Analysis and identification of the composition of ancient glass objects based on an improved DBSCAN model

Ruijin Li[1,a], Kaihao Zhang[1,b] and Baoyi An[2,c*]

rjli2020@lzu.edu.cn[a], zhangkh20@lzu.edu.cn[b], anby20@lzu.edu.cn[c*]

[1]School of Management, Lanzhou University, Lanzhou730000, China
[2]School of Physical Science and Technology, Lanzhou University, Lanzhou 730000, China

Abstract: A scientific approach to classifying cultural objects is conducive to more efficient management of historical artefacts. According to existing research, the chemical composition content of artefacts can determine the type of composition, degree of weathering, and production age. This paper provides a method for classifying and identifying ancient glass artefacts based on their chemical composition content, which has not yet been investigated. An improved DBSCAN model is used to classify marked glass artefact samples. A support vector machine and a combined binary logistic regression model are used to classify and identify unknown glass artefact samples, achieving an accuracy of 90%. The method is used to classify two common types of ancient glass artefacts and their degree of weathering in China, thus providing a new approach to dating artefact production.

Keywords: Heritage identification; Support vector machine; DBSCAN model; Binary logistic regression model

1 INTRODUCTION

The vast array of magnificent cultural heritage is a self-evident jewel of human history and a vital treasure shared by all humanity. With the continuous development of modern heritage conservation and management technology, the use of digital technology for heritage conservation has become an important heritage conservation and management method for many heritage conservation and research units. Among them, glass artifacts, one of the most beautiful existences, have been in China for about 2500 years. This paper focuses on the composition analysis and identification of glass artifacts for conservation, where glass is a material made from the inorganic minerals quartz sand, borax, and barite, with SiO2 being the main chemical component of glass [4].

The two main types of glass relics commonly found in China in ancient times are lead-barium glass and potassium glass. The lead-barium glass, the result of the wisdom of our ancient people, was made by melting quartz sand with the addition of lead ore as a solubilizing agent, resulting in a product with lead oxide and barium oxide as the main components. Another common type of glass is potassium glass, fueled by a substance with a high potassium content, such as grass ash. At the same time, the high temperature of the refining process caused the stabilizer limestone to be converted into calcium oxide. Ancient glass objects were buried under different

EMIS 2023, February 24-26, Chengdu, People's Republic of China

DOI 10.4108/eai.24-2-2023.2330619

conditions of temperature, humidity, acidity, and alkalinity, which led to different chemical reactions on their surfaces, resulting in changes in the chemical content of the glass, known as 'weathering.' The damage caused by the 'weathering' will impact the archaeologist's research observations.

The DBSCAN model, as a density-related clustering algorithm, was first proposed by German scholar M. Ester et al. in KDD96. The model has subsequently been continuously optimized. In 2012, the American scholar Patwary proposed the improved algorithm PDS-DBSCAN, which combines the concept of graph algorithms to solve the problems of workload imbalance and reduced operational efficiency caused by existing master-slave strategies, achieving better workload distribution and shared memory acceleration. In 2013, the German scholar Kellner proposed the grid-based DBSCAN, applied to the analysis of radar data, focused on adjusting to adaptive clustering neighborhood radius that varies according to the target distance to avoid missed and false detections due to different distant and near targets [5]. 2011 A new hybrid method, PACA-DBSCAN, combines partition-based PDBSCAN with ant clustering [3], which can handle multidimensional data.

This paper analyzes the raw data of glass artifacts and uses methods such as data feature selection, missing value processing, and variable discretization, and finally uses the DBSCAN model to cluster the glass artifacts. Based on this, a support vector machine for category prediction of artifacts is constructed, and category prediction, as well as effect testing, is carried out. The DBCSAN model focuses on the rational division of data with uneven density and the specification of the two core numbers of the DBSCAN model, Eps and $MinPts$, in a categorical calculation.

2 CLUSTERING MODEL

The DBSCAN algorithm is a representative density-based clustering algorithm. The algorithm can divide regions with sufficient density into clusters and find clusters of arbitrary shape in a noisy spatial database. Therefore, based on its basic principles, the two core parameters of the DBSCAN algorithm are the scan radius Eps, which is the distance used to determine whether two points are of the same class, and the minimum number of inclusion points $MinPts$, which refers to the minimum number of samples required to be able to determine a cluster. Also, for the sample points to be classified, there are boundary points located at the classification boundary and outlier points located in low-density areas, etc. [2]

However, the settings of the two parameters Eps and $MinPts$ control the categories' size range and the number of clusters, outliers, etc. An objective and reasonable determination of the optimal Eps and the optimal $MinPts$ plays a vital role in the classification results of the dataset. To improve the accuracy of classifying different kinds of glass, we consider an improved DBSCAN model to optimize the problem that the DBSCAN algorithm generates many outliers when dealing with high-dimensional density inhomogeneous data. The basic ideas are: (1) detecting the density of the original data and forming different density partitions of arbitrary shapes and (2) automatically setting parameter values in different density partitions for DBSCAN algorithm clustering. The clustering results are then combined so that the different types of glass can be divided into appropriate subclasses according to their chemical content.

2.1 Density testing analysis

According to the formula step by step to calculate the density formula of node i (Number of elements / Area), the node circular neighborhood of other points k, and finally the average density ρ_i.

$$\rho_i = \frac{|Pts(i)|}{\pi \cdot Eps^2} \tag{1}$$

$$\rho_k = get_\rho(Pts_k), Pts_k \in Pts(i) \tag{2}$$

$$\overline{\overline{\rho_i}} = \frac{\sum_{k=1}^{|Pts(i)|} \rho_k}{|Pts(i)|} \tag{3}$$

$Pts(i)$ is the set of points with center i and radius Eps, $|Pts(i)|$ denotes the number of elements, ρ_i denotes the number of elements / area, k is the other points in the circular neighborhood of node i, and Pts_k is the circle with k as the centre and Eps as the radius. Where ρ_k denotes the density of the Eps neighborhood of Pts_k.

Calculate the density variance and density coefficient of variation for the neighborhood of node i with the following equations:

$$s^2 = \frac{1}{n-1}\left[\sum \rho_i^2 - n(\overline{\overline{\rho_i}})\right], n = |Pts(i)| \tag{4}$$

$$cv_1 = \frac{s}{\rho_i}, s = \sqrt{s^2} \tag{5}$$

The Eps neighborhood of the n nodes is queried to obtain the cv value of each node. The cv values are used to partition the data into several more uniform-density regions of more uniform density for further clustering.

2.2 Calculation of regional parameters

The range rate of change is the range value of a partition divided by the value of the partition behind it. Set the threshold λ. If the range rate of change is more significant than λ, the division point is the critical point between the two data partition boxes; conversely, no data partition exists when the range rate of change is less than λ.

(1) Determine the value of $MinPts$ from the value of cv

(2) Calculate the number of nodes in the Eps neighborhood of each point from Eq. (6)

(3) Calculate the distance between each node and its first $|Pts(i)|$ nearest node according to Eq. Consider the average of Eps, i.e., $E(Eps)$, as the radius of each partition.

$$Eps(i) = \frac{|Pts(i)|}{MinPts} Eps, E(Eps) = \frac{\sum_{i=1}^{n} Eps(i)}{n} \tag{6}$$

3 DATA PROCESSING

3.1 Introduction to the data set

The data set used in this paper is derived from the Chinese Society for Industrial and Applied Mathematics data, with a total of 69 glass artifacts. Each artifacts contains data on the content of 14 chemical components, such as silica, potassium oxide, and calcium oxide. The content statistics are based on a percentage system, i.e., compositional data.

3.2 Treatment of missing values

Based on the data in the dataset, we found that some of the artifacts data need to be completed. We removed data with less than 85% of the total chemical content based on the reasonableness and feasibility of the artifacts analysis, making the final results true and reliable.

3.3 Central logarithmic ratio transformation

As only component data are given in the dataset, making the relevant data meaningless for direct group comparisons, we process the dataset in advance. Using the central log-ratio transformation is possible to move from a quantitative description of the individual components of a whole to a form of quantitative description that allows easy comparison of the components between wholes. The way this change is as follows：

$$clr(x) = \left[\ln\frac{x_1}{g(x)}; \ln\frac{x_2}{g(x)}; \cdots; \ln\frac{x_N}{g(x)}\right] \tag{7}$$

The $g(x)$ denotes the geometric mean.

$$g(x) = \sqrt[N]{x_1 x_2 \cdots x_N} \tag{8}$$

3.4 Data dimensionality reduction

After a correlation analysis and other steps, we calculated correlations between the 14 dimensions of ancient glass artifacts, as shown by the results in Table 1. Noting the significant correlation coefficients between some variables and considering the influence of chemical composition on the classification of artifacts, we used principal component analysis to distill the crucial determining variables from the many variables to measure the influence that chemical composition has on the category of artifacts.

4 CLUSTERING CLASSIFICATION IMPLEMENTATION

4.1 Clustering results

The two core data values for this problem using the DBSCAN model to classify high potassium glass and lead-barium glass were calculated according to the improved algorithm. For $MinPts$ both were 2, i.e., at least two sample points were required to classify into a cluster. For Eps, the value for high potassium glass was 1.2, and for lead-barium glass 0.33, i.e., the distances to determine whether the two points were of the same type were 1.2 and 0.33 respectively. During the data pre-processing process, we extracted the four main data with the most distinctive

features from the fourteen variables utilizing principal component analysis as new variables to analyze the material composition of the two classes of glass, from which we obtained the sub-classification results for high potassium glass and lead-barium glass as shown in Figure 1.

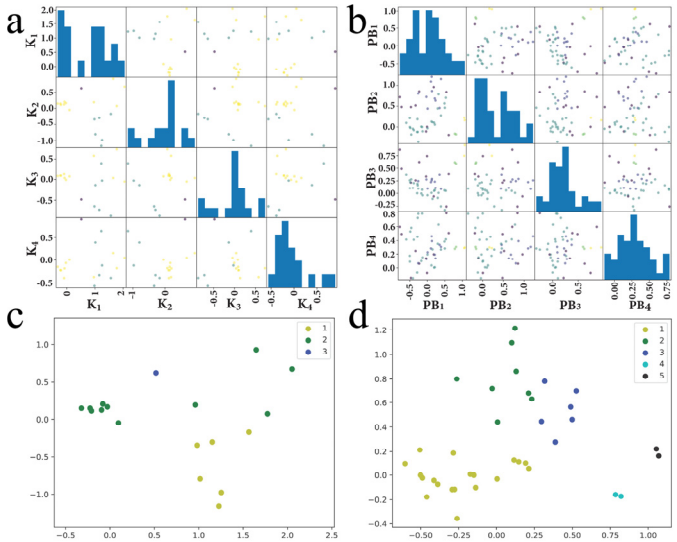


Figure 1: The figure shows the cluster result analysis. Figures **a** and **c** are the four-variable clustering results and final results of high-potassium glass, and Figures **b** and **d** are the four-variable classification results and final results of lead-barium glass, respectively.

4.2 Support vector machine classification

Support Vector Machine (SVM) is a supervised machine learning algorithm, and in this question, we use it for the classification of glass artifacts types. In the general idea, we treat each data as a point in an n-dimensional space, where each feature value is a specific coordinate value, and then we classify the data by finding the hyperplane used to distinguish the two classes. The process is explained in detail on the internet so that we will only go into a bit of detail. The problem was directly analyzed using the "svm" package within the "sklearn" package in Python, and by randomly selecting the training and test sets, we obtained an accuracy of 100% for the classification model, with an F1 score close to 1. The classification results are shown in Table 1 and Table 2.

4.3 Binary Logistic Multi-classification Methods

Logistic regression models can be seen as generalized linear models. For dichotomous logistic regression models, generally, map the linear regression to (0,1) by means of a Sigmoid activation function, e.g., a mapped value greater than 0.5 is one category. Otherwise, it is another. Binary logistic regression is a non-linear regression that introduces a multi-factor probabilistic type of regression, which can investigate the interrelationship between the dependent variable and multiple independent variables. The idea behind the construction of the binary logistic regression model is as follows: we need to continuously train a predictor using the category of glass artifacts of known classification type in Table 1 and Table 2 as the target and the

corresponding 14 chemical components as the parameter variables, and then use the predictor to predict the glass type of the unknown category based on its chemical component data.

4.3.1 Binary Logistic Regression

High potassium glass and lead-barium glass are mutually exclusive events. We take whether the category is high potassium glass as the objective function and set the dichotomous objective to 0 and 1. 0 means it is not high potassium glass, i.e., lead-barium glass, 1 means it is high potassium glass, and the dichotomous dependent variable as a whole can be expressed as $y^{(i)} \in \{0,1\}$. The regression model is：

$$g_\theta(x) = \frac{1}{1 + e^{-\theta^T x}} \tag{9}$$

The equation for the loss function that allows it to be minimized as Eq. (10)

$$h_\theta(x) = -\frac{1}{m}\left[\sum_{i=1}^{m} log\, h_\theta\left(x^{(i)}\right) + \left(1 - y^{(i)}\right) log\left(1 - h_\theta\left(x^{(i)}\right)\right)\right] \tag{10}$$

From this equation, we can distinguish the type of unknown glass (high potassium or lead barium) based on the probability values obtained from the regression results.

4.3.2 Multi-classification tasks

Suppose we want to perform a K-classification task. First divide the task M times, convert the one K-classification task into M classifier tasks for solving, and calculate the encoding $M_i(x)$ for each category corresponding to M classifiers. This is the encoding process. where $i = 1,2,\cdots,K$; $x = -1$or 1. This process is the encoding process.

The samples to be predicted are then provided to these M classifiers, and M results are obtained, producing a string of codes, e.g., $M_{pred}(-1,1,-1,1,-1)$; this string of codes is then compared with the distance between the codes produced by each category, and the category with the lowest distance is considered to be the final result. Common distance calculation methods such as Hemming distance and Euclidean distance are used. The calculation method is referred to Figure 2.

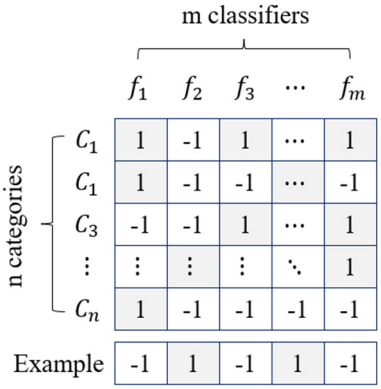


Figure 2: Binary encoding classifier instance

5 EXPERIMENTAL RESULTS AND EVALUATION

Based on the clustering model, this paper results in a subclass classification of high potassium glass and lead-barium glass, as shown in Table 1 and Table 2. In the following, we selected eight unknown classified glass artifacts and applied the above method to predict the artifacts types. The unknown artifacts properties are shown in Appendix 2.

Table 1: Subclassification in high-potassium glass types

Relic number	Category	Relic number	Category
01	First	03 place 1	Second
04	First	03 place 2	Second
05	First	06 place 1	Second
13	First	06 place 2	Second
14	First	22	Second
16	First	27	Second
21	First	07	Second
09	Second	10	Second
12	Second		
12	Second		

Table 2: Subclassification of lead-barium glass types

Relic number	Category	Relic number	Category
20	First	24	First
30 place 1	First	31	First
37	First	29	First
49	First	48	First

51 place 2	First	54	First
30 place 2	Second	50 place 2	Second
2	Second	11	Second
19	Second	41	Second
43 place 2	Second	49	Second
51 place 1	Second	58	Second
32	Third	33	Third
35	Third	45	Third
46	Third	47	Third
55	Third	23	Third
25	Third	28	Third
42 place 1	Third	42 place 2	Third
44	Third	53	Third
34	Third	36	Third
38	Third	39	Third
40	Third	43 place 1	Third
50 place 1	Third	52	Third
54	Third	56	Third
57	Third	8 place 1	Fourth
26 place 1	Fourth	8 place 2	Fifth
26 place 2	Fifth		

5.1 Key performance indicators

After building the support vector machine classifier as well as the binary logistic multi-classifier, for the glass artifacts chemical composition content dataset, we used 90% of the samples as the training set and 10% of the samples as the test set for training and finally successfully output the classification of each glass artifacts.

5.2 Evaluation of generalization capability

In order to prevent possible defects such as single distribution and over-fitting brought by one dataset, we used the data on the chemical composition of labeled glass artifacts intercepted from the network for generalization performance testing. It was found that the prediction accuracy of the new dataset was better, with the accuracy and completeness of the check reaching 100%. The predicted values were compared with the true values, and the fundamental performance indicators of the network were derived, as shown in Table 3.

Table 3: Performance indicators

	Accuracy	Recall	F1 Score
SVM	0.90	0.81	0.95
Logistic	0.91	0.79	0.88

6 CONCLUSION

In the digital era, various industry sectors have digital information technology applications. This paper proposes a classification method for glass cultural relics that can effectively integrate cultural relics data resources for efficient and accurate classification and fast and orderly management of cultural relics. In this paper, we study a classification method for ancient cultural relics by improving DBSCAN model clustering, support vector machine and binary logistic classification methods, and realize the classification of unknown categories of cultural relics through intelligent algorithms and other means. As a whole, the model we use has a high degree of generalisability as well as stability.

For the identification of cultural artifacts, the model algorithm can be used to distinguish between categories of artifacts according to their compositional content and to predict unknown categories of artifacts, which can be helpful for the excavation and classification of ancient glass artifacts. For heritage management, the model can be used to build a database of heritage objects and provide digital support for heritage conservation. For heritage conservation, the model developed in this paper allows for a certain degree of use and organisation of heritage information resources [1], and through the classification of heritage archiving management, can effectively avoid the loss and damage of heritage that may occur in traditional heritage conservation methods.

Acknowledgments: The use and improvement strategies of DBSCAN are mainly carried out by referring to the following articles: Patwary M , Palsetia D , Agrawal A , et al. A new scalable parallel DBSCAN algorithm using the disjoint-set data structure[C]// International Conference for High-Performance Computing, Networking, Storage and Analysis; SC 12. 0.

REFERENCES

[1] Huang Jianli. (2018). Discuss the cultural relics protection measures in the cultural relics management of museums. Comparative Research on Cultural Innovation (04), 37-38.

[2] Jiang, J., Bian, H., & Yang, Y. (2014). Ddbscan: a density detection dbscan algorithm in e-commerce sites evaluation. Biotechnology: An Indian Journal, 10.

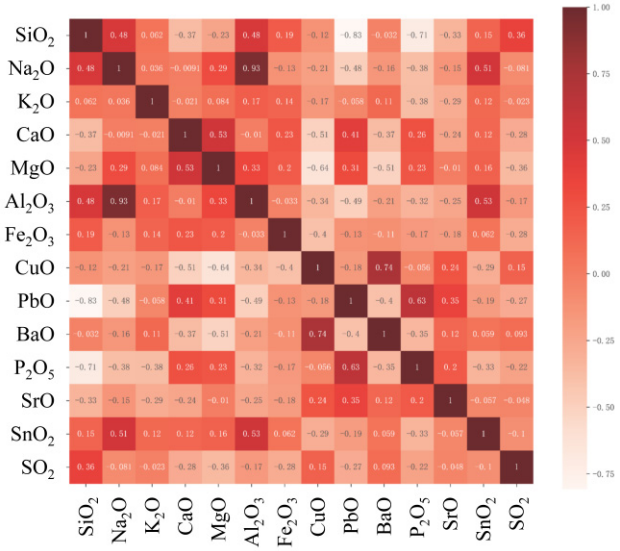
[3] Jing Li. (2011). Research combining ant colony algorithm and division-based DBSCAN clustering algorithm (Master's dissertation, Northeast Normal University).

[4] Liu Juying. (2020). Analysis of museum digital construction and cultural relic management characteristics. Cultural Relics Identification and Appreciation (23), 110-112.

[5] Zhang Changyong & Han Liang. (2022). Lidar obstacle detection based on the optimized DBSCAN. Progress in laser and optoelectronics (12), 516-524.

APPENDIX

Appendix 1: Correlation coefficient heat map.



Appendix 2: The unknown artifacts properties.

Relic number	Weathering	SiO_2	K_2O	PbO	BaO	⋯	P_2O_5
A1	No	78.45	0	0	0	⋯	1.06
A2	Yes	37.75	0	34.30	0	⋯	14.27
A3	No	31.95	1.36	39.58	4.69	⋯	2.68
A4	No	35.47	0.79	24.28	8.31	⋯	8.45
A5	Yes	64.29	0.37	12.23	2.16	⋯	0.19
A6	Yes	93.83	1.35	0	0	⋯	0.21
A7	Yes	90.83	0.98	0	0	⋯	0.13
A8	No	51.12	0.23	21.24	11.34	⋯	1.46