Modeling and Simulation 建模与仿真, 2022, 11(6), 1627-1635
Published Online November 2022 in Hans. http://www.hanspub.org/journal/mos
https://doi.org/10.12677/mos.2022.116153

Hans汉斯

古代玻璃制品成分分析和鉴定的研究

马双宇，朴凤贤，李郡霆，汪义坤，袁宇航，臧陶亮

沈阳航空航天大学，辽宁 沈阳

收稿日期：2022年10月24日；录用日期：2022年11月17日；发布日期：2022年11月24日

摘　要

古代玻璃种类繁多且易受环境影响而风化，因此需要对古代玻璃制品的化学成分数据分析，研究有无风化玻璃制品成分的变化规律，以并探索亚分类方法，进而可以根据未知分类的文物化学成分对文物进行准确的分类。本文通过使用K-means算法和BP神经网络结合的方式对玻璃制品进行亚分类划分，之后根据亚分类种类进行风化前后成分的预测；通过RUSBoost机械学习算法，70%的数据作为训练集，15%的数据作为测试集，其余部分作为预测集，来进行玻璃制品的种类鉴定。这些模型相互之间配合紧密，所得结果依次递进，使最终求解真实可靠。模型充分联系实际，具有很好的通用性和推广性。

关键词

K-Means算法，BP神经网络，RUSBoost机械学习

Study on Composition Analysis and Identification of Ancient Glass Products

Shuangyu Ma, Fengxian Piao, Junting Li, Yikun Wang, Yuhang Yuan, Taoliang Zang

Shenyang Aerospace University, Shenyang Liaoning

Received: Oct. 24th, 2022; accepted: Nov. 17th, 2022; published: Nov. 24th, 2022

Abstract

There are many kinds of ancient glass and it is easy to be weathered by the environment. Therefore, it is necessary to analyze the chemical composition data of ancient glass products, study the changing rules of the composition of weathered glass products, and explore subclassification methods, so as to accurately classify cultural relics according to the chemical composition of cultural relics unknown. This paper uses K-means algorithm and BP neural network to subclassify glass products, and then predicts the composition before and after weathering according to the subclassification types. Through the RUSBoost mechanical learning algorithm, 70% of the data is used

文章引用: 马双宇, 朴凤贤, 李郡霆, 汪义坤, 袁宇航, 臧陶亮. 古代玻璃制品成分分析和鉴定的研究[J]. 建模与仿真, 2022, 11(6): 1627-1635. DOI: 10.12677/mos.2022.116153

as the training set, 15% of the data is used as the test set, and the rest of the data is used as the prediction set to identify the types of glass products. The classification rules and identification of ancient glass products are divided. These models cooperate closely with each other, and the results are progressive in turn, which makes the final solution true and reliable. The model is fully connected with practice and has good generality and popularization.

Keywords

K-Means Algorithm, BP Neural Network, RUSBoost Mechanical Learning

Copyright © 2022 by author(s) and Hans Publishers Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access

1. 引言

丝绸之路是古代中西方文化交流的通道，其中玻璃是早期贸易往来的宝贵物证。古代玻璃种类繁多且易受环境影响而风化，因此需要对古代玻璃制品的化学成分数据分析，研究有无风化玻璃制品成分的变化规律，以及高钾、铅钡两种玻璃类型的化学成分统计规律，并探索亚分类方法，进而可以根据未知分类的文物化学成分对文物进行准确的分类。

由于玻璃文物表面风化与其类型、纹饰和颜色具有一定关系。首先确定各因素对表面风化的影响程度。之后采用定量分析，根据化学成分将玻璃文物进行分类，并根据不同玻璃类型进行风化前后的成分分析。关于对玻璃文物风化前的成分预测，本论文采用先分类后预测的方法，提高预测准确性。通过RUBoost 机械学习算法根据玻璃文物的化学成分推断玻璃文物的类别。

2. 基于 K-Means 算法和 BP 神经网络的亚分类方法

2.1. 实验数据及处理

本文研究所用的实验数据包括三个：表 1：58 个文物的纹饰、类型、颜色和表面风化情况的信息；表 2：58 个文物不同部位的化学成分；表 3：8 个已知化学成分和风化情况，但未知玻璃类型的文物。

将表单 2 中的化学成分百分比之和大于 100%和小于 85%的数据去除，将空白项进行填零处理。

表单 1 部分数据：

Table 1. Form 1 partial data
表 1. 表单 1 部分数据

文物编号	纹饰	类型	颜色	表面风化
01	C	高钾	蓝绿	无风化
02	A	铅钡	浅蓝	风化
03	A	高钾	蓝绿	无风化
04	A	高钾	蓝绿	无风化
05	A	高钾	蓝绿	无风化
06	A	高钾	蓝绿	无风化

Continued

07	B	高钾	蓝绿	风化
08	C	铅钡	紫	风化
09	B	高钾	蓝绿	风化

表单 2 部分数据：

Table 2. Form 2 partial data
表 2. 表单 2 部分数据

文物采样点	二氧化硅(SiO_2)	氧化钾(K_2O)	氧化钙(CaO)	氧化镁(MgO)	氧化铝(Al_2O_3)	氧化铁(Fe_2O_3)	氧化铜(CuO)	氧化铅(PbO)	五氧化二磷(P_2O_5)
01	69.33	9.99	6.32	0.87	3.93	1.74	3.87	0	1.17
02	36.28	1.05	2.34	1.18	5.73	1.86	0.26	47.43	3.57
03 部位 1	87.05	5.19	2.01	0	4.06	0	0.78	0.25	0.66
03 部位 2	61.71	12.37	5.87	1.11	5.5	2.16	5.09	1.41	0.7
04	65.88	9.67	7.12	1.56	6.44	2.06	2.18	0	0.79
05	61.58	10.95	7.35	1.77	7.5	2.62	3.27	0	0.94
06 部位 1	67.65	7.37	0	1.98	11.15	2.39	2.51	0.2	4.18
06 部位 2	59.81	7.68	5.41	1.73	10.05	6.04	2.18	0.35	4.5
07	92.63	0	1.07	0	1.98	0.17	3.24	0	0.61

表单 3 部分数据：

Table 3. Form 3 partial data
表 3. 表单 3 部分数据

文物编号	表面风化	二氧化硅(SiO_2)	氧化钙(CaO)	氧化铝(Al_2O_3)	氧化铁(Fe_2O_3)	氧化铜(CuO)	氧化铅(PbO)	氧化钡(BaO)	五氧化二磷(P_2O_5)
A1	无风化	78.45	6.08	7.23	2.15	2.11	0	0	1.06
A2	风化	37.75	7.63	2.33	0	0	34.3	0	14.27
A3	无风化	31.95	7.19	2.93	7.06	0.21	39.58	4.69	2.68
A4	无风化	35.47	2.89	7.07	6.45	0.96	24.28	8.31	8.45
A5	风化	64.29	1.64	12.75	0.81	0.94	12.23	2.16	0.19
A6	风化	93.17	0.64	1.52	0.27	1.73	0	0	0.21
A7	风化	90.83	1.12	5.06	0.24	1.17	0	0	0.13
A8	无风化	51.12	0.89	2.12	0.00	9.01	21.24	11.34	1.46

2.2. 相关性分析

为判断玻璃文物表面风化与其类型、纹饰和颜色是否具有关系，需要在计算皮尔逊相关系数前计算

各个变量之间的显著关系 p 值。两者之间的相关关系可能只是偶然因素引起的，所以我们要对两个变量之间的相关关系的显著性水平进行判断。

将表 1 中的文字数据替换为数值矩阵形式。首先对对两变量之间是否具有统计上的显著关系(p 值严格小于 0.01)进行检验。通过表 4 可以得出纹饰对表面风化和颜色对表面风化无显著关系。之后在计算每两个变量之间的皮尔逊系数，之后画出相关性热力图。

$$\rho_{X,Y}=\frac{cov(X,Y)}{\sigma_X\sigma_Y}=\frac{E\left((X-\sigma_Y)(Y-\sigma_X)\right)}{\sigma_X\sigma_Y}=\frac{E(XY)-E(X)E(Y)}{\sqrt{E\left(X^2\right)-E^2(X)}\sqrt{E\left(Y^2\right)-E^2(Y)}} \tag{4}$$

Table 4. P values
表 4. P 值

P 值	纹饰	玻璃类型	颜色	表面风化
纹饰	0.000	0.006	0.001	0.384
玻璃类型	0.006	0.000	0.000	0.008
颜色	0.001	0.000	0.000	0.885
表面风化	0.384	0.008	0.885	0.000

通过图 1 我们不难看出：玻璃类型对于表面风化和颜色具有较强的相关性。根据古代玻璃的制作，是根据不同的玻璃类型和颜色确定纹饰进行加工的，因此纹饰应该与颜色和玻璃类型呈负相关，即颜色和玻璃类型影响纹饰。

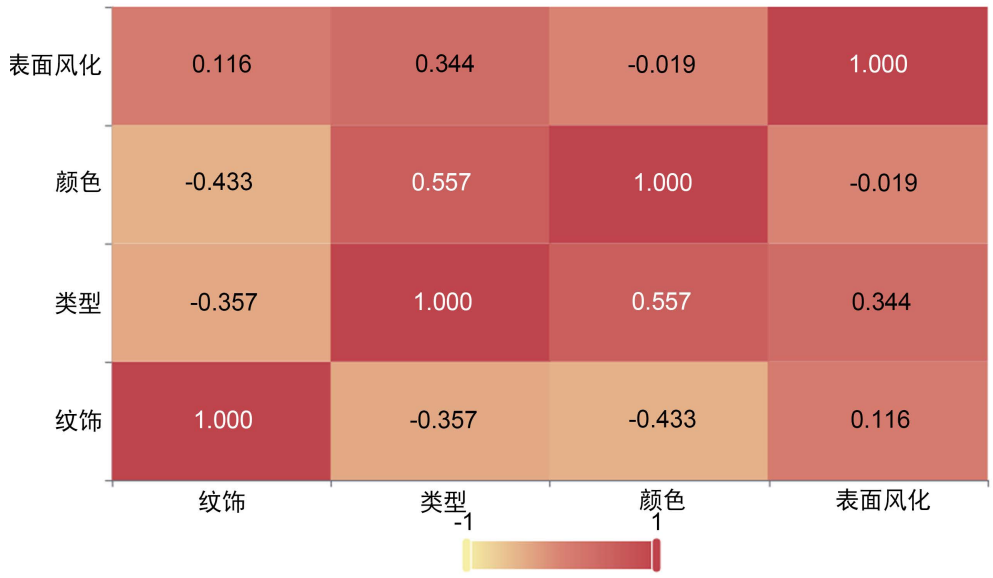

Figure 1. Correlation thermal map
图 1. 相关性热力图

2.3. 数据分析

首先我们根据表 2 中的数据将玻璃分为高钾和铅钡对其化学成分均值以三维柱状图形式呈现。

通过图 2 我们可以发现高钾玻璃主要是由大量二氧化硅元素组成；铅钡玻璃是由基本等量的二氧化

硅和氧化铅元素以及部分氧化钡元素组成，因此可以通过二氧化硅和氧化铅以及氧化钡元素含量来对高钾玻璃和铅钡玻璃进行区分。

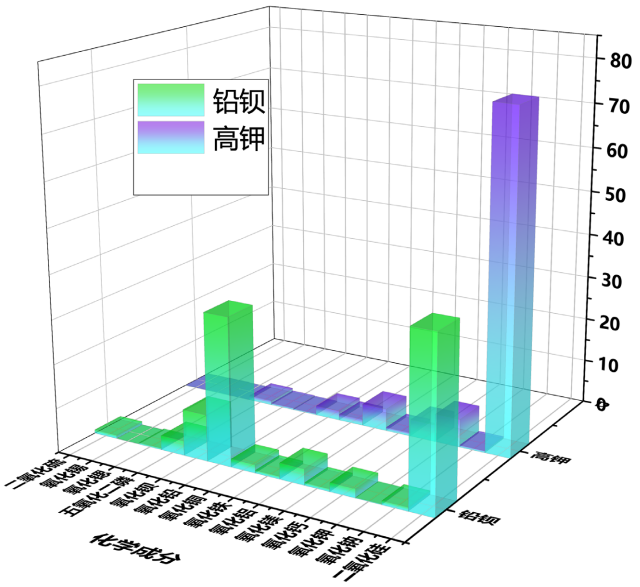

Figure 2. Three-dimensional bar chart of high potassium lead and barium content
图 2. 高钾铅钡元素含量三维柱状图

2.4. 亚类划分及预测

在高钾和铅钡玻璃的前提下，进行亚类分析。通过 K-Means 和 BP 神经网络在高契合度的情况下，兼顾分类的多样性和准确性下，得到 $K = 8$ 时的效果最好，BP 神经网络的检测准确度为 95.5%。

2.4.1. 模型建立

目前经常被人们使用的检验聚类有效性的指标有 XB 指标(Xie-Beni)、KL 指标(Krzanowski-Lai)、Sil 指标 (Silhouette)、DB 指标(Davies-Bouldin)、IGP 指标(In-Group Proportion)、BWP 指标(Between-Within Proportion)等。其中 BWP 是一种有效的聚类指标，该指标依据簇内相似与簇间相异两方面来评判聚类结果[1] [2]。如果 $X = \{X_1, X_2, \cdots, X_n\}$ 为聚类数据集，$k = (X, R)$为聚类空间，如果数据集中的 n 个样本分为 c 类，则最小类间距离、类内距离，BWP 指标判别函数公式如下。

最小类间距离：

$$b(i,j) = \min_{1\le m\le c, m\ne j}\left(\frac{1}{n_m}\sum_{p=1}^{n_m}\left\|x_p^m - x_i^j\right\|^2\right) \tag{2}$$

类内距离：

$$w(i,j) = \frac{1}{n_j - 1}\sum_{q=1,q\ne i}^{n_j}\left\|x_p^m - x_i^j\right\|^2 \tag{3}$$

BWP 指标：

$$\mathrm{BWP}(k) = \sum_{j=1}^{k}\sum_{i=1}^{n_i}\mathrm{bwp}(i,j)\ \ (K\text{ 为类簇},\ \ \mathrm{bwp}(i,j) = \frac{b(i,j) - w(i,j)}{b(i,j) + w(i,j)}) \tag{4}$$

其中 BWP 指标值越高，聚类性能越好，BWP 指标确定 k 值。

对于单个样本来说，BWP 指标值越高，则意味着聚类性能就越好。对大量样本来说，取 BWP 的平均值来观察聚类性能，BWP 平均值越高，数据集的聚类性能就会越好。其中 $\text{avg}_{\text{BWP}(k)}$ 为所有样本的 BWP 平均值，k_{opt} 为聚类数，计算公式如下：

$$\text{avg}_{\text{BWP}(k)} = \frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n_i}\text{bwp}(i,j) \tag{5}$$

$$k_{opt} = \arg\max_{2\le k<n}\left(\text{avg}_{\text{BWP}(k)}\right) \tag{6}$$

本文提出一种基于最大最小距离和 BWP 指标相结合的 K-Means 算法[1] [2]，以下命名为 UPK-Means，改进后算法的基本思路如下：假如数据集 $X=\{X_1,X_2,X_3,\cdots,X_n\}$ 可分为 K 类，K 数目未知。首先在 X 中找出距离最远的两个样本点 C_1、C_2，结合给定的 θ 值，根据最大最小距离原则确定余下的 K-2 个聚类中心，然后把样本点分配给距离最近的聚类中心，当数据集中的每个样本点都有确定的一个类别时，聚类结束。参数值 θ 不同，聚类效果也会有所不同，最后利用 $\text{avg}_{\text{BWP}(k)}$ 评价聚类性能，$\text{avg}_{\text{BWP}(k)}$ 值越高，聚类性能越好。

2.4.2. 模型求解

将表 2 中处理后的数据导入 SPSSPRO 中，使用 K-Means 聚类算法后得出玻璃文物的亚分类规律。之后将各分类的化学成分占比以图 3 形式表示。

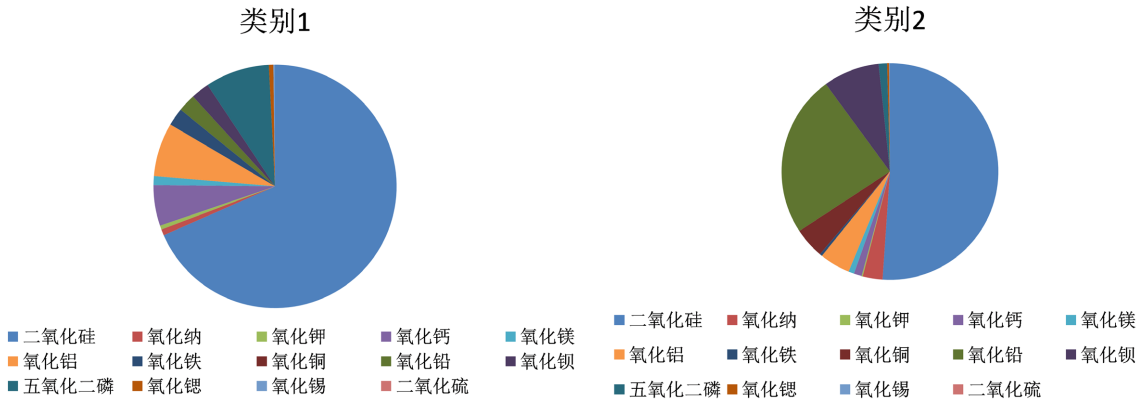

Figure 3. Partial results of subclassification
图 3. 亚分类的部分结果

在根据分化点检测的数据预测风化前的化学成分之前，需要将该风化点根据检测数据划分到不同亚类当中，之后根据再根据该亚类的风化前后化学成分变化规律来进行预测。

风化前后变化率 η

$$\eta = \frac{\alpha_{\text{未风化高钾}} - \alpha_{\text{风化高钾}}}{\alpha_{\text{未风化高钾}}} \tag{7}$$

由于铅钡玻璃中二氧化硅，氧化铅和氧化钡含量在各亚类中无明显差别，因此为了提高预测精度，这几项含量采用所有铅钡玻璃进行变化率的计算；在亚类中有明显区别的采用同类进行变化率计算。使用此方法得到的预测精度较传统方法有所提高。

3. 基于 RUSBoost 算法的玻璃制品鉴定

3.1. RUSBoost 模型的建立

RUSBoost 算法首先利用初始训练集训练出一个基学习器；再根据当前基学习器的训练误差调整每个

训练样本的分布权重，通过增大被误分类样本的权重，使其在后续训练过程中获得更多关注；然后利用调整后的样本训练出下一个基学习器；如此反复迭代，直至生成 T 个基学习器；最后将根据上述 T 个基学习器的加权投票结果来预测未标记样本[3] [4]。

假设训练数据集为 $S=\{(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)\}$，其中 (x_i,y_i) 是一组样本数据， RUSBoost 算法在第 t 次迭代中生成的基学习器为 h_t，每组样本 X_i 对应的预测结果为 $h_t(x_i)$，并且该样本在该次迭代中对应的权重为 $D_t(i)$。因此，在第 t 次迭代中所有样本构成离散分布。

离散分布为 $\Lambda_t=\{D_t(1),D_t(2),\cdots,D_t(m)\}$，其中 $Z_t=\sum_{i=1}^{m}D_t(i)=1$。同时，每个基学习器 h_t 的错误率定义为在分布 Λ_t 下其预测错误的概率，即

$$\epsilon_t=P_{x\sim\Lambda_t}\left[h_t(x)\neq y\right] \tag{8}$$

其中， α_t 是基学习器 h_t 的权重。而 RUSBoost 算法的预测结果表示为 $\operatorname{sign}\left[H(x)\right]$。

RUSBoost 算法是通过最小化指数损失函数 $l_{\exp}(H\mid\Lambda)$ 达到贝叶斯最优错误率：

$$\begin{aligned}l_{\exp}(\alpha_t h_t\mid\Lambda_t)&=E_{x\sim\Lambda_t}\left\{\mathrm{e}^{-y\alpha_t h_t(x)}\right\}\\&=\mathrm{e}^{-\alpha_t}\mathbf{P}_{x\sim\Lambda_t}\left[h_t(x)=y\right]+\mathrm{e}^{\alpha_t}\mathbf{P}_{x\sim\Lambda_t}\left[h_t(x)\neq y\right]\\&=\mathrm{e}^{-\alpha_t}(1-\epsilon_t)+\mathrm{e}^{\alpha_t}\epsilon_t\end{aligned} \tag{9}$$

令导数为零：

$$\frac{\partial l_{\exp}(\alpha_t h_t\mid\Lambda_t)}{\partial\alpha_t}=0 \tag{10}$$

则最优权重计算为：

$$\alpha_t=\frac{1}{2}\ln\left(\frac{1-\epsilon_t}{\epsilon_t}\right) \tag{11}$$

RUSBoost 算法在 $t+1$ 次迭代之后获得的基学习器线性组合为 $H_{t+1}=H_t(x)+a_{t+1}h_{t+1}$，其指数损失函数通过泰勒展开可以近似为：

$$\begin{aligned}l_{\exp}(H_{t+1}\mid\Lambda)&=E_{x\sim\Lambda}\left\{\mathrm{e}^{-y\left[H_t(x)+a_{t+1}h_{t+1}(x)\right]}\right\}\\&\cong E_{x\sim\Lambda}\left\{\mathrm{e}^{-yH_t(x)}\left[1-y\alpha_{t+1}h_{t+1}(x)+\frac{1}{2}y^2\alpha_{t+1}^2h_{t+1}^2(x)\right]\right\}\\&=E_{x\sim\Lambda}\left\{\mathrm{e}^{-yH_t(x)}\left[1-y\alpha_{t+1}h_{t+1}(x)+\frac{1}{2}\alpha_{t+1}^2\right]\right\}\end{aligned} \tag{12}$$

因此，$t+1$ 次迭代的最优基学习器为：

$$\begin{aligned}h_{t+1}^{*}(x)&=\arg\min_{h_{t+1}}l_{\exp}(H_{t+1}\mid\Lambda)\\&=\arg\min_{h_{t+1}}E_{x\sim\Lambda}\left\{\mathrm{e}^{-yH_t(x)}\left[1-y\alpha_{t+1}h_{t+1}(x)+\frac{1}{2}\alpha_{t+1}^2\right]\right\}\\&=\arg\max_{h_{t+1}}E_{x\sim\Lambda}\left[\mathrm{e}^{-yH_t(x)}y\alpha_{t+1}h_{t+1}(x)\right]\\&=\arg\max_{h_{t+1}}E_{x\sim\Lambda}\left\{\frac{\mathrm{e}^{-yH_t(x)}}{E_{x\sim\Lambda}\left[\mathrm{e}^{-yH_t(x)}\right]}y\alpha_{t+1}h_{t+1}(x)\right\}\end{aligned} \tag{13}$$

其中， $E_{x\sim\Lambda}\left[\mathrm{e}^{-yH_t(x)}\right]$ 是一个常数。令：

$$\Lambda_{t+1}(x)=\frac{\Lambda(x)\mathrm{e}^{-yH_t(x)}}{E_{x\sim\Lambda}\left[\mathrm{e}^{-yH_t(x)}\right]} \tag{14}$$

$\Lambda_{t+1}(x)$ 定义了一种新分布，故：

$$h_{t+1}^{*}(x)=\arg\max_{h_{t+1}} E_{x\sim\Lambda}\left[y\alpha_{t+1}h_{t+1}(x)\right] \tag{15}$$

则有：

$$h_{t+1}^{*}(x)=\arg\max_{h_{t+1}} E_{x\sim\Lambda+1}\left\{\left[1-2I\left(h_{t+1}(x)\neq y\right)\right]\alpha_{t+1}\right\}=\arg\min_{h_{t+1}} E_{x\sim\Lambda+1}\left[I\left(h_{t+1}(x)\neq y\right)\right] \tag{16}$$

所以，h_{t+1} 是在分布 Λ_{t+1} 下以最小化分类误差为优化目标而训练得到的基学习器。而分布 Λ_{t+1} 可通过如下递推公式计算获得：

$$\begin{aligned}\Lambda_{t+1}(x)&=\frac{\Lambda(x)\mathrm{e}^{-yH_t(x)}}{E_{x\sim\Lambda}\left[\mathrm{e}^{-yH_t(x)}\right]}=\frac{\Lambda(x)\mathrm{e}^{-yH_t(x)}\mathrm{e}^{-y\alpha_tH_t(x)}}{E_{x\sim\Lambda}\left[\mathrm{e}^{-yH_t(x)}\right]}\\&=\Lambda_{t+1}(x)\mathrm{e}^{-y\alpha_tH_t(x)}\frac{E_{x\sim\Lambda}\left[\mathrm{e}^{-yH_{t-1}(x)}\right]}{E_{x\sim\Lambda}\left[\mathrm{e}^{-yH_t(x)}\right]}\end{aligned} \tag{17}$$

3.2. 模型求解

将表单 2 中的空缺化学成分进行填零处理，将玻璃类型为高钾和铅钡两种类型。将表单 2 中的数据带入 Matlab 工具箱中的 RUSBoost 模型中，通过混淆矩阵得知，训练集和测试集的平均准确度可以达到 100%

由图 4 中的 ROC 曲线可知，AUC = 1.00 > 0.95，说明分类预测结果十分可靠，可以进行下一步求解。

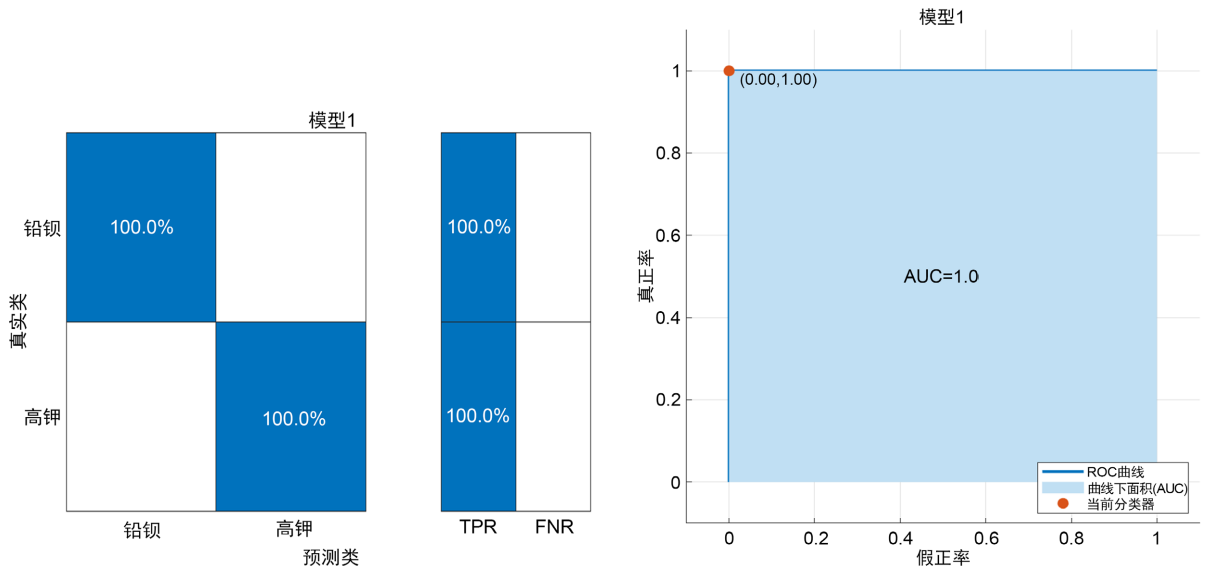

Figure 4. Confusion matrix and ROC curve
图 4. 混淆矩阵和 ROC 曲线图

表 3 中各未知玻璃类型的文物预测结果如下表 5：

Table 5. Prediction result table
表 5. 预测结果表

文物编号	A1	A2	A3	A4	A5	A6	A7	A8
预测结果	高钾	铅钡	铅钡	铅钡	铅钡	高钾	高钾	铅钡

4. 结论

本文根据已有数据建立小样本数学模型，能够准确的预测出玻璃文物风化前后的化学成分变化，并且能够对玻璃文物类别进行亚类划分以及鉴别。在预测风化成分方面，相较于传统方法，采用先分类后预测的方法，能够大幅提高预测准确性。在亚类划分方面，采用 K-Means 和 BP 神经网络结合的方法，能够在兼顾分类多样性和准确性下进行亚类划分。在鉴别玻璃类型方面，对小样本数据采用 RUBoost 机械学习方法，可以准确的鉴别玻璃类型。综上，本篇论文通过建立数学模型对鉴别古代玻璃类型，研究有无风化玻璃制品成分的变化规律，以及高钾、铅钡两种玻璃类型的化学成分统计规律和探索亚类方法方面做出了合理解答。

参考文献

[1] 韩存鸽, 刘长勇. 一种改进的 K-Means 算法[J]. 闽江学院学报, 2019, 40(5): 1-5.

[2] 王菲菲, 李秦, 张梦佳. k-means 聚类算法的改进研究[J]. 甘肃科技纵横, 2017, 46(3): 68-70+83.

[3] 尹化荣, 陈莉, 张永新, 等. 基于 RUSBoost 和积矩系数的神经网络优化算法[J]. 计算机应用研究, 2018, 35(9): 2592-2596.

[4] 尹絮童. RUSBoost 算法在不平衡数据集上的应用[D]: [硕士学位论文]. 大连: 大连理工大学, 2018.