ISSN 2616-5775 Vol. 5, Issue 13: 89-95, DOI: 10.25236/AJCIS.2022.051314

A study on the composition analysis and identification of ancient glass products based on SVM model

Hui Xu*

Guangzhou Vocational and Technical University of Science and Technology, Guangzhou, Guangdong, 510000, China
**Corresponding author: 1015817369@qq.com*

Abstract: *This paper focuses on the composition of ancient glass artifacts and the identification of types based on composition. Firstly, this study is to classify and summarize unknown ancient glass artifacts and perform sensitivity analysis, and secondly, to analyze the relationship between the chemical composition of different glass types and to derive the correlation differences in relation to the glass types. For problem one, a support vector machine classification model was developed, the sample data set was trained, the model obtained after training was saved, and then the data related to each chemical composition of the unknown type of ancient glass artifacts were imported, and finally their classification results were obtained, and a sensitivity analysis was performed in 10 groups for the highly significant indicator factor of silica content, when the silica content changed , while other factors remained constant, it was found that the classification results showed a great difference and the prediction probability changed significantly, so the established support vector machine classification model was considered to be highly sensitive. For the second problem, the internal relationship between the various chemical components of different types of ancient glass artifacts was analyzed, and here the Pearson correlation analysis was done between the different types of chemical components and each other, and several types were combined and the cross-sectional comparison was done by the heat diagram between them, and it was concluded that the correlation between barium oxide and copper oxide, phosphorus pentoxide and calcium oxide, barium oxide and sulfur dioxide in the type of lead barium The color effect of the thermogram is significant, while the correlation between phosphorus pentoxide and iron oxide, calcium oxide and potassium oxide, and strontium oxide and phosphorus pentoxide in the high potassium type is significant, and the color effect of the thermogram is significant.*

Keywords: *Classification aggregation; Correlation analysis; Support vector machines; Classification models*

1. Introduction

The main cause of weathering is the effect of water. If glass and glass products are stored in a humid environment for a long time, or if they are not dried after rain and water immersion, the water that adheres to the surface of the glass hydrolyzes the soluble component of the glass, sodium silicate, producing sodium hydroxide and silica colloids. The silica colloid is a colloid that has very little solubility in water and acid, so it no longer participates in chemical reactions and settles on the surface of the glass. As weathering becomes more severe, white frost is generated on the surface of the glass, thus losing transparency and even producing flat glass sticking to each other. When glass is weathered, its surface will show different physical changes with the degree of weathering, such as fogging, mildew, iridescence, glass bonding and reduction of visible light transmission. The weathering mechanisms include the adsorption of water on the glass surface, the exchange of hydrogen ions or hydrated hydrogen ions with alkali metal ions in the glass, and the combination of alkali metal ions on the glass surface with hydroxide ions to form an alkali solution, which in turn destroys the structure of the glass itself [1].

This study intends to address the following questions.

Here is a batch of data related to ancient glass products in China, which archaeologists have classified into two types of high potassium glass and lead-barium glass based on the chemical composition of these artifact samples and other testing means. We need to establish a mathematical model to analyze and model the relevant data to solve the following problems.

1) Classification and sensitivity analysis for unknown categories of glass artifacts.

2) Analyze the correlation between the chemical composition of different types of glass artifacts and compare the differences.

2. Model building and solving

2.1 Modeling and solving for Problem 1

2.1.1 Establishment of SVM model for SVM classification

Support Vector Machine [2][3](SVM) is a kind of generalized linear classification machine that classifies data according to supervised learning. It assigns labels to objects through example learning, which is one of the traditional machine learning algorithms. Its decision boundary is the maximum margin hyperplane [4][5][6] solved for the learning sample(Figure 1). It is a machine learning method based on statistical learning theory VC, which is based on dimension theory and structural risk minimization principle. It shows many unique advantages in solving small sample, nonlinear and high-dimensional pattern recognition problems, and to a large extent overcomes problems such as dimensional disaster and overlearning [7][8]. Given the training sample dataset $Q=\{[v_1,c_1],[v_2,c_2],U,[v_i,c_l]\}\epsilon(\varrho\times Y)^l$, where $v_i\epsilon\varrho=R^n$, each point vi consists of n attribute features, $c_i\epsilon Y=\{-1,1\},i=1,U,l$. Find a real-valued function t(x) on R to facilitate the classification function

$$g(x)=sign(t(x)) \tag{1}$$

The problem of inferring the value c corresponding to any one pattern x is a classification problem [9].

For the labeled two groups of vectors, an optimal segmentation hypersurface is given to divide the two groups of vectors into two sides, so that the nearest vector in the two groups of vectors is as far away from the hyperplane as possible. The length of the vector $\omega=[w_1\ w_2\ \cdots\ w_n]^T$ is:

$$\|\omega\|=\sqrt{w_1^2+w_1^2+\cdots+w_n^2} \tag{2}$$

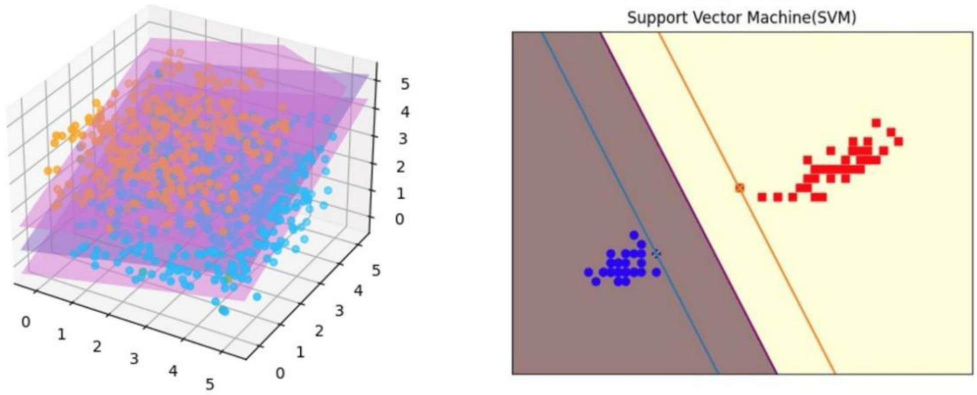

Figure 1: Support vector machine classification chart

Since all vectors on the hyperplane have the same projection length in the normal vector direction, denoted as I, the distance from the origin to this hyperplane, the absolute value of the inner product of any vector $v=[v_1\ v_2\ \cdots\ v_n]^T$ on the hyperplane and this normal vector ω is the product of its projection in the normal vector direction and the modulus length of the normal vector, i.e.

$$|\omega^T v|=|w_1v_1+w_2v_2+\cdots+w_nv_n|=aI \tag{3}$$

Since both a and I are fixed values once the normal vector and hyperplane are fixed, aI in the above equation can be rewritten as -b, and the meaning given to b>0 is that the vector v projects in the direction opposite to the hyperplane in the normal vector, and the hyperplane equation is obtained as follows

$$\omega^T v + b = 0 \tag{4}$$

and call where b is the intercept of the hyperplane[10].

Assuming that the hyperplane can correctly classify the sample data set training into two classes, i.e., $[v_i, c_i] \epsilon Q$ the samples of the same class all fall on the same side of the classification hyperplane, the sample set is said to be linearly divisible, i.e., it satisfies

$$\begin{cases} \omega^T v_i + b \geqslant 1 & c_i = 1(i = 1, 2, \cdots, l) \\ \omega^T v_i + b \leqslant -1 & c_i = -1(i = 1, 2, \cdots, l) \end{cases} \tag{5}$$

These training sample points are called support vectors, and sample points vi are defined. The interval from the classification hyperplane referred to in Equation (5) is

$$\psi_i = c_i(\omega v_i + b) = |\omega v_i + b| \tag{6}$$

The and in equation (7) are normalized, i.e., the original w and b are replaced by $\frac{\omega}{\|\omega\|}$ and $\frac{b}{\|b\|}$, respectively, and the normalized interval defined $d = \frac{\omega v_i + b}{\|\omega\|}$ is called the set interval. The sum of the distances of two dissimilar support vectors to the hyperplane is $d = \frac{2}{\|\omega\|}$.

By finding the partitioned hyperplane with the maximum set interval, i.e.

$$\begin{cases} \max\limits_{\omega, b} \frac{2}{\|\omega\|} \\ s.t.\ c_i(\omega^T v_i + b) \geqslant 1, i = 1, 2, \cdots, m \end{cases} \tag{7}$$

This equates to

$$\begin{cases} \min\limits_{\omega, b} \frac{\|\omega\|^2}{2} \\ s.t.\ \ c_i(\omega^T v_i + b) \geqslant 1, i = 1, 2, \cdots, m \end{cases} \tag{8}$$

This is the basic model of support vector machine classification SVM.

However, in practical applications, most of the problems are nonlinear, which can not be helped by linearly separable SVMs. For such linearly indistinguishable problems, the common method is to transform the sample mapping of the original infant space into the high-dimensional $Hilbert$ feature space H through the nonlinear mapping $\varPhi: R^d \to \gamma(x)$, and then construct the optimal classification hyperplane in the high-dimensional $Hilbert$ feature space. In addition, as in the case of linearly separable SVMs, considering that there are still linearly indistinguishable cases caused by a small number of samples after the nonlinear mapping to the high-dimensional feature space, we also consider introducing relaxation variables, i.e., transforming them into a new training set in the $Hilbert$ feature space.

$$Q'=\{[v_1,c_1],[v_2,c_2],U,[v_i,c_l]\}=\{[\gamma(v_1),c_1],U,[\gamma(v_i),c_l]\} \tag{9}$$

such that it is linearly divisible in the Hilbert feature space H. Then the hyperplane H is found on the feature space Hilbert, and this hyperplane $[\omega\times\gamma(x)]+b=0$ can be rigidly partitioned into the training set Q'. The original problem is transformed into

$$\min\ \frac{1}{2}\|\omega\|^2 \tag{10}$$

$$s.t.\ \ c_i\{[\omega\times\gamma(v_i)]+b\}\geqslant 1,i=1,U,l \tag{11}$$

Using the kernel function K satisfies

$$K(v_i,v_j)=[\gamma(v_i)\cdot\gamma(v_j)] \tag{12}$$

2.1.2 Specific solution steps and results

Analysis steps

1) Build the support vector machine classification model by applying it to the training set data.

2) Apply the established support vector machine classification model to the training and test data to obtain the classification evaluation results of the model.

3) Because of the randomness of the support vector machine classification model, the result of each calculation is not the same, if the training model is saved, the subsequent sample data can be directly substituted into the training model for classification.

Note: The support vector machine classification model cannot get a definite equation like the traditional model, so the model is usually evaluated by testing the data classification effect.

Table 1 is the parameters of the model.

Table 1: SVM parameters

Parameter Name	Parameter Value
Training time (duration)	0.012s
Data shuffle	Yes
Data Slicing	0.7
Penalty Factor	1
Cross-validation	No
Kernel functions	$linear$
Kernel function coefficients	$scale$
Maximum number of terms in the kernel function	3
Nuclear function constants	0
Error convergence condition	0.001
Multi-category integration strategy	ovr
Maximum number of iterations	1000

From Figure 2 and Table 2, it can be seen that the prediction results have a high degree of confidence and are more reasonably accurate.

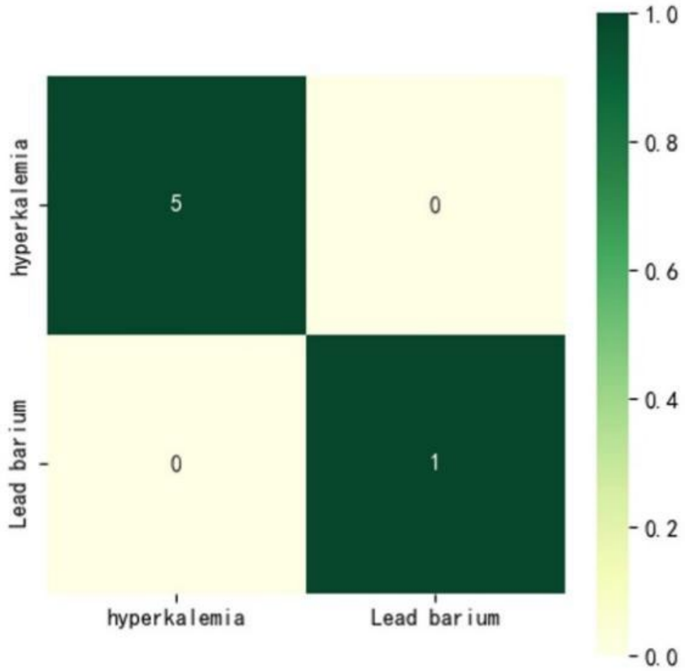

Figure 2: Confusion matrix heat map

Table 2: Prediction evaluation results of training test data

Predicted results X	Type	Predicted outcome probability_Barium lead	Predicted outcome probability_high potassium
Lead Barium	Lead Barium	0.997687	0.002313
Lead Barium	Lead Barium	0.999995	0.000005
high potassium	high potassium	0.106106	0.893894
Lead Barium	Lead Barium	0.999158	0.000842
Lead Barium	Lead Barium	0.897512	0.102488
Lead Barium	Lead Barium	0.999541	0.000459

Note: The probability of the predicted result is retained to 6 decimal places.
Table

SiO_2	Na_2O	K_2O	CaO	MgO	Al_2O_3
3.72	0	0.4	3.01	0	1.18
16.71	0	0	1.87	0	0.45
87.05	0	5.19	2.01	0	4.06
29.15	0	0	1.21	0	1.85
53.33	0.8	0.32	2.82	1.54	13.65
35.78	0	0.25	0.78	0	1.62
Fe_2O_3	CuO	PbO	BaO	P_2O_5	
0	3.6	29.92	35.45	6.04	
0.19	0	70.21	6.69	1.77	
0	0.78	0.25	0	0.66	
0	0.79	41.25	15.45	2.54	
1.03	0	15.71	7.31	1.1	
0.47	1.51	46.55	10	0.34	

2.1.3 Perform specific identification and sensitivity analysis

Subsequently, we save this support vector machine SVM model obtained after training, and then import the sample data given in the question for actual identification, first by filling in the vacant values, and then by direct import of the identification. By comparing the probabilities of the predicted results, it can be concluded that they are more reliable and more accurate (Table 3).

Table 3: Sensitivity analysis results

Artifact Number	Predicted Results_N	Predicted results_Barium lead	Predicted results_high potassium	SiO_2	Na_2O	K_2O	CaO
A1	High potassium	0.14803	0.85197	78.45	0	0	6.08
A2	Barium lead	0.98932	0.01068	37.75	0	0	7.63
A3	Barium lead	0.99678	0.00322	31.95	0	1.36	7.19
A4	Barium lead	0.97129	0.02871	35.95	0	0.79	2.89
A5	Barium lead	0.77455	0.22545	64.29	1.2	0.37	1.64
A6	High potassium	0.07169	0.92831	93.17	0	1.35	0.64
A7	High potassium	0.09681	0.90319	90.83	0	0.98	1.12
A8	Barium lead	0.93008	0.06992	51.12	0	0.23	0.89

2.1.4 Sensitivity test analysis for the model

By SVM support vector machine classification model to predict the probability of unknown samples results, then here will be selected previously studied and analyzed for the results of the existence of a large number of significant impact of silica content as an indicator factor of the significance analysis, other indicator factors remain unchanged, then for A1, you can directly follow each time minus 5 to do 10 groups of data for analysis(Table 4).

Table 4: Sensitivity test analysis results

Artifact Number	Predicted Results_N	Predicted results_Barium lead	Predicted results_high potassium	SiO_2	Na_2O	K_2O	CaO
A1	High potassium	0.14803	0.85197	78.45	0	0	6.08
A1	High potassium	0.14803	0.85197	73.45	0	0	6.08
A1	High potassium	0.14803	0.85197	68.45	0	0	6.08
A1	High potassium	0.14803	0.85197	63.45	0	0	6.08
A1	High potassium	0.14803	0.85197	58.45	0	0	6.08
A1	High potassium	0.14803	0.85197	53.45	0	0	6.08
A1	High potassium	0.14803	0.85197	48.45	0	0	6.08
A1	High potassium	0.14803	0.85197	43.45	0	0	6.08
A1	High potassium	0.14803	0.85197	38.45	0	0	6.08
A1	High potassium	0.14803	0.85197	33.45	0	0	6.08

3. Conclusion

The model developed in this study, support vector machine SVM it is a machine learning algorithm with a solid theoretical foundation. It simplifies the usual classification and regression problems. The complexity of the operation depends on the number of support vectors rather than the dimensionality of the sample data space set. A small number of support vectors determines the final result and is insensitive to outliers, which not only helps us to get hold of key samples and eliminate a large number of redundant samples, but also predestines the method to be not only simple, but also robust and generalizable. Model in the image processing and presentation, we use python and SPSS in the visualization processing algorithm and tools, combined with the data based on the results of the problem graphically, more intuitive, clear and organized.

References

[1] Liu Yuanxin, Jiang Quan. Selection of evaluation method for weathering degree of flat glass [J]. China Building Materials Science and Technology,1992,1(6):29-33.

[2] Cristianini N, Taylor J S. An introduction to support vector machines and other kernel-based learning methods [M]. Translated by LI Guo-zheng, WANG Meng, ZENG Hua-jun. Beijing: Publishing House of Electronics Industry, 2004

[3] ZHANG Xue-gong. Statistical learning theory and support vector machines [J]. Acta Automatica Sinica, 2000, 26(1): 32-41.

[4] Vapnik, V. Statistical learning theory. 1998(Vol.3).New York, NY:Wiley, 1998: Chapter 10-11, pp. 401-492

[5] Zhou Zhihua. Machine Learning. Beijing: Tsinghua University Press, 2016: pp 121-139, 298-300

[6] Li Hang. Statistical Learning Methods. Beijing: Tsinghua University Press, 2012: Chapter 7, pp.95-135

[7] VAPNIK V N. The nature of statistical learning theory [M]. Translated by ZHANG Xue-gong. Beijing: Tsinghua University Press, 2000.

[8] VAPNIK V N. Statistical Learning Theory [M]. Translated by XU Jian-hua, ZHANG Xue-gong. Beijing: Publishing House of Electronics Industry, 2004.

[9] Mao Shisong, Wang Jinglong, Pu Xiaolong, et al. Advanced Mathematical Statistics (Second Edition) [M]. Beijing: Higher Education Press, 2006

[10] Byun H, Lee S W. Applications of support vector machines for pattern recognition: A survey [C]//International workshop on support vector machines. Springer, Berlin, Heidelberg, 2002: 213-236.