Advances in Computer, Signals and Systems (2022)
Clausius Scientific Press, Canada

DOI: 10.23977/acss.2022.060601
ISSN 2371-8838 Vol. 6 Num. 6

Classification and identification of glass artifacts based on high-dimensional clustering and XGBoost algorithm

Yilin Zeng

School of Economics &Management, Shanghai Maritime University, Shanghai, 201306, China

Keywords: Integrated learning; high-dimensional clustering; XGBoost; glass classification

Abstract: After the weathering of glass artifacts, a large number of environmental elements are exchanged with elements inside the glass artifacts. In this paper, based on the chemical composition of the artifact samples and other detection means, they are classified into two types: lead-barium glass and high-potassium glass. A sub-classification model based on high-dimensional clustering is introduced, and an identification model is established by optimizing the K-means algorithm. Finally, a feature classification model based on XGBoost algorithm is used to analyze the chemical composition of the unknown class of glass artifacts.

1. Introduction

The Silk Road was an important way of cultural exchange between China and the West in ancient times, of which glass was a witness to the early trade between China and the West[1-2]. The main source of glass is quartz sand, whose main component is silicon dioxide (SiO2)[3]. In the process of making glass, in order to reduce the melting temperature of quartz sand, it is often necessary to add fluxes. Commonly used fluxes are grass ash, natural bubble soda, lead ore and saltpeter, etc., while limestone needs to be added as a stabilizer[4-5]. In ancient times, glass was highly susceptible to weathering due to the role of the buried environment. After weathering, a large number of environmental elements are exchanged with the internal elements of glass artifacts, thus changing their composition ratios and affecting the judgment of their categories[6]. Therefore, this paper attempts to establish a classification and identification model of glass artifacts through high-dimensional clustering and XGBoost algorithm.

2. Building and solving sub-classification models based on high-dimensional clustering

In order to select the appropriate chemical components for each category to subclassify them, we need to use the K-means clustering algorithm to build a high-dimensional clustering model.

The K-means algorithm, also known as the mean-equivalence method, has the central idea of dividing several data objects in Euclidean space and implementing object selection by an initial center strategy to make them cluster centers. Then, by calculating the distance between the other objects and each center of mass, the nearest categorization is used, and again the mean value of each cluster data is calculated accordingly to obtain a brand new cluster center, and this process is repeatedly iterated until all clusters converge.

Although the K-means algorithm is simple, fast and relatively efficient for handling large data sets. However, it requires the user to give the number K of clusters to be generated in advance, and is sensitive to the initial values and to isolated point data.

Therefore, on this basis, we choose to use a partitioning-based method for clustering categorical datasets for subclassification, using which we can improve the speed of clustering categorical datasets. Because the classification dataset is divided into several subsets, each subset includes less data, these subsets can be clustered in parallel, and the sensitivity of the model algorithm to special data can be reduced, and the final clustering results can be obtained by fusing the individual clustering results.

2.1 Building a sub-classification model based on high-dimensional clustering

Selection of clusters first, followed by subclassification of data.Assumptions$X = [x_1, x_2, \ldots . x_n]$ is a set of data tuples, Among them $X_i = [x_{i1}, x_{i2}, \ldots . x_{im}]$ represents the data objects with m attributes. Assume that K is a positive integer while dividing n objects in X among K categories using the following cost function minimum as the clustering criterion:

$$E_c(W,Q) = \sum_{l=1}^{k}\sum_{i=1}^{n} \omega_{li}^{a} d_c(Q_l, X_i) \tag{1}$$

By observation of the data. Here, dc is the difference measure and a > 1 is the weighted index, and for categorical data, the difference measure is defined as

$$d_c(Q_l, X_i) = \sum_{j=1}^{m} \delta\left(q_{lj}^{c}, x_{ij}^{c}\right) \tag{2}$$

The update method of the clustering center is shown below: Set $A_j = \{d_j^{(1)}, d_j^{(2)}, \ldots . , d_j^{(nj)}]$, where nj is the number of values that can be taken for the jth classification attribute and the cluster center $Q_l = [q_{l1}, q_{l2}, \ldots . q_{lm}]$ There is the following theorem, formulas (3) Obtain minimum, when and only when $q_{ij} = a_j^{(r)} \in A_j, j = 1,2,...,m$ satisfies the following equation:

$$\sum_{i, x_{i,j} = a_j^{(r)}} \omega_{li}^{a} \geq \sum_{i, x_{i,j} = a_j^{(t)}} \omega_{li}^{a} \quad ,1 \leq t \leq n_j \tag{3}$$

2.2 Solving subclassification models based on high-dimensional clustering

By analyzing the data of this batch of glass artifacts through spss software, the clustering centers were selected and these data were solved by subclassifying.

2.3 Rationalization and sensitivity analysis of classification results

2.3.1 Rational Analysis

By establishing a subclassification model based on high-dimensional clustering, the data of glass artifacts given in the question are organized and analyzed, and the clustering centers and clusters are identified by algorithmic rules, and then the data are further processed to divide the final clustering range and complete the subclassification process for each category. Based on the K-means algorithm, the subclassification model based on high-dimensional clustering inherits its advantages, and the algorithm is simple and fast, which is relatively efficient in dealing with large data sets. At the same time, it improves the large data processing capability of the model in performing the division of

subspace sets, accelerates the model response speed, and provides strong support for the analysis of multiple sample data.

2.3.2 Sensitivity Analysis

Adding a slight perturbation (adding 0.1 for each chemical component value), its classification results remain unchanged with a small change in the clustering center. It means that its sensitivity is low and the minor perturbation will not disturb the clustering of the model and thus will not affect the subclassification results.

3. Establishment of feature classification model based on XGBoost algorithm and component identification

3.1 Model Introduction

For data classification problems, we usually use the GBDT (Gradient Boosting Decision Tree) algorithm to build a classification model.

The GBDT model is an additive model that serially trains a set of CART regression trees and eventually sums the predictions of all the regression trees, resulting in a strong learner where each new tree fits the negative gradient direction of the current loss function. For the classification task, each base learner is actually fitting the residuals (probability residuals) of the current GBDT output to derive the classification result.

However, the GBDT model has the problem of being sensitive to outliers, and it is difficult to perform parallel computation because of the dependencies between classifiers. Therefore, we choose the XGBoost (Extreme Gradient Boos-ting) algorithm for the model construction. Compared with the GBDT algorithm, the XGBoost algorithm supports both categorical and linear regression trees, performs a second-order Taylor expansion of the cost function, uses both first-order and second-order derivatives, and adds a regular term to control the complexity of the model, which can prevent overfitting of the model.

The procedure is as follows.

Step1.Build XGBoost classification model by training set data.

Step2. Calculate the feature importance by the established XGBoost.

Step3. Apply the established XGBoost classification model to the training and testing data to get the classification evaluation results of the model.

Step4. Since XGBoost has randomness, the result of each operation is not the same, if this training model is saved, the subsequent data can be directly uploaded to this training model for calculation of classification.

Overall, XGBoost algorithm has the advantages of high accuracy, not easy to overfit, scalability, and high performance.

Overall, the XGBoost algorithm has the advantages of high accuracy, less over-fitting, scalability, and high performance.

3.2 Parameter determination

Table 1: Model parameters

Parameter Name	Parameter Value
Training time	1.056s
Data slicing	0.7
Data shuffling	Negative
Cross-validation	10
Base learners	gbtree
Number of base learners	100
Learning rate	0.1
L1 regular term	0
L2 regular term	1
Sample feature sampling rate	1
Tree feature sampling rate	1
Node feature sampling rate	1
Minimum weight of samples in leaf nodes	0
Maximum depth of the tree	10

3.3 Model evaluation

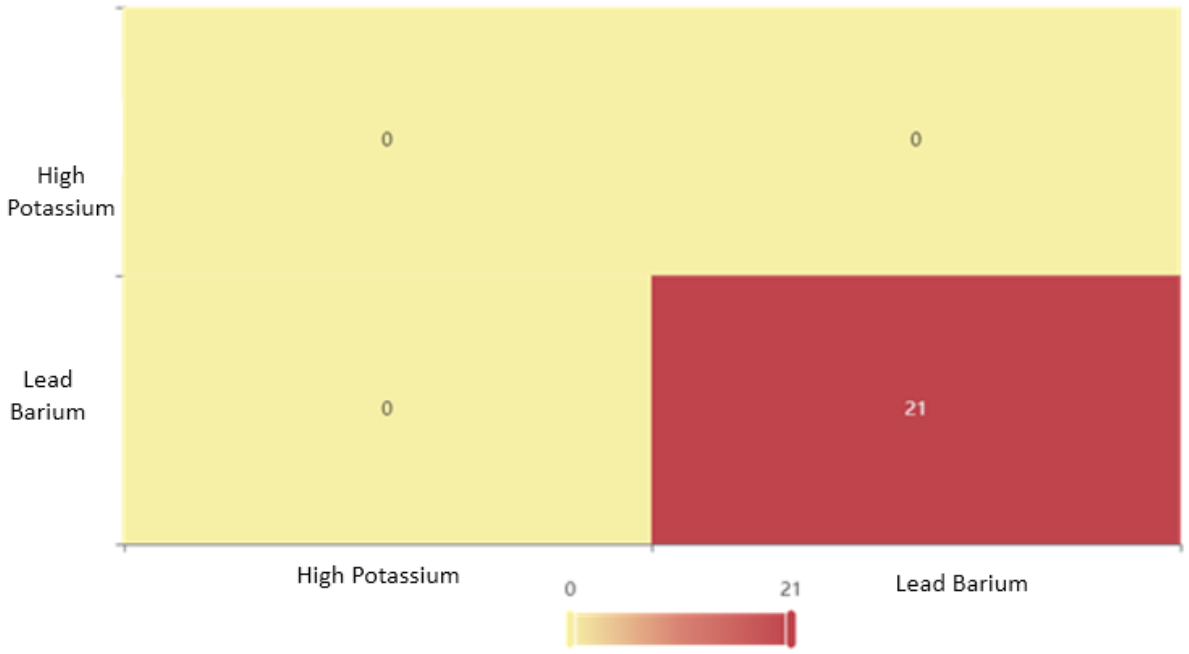

Figure 1: Chaos matrix

Figure 1 shows the confusion matrix in the form of a heat map.

Table 2: Predictive assessment indicators

	Accuracy	Recall	Accuracy	F1
Training set	1	1	1	1
Cross-validation set	1	1	1	1
Test set	1	1	1	1

The prediction evaluation metrics for the cross-validation set, training set and test set are shown in Table 1 and Table 2, and the prediction effectiveness of XGBoost is measured by quantitative metrics. Among them, the evaluation metrics of the cross-validation set can continuously adjust the hyperparameters to obtain a reliable and stable model.

● Accuracy: The proportion of correct predicted samples to the total samples, and the larger the accuracy, the better.

● Recall rate: The proportion of results that are actually positive samples that are predicted to be positive samples, the larger the recall rate, the better.

● Precision rate: The proportion of the predicted positive samples to the actual positive samples, and the larger the precision rate, the better.

●F1: The sum of precision and recall. Precision and recall affect each other, and although a high level of both is the desired ideal, in practice it is often the case that a high precision rate results in a low recall rate, or a low recall rate results in a high precision rate. If a balance of both is needed, then the F1 metric can be used.

3.3.1 Model Building

XGBoost sums the k trees according to the additive model of the Boosting algorithm, and the data set containing m features with a capacity of n is denoted as, $D = \{(x_i, y_i)(|D| = n, x_i \in R^m, y_i \in R)\}$, The model function can be expressed in equation (4).

$$\hat{y} = \phi(x_i) = \sum_{k=1}^{k} f_k(x_i), f_k \in F \tag{4}$$

where f(x) is the CART regression tree, $F = \{f(x) = w_{q(x)}\}(q: R^m \to T, w \in R^T$ is the set space of the regression tree. the leaf nodes are partitioned again in the XGBoost algorithm according to the new features, and the shape of the process tree in Figure 2 is obtained.

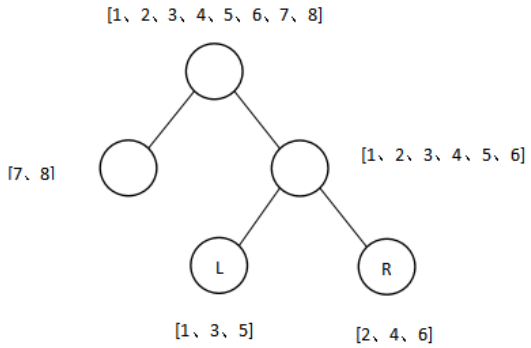

Figure 2: Process tree

The objective function of XGBoost is the following equation, which can avoid overfitting the model by limiting the complexity of each regression tree.

$$L\ （\phi）=\sum_{i} l(\hat{y}_i, y_i)+\sum_{k}\Omega(f_k) \tag{5}$$

The data in the question is processed through the construction of the feature classification model of the XGBoost algorithm, and the results are solved and analyzed.

3.3.2 Solution of feature classification model based on XGBoost algorithm

The data analysis by using SPSS software leads to Table 3.

Table 3: Predicted probabilities

Actual category	Lead and barium predicted probability	High potassium prediction probability	Predicted results	Accuracy
High Potassium	0.93%	99.07%	High potassium	√
Lead Barium	99.37%	0.63%	lead barium	√
Lead barium	99.37%	0.63%	lead barium	√
Lead barium	99.37%	0.63%	lead barium	√
Lead barium	99.37%	0.63%	Lead barium	√
High Potassium	0.93%	99.07%	High Potassium	√
high potassium	0.93%	99.07%	high potassium	√
lead barium	99.37%	0.63%	lead barium	√

3.3.3 Sensitivity Analysis

When the three chemical compositions of silica, sodium oxide, and lead oxide were treated to simulate the case of receiving interference, the XGBoost model predicted the same glass type as the original data for these three cases, and the probability value of predicting high potassium glass or lead-barium glass was as high as 99%, indicating that the XGBoost model is resistant to interference.

4. Conclusions

In ancient times, glass was extremely susceptible to weathering due to the effects of the burial environment. After weathering, environmental elements exchange with the internal elements of glass artifacts in large quantities, thus changing their composition ratios and affecting the judgment of their categories. To this end, this paper divides the glass samples into two categories: high-potassium glass and lead-barium glass, introduces a sub-classification model based on high-dimensional clustering, and divides these glass samples into six categories by optimizing the K-means algorithm, and finds the corresponding clustering centers so as to establish a recognition model. Immediately afterwards, a feature classification model based on the XGBoost algorithm was used to analyze the chemical composition of the unknown class of glass artifacts. A sensitivity analysis was also performed on the analysis results, and the probability value of predicting high potassium glass or lead-barium glass was as high as 99%, indicating that the XGBoost model is highly resistant to interference.

References

[1] Liu Z. X., Deng P. F., Pan S. H., Luo W. D., Zhang Shilin, and Li Tao Ming. Classification and research status of fireproof glass [J]. Guangzhou Chemical, 2021, 49(15):16-18.
[2] Lu Yue. Deep learning-based classification and detection of cell phone glass defects [D]. Zhengzhou University., 2019.
[3] Shi Youzhou, A kind of glass sorting storage rack. Anhui Province, Anhui Youtong Glass Co., 2019-04-19.
[4] Cheng Wei. Research on machine vision-based defect detection and classification system for glass fiber electronic fabric [D]. Xi'an University of Engineering, 2018.
[5] Xue Yuan. Research on Classification and Recognition of Glass Defects Based on Machine Vision [D]. Hefei University of Technology, 2018.
[6] ZHOU Xin, DENG Wenyi, LIU Lishuang. Research on the classification of glass defects for rapid detection[J]. Microcomputer Information, 2008(27):304-305+20.