optimization framework pseudo-code, as shown in table 6.

**Table 6.** Bayesian optimization framework pseudo-code

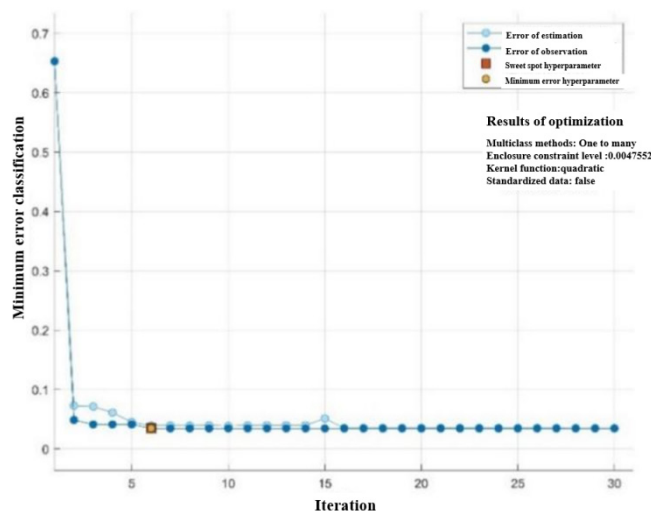
### **Algorithm 1: Bayesian optimization framework [8]**

- 1: **for**  $t=1,2,\dots$  **do**
- 2:     Maximize the collection function to get the next evaluation point:  

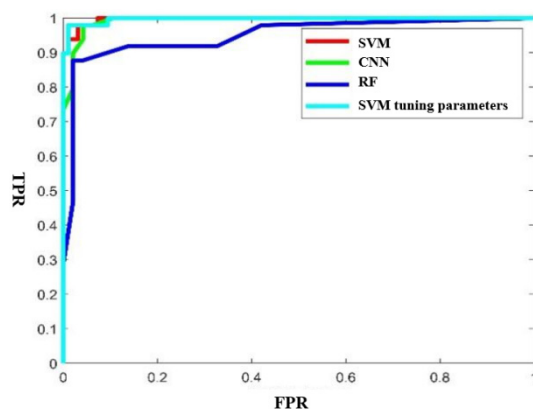
$$x_t = \arg \max_{x \in \mathcal{X}} \alpha(x|D_{1:t-1});$$
- 3:     Evaluate the objective function value  $y_t = f(x_t) + \varepsilon_t$ ;
- 4:     Integrate the data:  $D_t = D_{t-1} \cup \{x_t, y_t\}$  and update the probabilistic proxy model;

The Bayesian tuning process is shown in Figure 4. According to the minimum classification error map of the SVM, the optimal solution for each parameter can be obtained after 30 iterations.

The SVM was used to reclassify typical artefact samples with an accuracy of 98.86%. The ROC curves are shown in Figure 5, which shows that the SVM with the tuned parameters has the best results compared with the original three models. In addition, considering the large randomness of the samples selected for cross-validation, we also conducted several validations on the samples, and the results also proved that the performance of the SVM after covariance has improved significantly.



**Figure 4.** SVM minimum classification error plot



**Figure 5.** Comparison of ROC curves of glass artefact classification models based on machine learning methods

In the process of building the machine learning model and subsequent optimisation, we used 10-fold cross-validation and performed multiple tests in order to avoid over-fitting due to the small amount of data.

Finally, we applied the tuned SVM for classification and the classification results are shown in Table 7.

**Table 7.** SVM classification results after tuning

| Heritage number   | A1 | A2 | A3 | A4 | A5 | A6 | A7 | A8 |
|-------------------|----|----|----|----|----|----|----|----|
| Forecast Category | HP | LB | LB | LB | LB | HP | HP | LB |

## 2.7 Sensitivity Analysis of the Results

In sensitivity testing, the control variable method is used, where the value of an important parameter in the classification process is changed while controlling for other parameters to remain constant, and then the change in the model results is observed [10]. If a change in a parameter has a significant effect on the result, this result is unstable and sensitive. We can determine the stability of the results by recording and evaluating the corresponding changes in the results.

For the machine learning classification model, to test the sensitivity of the results, we chose the metric SiO<sub>2</sub>, which has the greatest impact. Ten sets of data were first set up, each with SiO<sub>2</sub> in an equal series, and the SiO<sub>2</sub> content of the previous set of data was increased or decreased by 0.5% in turn, i.e. the SiO<sub>2</sub> content varied in the range of -2.5% to 2.5%, and the changed data were input into

the grid search method optimised SVM and the Bayesian optimised SVM respectively, and the classification was performed again.

We found that for small changes in SiO<sub>2</sub> content, the classification results of both tuned SVMs for glass did not change significantly, with small deviations within an acceptable range. We can conclude that our results are insensitive to changes in SiO<sub>2</sub> content, have good stability and are less influenced by other factors, therefore, our classification results are stable.

### 3. Conclusion

We used three machine learning algorithms, logistic regression and support vector machines (SVM), simple Bayesian (NBC) and random forest (RF), to classify the glass in Form 2. In order to further improve the reliability and accuracy of the classification, we used grid search and Bayesian optimization to optimize the SVM. The grid search and the Bayesian optimization of the SVM were used to classify Form 2 again with an accuracy of 99.25% and 98.86%, respectively. Finally, we applied the SVM with grid search and Bayesian optimization to the data in Form 3 and predicted the glass type with no significant difference, demonstrating the robustness of the prediction results.

### References

- [1] Zi Zi Yan. An analysis of the influence of western glass on early Chinese glass[J]. Newspaper Aloe: Next, 2018(4):1.
- [2] El-Khatib M J, Abu-Nasser B S, Abu-Naser S S. Glass Classification Using Artificial Neural Network[J]. 2019.
- [3] Kaspi O, Girshevitz O, Senderowitz H. PIXE based, Machine-Learning (PIXEL) supported workflow for glass fragments classification. [J]. Talanta, 2021, 234:122608.
- [4] Bamer F, Alshabab S S, Paul A, et al. Data-driven classification of elementary rearrangement events in silica glass[J]. Scripta Materialia, 2021, 205:114179-.
- [5] Devi, M.S., Aruna, R., Kumar, S.V., Chowdary, G.V., Raju, B.V.S.S.K., Prasad, M.S. (2022). Statistical Oversampling Classification Based Glass Type Identification Through Oxide Content. In: Mekhilef, S., Shaw, R.N., Siano, P. (eds) Innovations in Electrical and Electronic Engineering. ICEEE 2022.
- [6] Du Wenxiu, Yang Juan, Gao Shuai, et al. A brief discussion on the characteristics and classification of fireproof glass[J]. Anhui Architecture, 2015, 22(4):3.
- [7] Zhou Zhihua, Machine learning, Beijing: Tsinghua University Press, 46, 2016.1.
- [8] Zhang Hang. Chinese text classification based on plain Bayesian and Python implementation[D]. Shandong Normal University, 2018.
- [9] Li, Yunfan, Hu, How-Cheng, Kang, Jiale. Application of plain Bayesian algorithm[J]. Computer Programming Skills and Maintenance, 2018(10):47+41. DOI: 10.16184/j.cnki.comprg.2018.10.001.
- [10] Chen, Juan, Wang, Xiangyu, Luo, Lingling and Cui, Jingjing. Missing value filling effect: A comparison of machine learning and statistical learning[J]. Statistics and Decision Making, 2020(17):5.