基于 K-Means 与支持向量机的玻璃成分分析与鉴别

高　鑫
重庆师范大学数学科学学院　重庆　401331

摘　要：古代玻璃的化学成分比例在其风化前后会发生较大变化，从而增大其鉴别难度。本文基于公开数据，建立 K-means 聚类、SVM 分类模型鉴别玻璃制品的类别，并采用统计分析、相关性分析等方法总结类别间的关联与规律。首先，基于卡方检验、相关性分析、方差分析探究玻璃特征的统计规律；其次，建立支持向量机模型，对未知类别的玻璃文物进行分类预测并分析模型的准确性；最后，基于 K-Means 对玻璃进行合理指标的选择，选取合适的指标进行亚类划分。最终得到玻璃表面风化与玻璃类型存在显著性差异，与纹饰、颜色不存在显著性差异；铅钡玻璃中氧化铅和氧化钡的含量占比明显高于高钾玻璃；高钾玻璃中氧化钾的含量明显高于铅钡玻璃。

关键词：玻璃成分鉴别；K-means 聚类；支持向量机；敏感性分析

Analysis and Identification of Glass Composition Based on K-Means and Support Vector Machine

Xin Gao
School of Mathematical Sciences, Chongqing Normal University, Chongqing, 401331

Abstract: The chemical composition ratios of ancient glass undergo significant changes before and after weathering, making it more difficult to identify. This paper utilizes publicly available data to establish K-means clustering and SVM classification models to identify categories of glass products. Statistical analysis, correlation analysis, and other methods are employed to summarize the associations and patterns between categories. Firstly, statistical regularities of glass characteristics are explored based on chi-square tests, correlation analysis, and variance analysis. Secondly, a support vector machine model is constructed to classify and predict unknown categories of glass artifacts, and the accuracy of the model is analyzed. Lastly, K-means is used to select reasonable indicators for glass and determine suitable indicators for subcategory divisions. The results reveal significant differences between the surface weathering of glass and its types, while no significant differences exist in terms of patterns and colors. The content ratios of lead oxide and barium oxide in lead-barium glass are significantly higher than those in high-potassium glass, whereas the content of potassium oxide in high-potassium glass is significantly higher than that in lead-barium glass.

Keywords: Identification of Glass Components; K-means Clustering; Support Vector Machine; Sensitivity Analysis

引言

石英砂是玻璃制作的主要原料之一，由于纯石英砂的熔点较高，需加入助熔剂和稳定剂以达到降低石英熔化温度和减慢反应的效果。随着玻璃制作工艺的精进，产生了许多不同类型的玻璃，如铅钡玻璃、钾玻璃。古代玻璃在长期埋藏过程中，极易受埋藏环境的影响而风化。风化的过程中，玻璃内部元素容易与环境元素大量交换，使得其原本的化学成分含量产生变化，从而影响其类别鉴别。

本文基于玻璃风化与否的化学成分含量数据，分析不同类别的玻璃文物样品特征间的统计规律与化学成分间的关联与差异，然后根据已有规律建立玻璃类别预测模型，在确定模型准确性后为不同类别选择适合的化学成分并进行亚类划分。

一、基于卡方检验的玻璃文物属性关系分析

基于玻璃表面有无风化、玻璃类型、颜色和纹饰的数据皆为非连续型定类变量，本文通过卡方检验判断某两个变量间有无关系，并进行统计描述，给出分析结果。

将数据的统计结果按照变量：纹饰、类型、颜色、表面风化进行分类，纹饰取值为：A、B、C，类型取值为：高钾、铅钡，颜色取值为：黑、蓝绿、浅蓝、浅绿、深绿、紫，表面风化取值为：风化、无风化。将纹饰与表面风化、类型与表面风化、颜色与表面风化的各种情况分别组合成$m\times n$列联表，构造统计量：

$$\chi^2=\sum_{ij}\frac{(w_{ij}-a_{ij})^2}{a_{ij}}\sim\chi^2(m-1)(n-1) \quad (1)$$

其中w_{ij}表示第i行第j列的值，$a_{ij}=\frac{w_i\cdot w_j}{w}$，现对纹饰类型与文物表面是否风化作出假设：

H_0：不同纹饰类型对文物表面是否风化不存在显著性差异

H_1： 不同纹饰类型对文物表面是否风化存在显著性差异。

根据纹饰类型与表面风化列联表分析，得到卡方检验结果$\chi^2=4.957$，$P=0.084>0.05$，因此拒绝接受假设H_1并认为不同纹饰类型对文物表面是否风化没有显著性差异。

同理，依次对类型与表面风化、颜色与表面风化作出假设，统计得出类型与表面风化、颜色与表面风化两变量之间的列联表，再结合 SPSS 软件进行卡方检验和效应量化分析，可得到如下结论：①不同纹饰、颜色对文物表面是否风化没有显著性差异；②不同玻璃类型对文物表面是否风化存在显著性差异；

二、特征变量间的关联与差异

本文为探究相同类别和不同类别下化学成分间的关联性与差异性，基于玻璃类型以及有无风化将数据分为 4 类，先通过相关性分析模型求解各化学成分间的相关系数，分析其关联性；再通过方差分析求得不同类别化学成分间的差异性。

同种类别化学成分间的关联分析

基于数据本身特性，本文采用 Pearson 线性相关系数描述两个变量间的相关性。通过求解化学成分之间的 Pearson 相关系数，求解各变量间的相关系数。

不同类别化学成分间关联分析的差异性

本文可采用方差分析模型对不同类别的各化学成分占比数据进行关联关系的差异性分析，从方差分析结果可以看出：不同类别的玻璃文物样品对于氧化钠、氧化铜、氧化锡、二氧化硫（共 4 项）不会表现出显著性差异；对于二氧化硅、氧化钾、氧化钙、氧化镁、氧化铝、氧化铁、氧化铅、氧化钡、五氧化二磷、氧化锶（共 10 项）呈现出显著性差异。

故此本文结合数据的统计规律与分类规律，对比不同化学成分的统计数据然后进行分类讨论，发现铅钡玻璃中氧化铅和氧化钡的含量占比明显高于高钾玻璃，高钾玻璃中氧化钾的含量明显高于铅钡玻璃。故将其作为区分玻璃类型的主要特征因素。

分类预测模型的构建与求解

SVM 是二分类机器学习方法，不需要基于成百上千的数据依旧可以表现出较好的泛化能力，达到满意的分类效果。其核心思想是找到一个超平面，使得它能够尽可能多地将两类数据点正确分开，同时使分开的两类数据点距离分类面最远，距离超平面最近的几个训练样本点被称为“支持向量”。

（一）基于拉格朗日乘子法的向量机模型

Step 1：将所有文物分为风化和无风化两大类预测其类型：用$i=1,2,\cdots\cdots,n$分别表示不同编号的文物，第i个文物的第j个化学成分的含量占比为a_{ij}（$j=1$，2，…，14）。$y_i=-1$表示第一类高钾，$y_i=1$表示第二类铅钡。

Step 2：构造文物的训练数据集为：$\{(x_1,\ y_1),\ (x_2,\ y_2),\ \cdots,\ (x_n,\ y_n)\}$，$x_i$表示第$i$个文物的特征向量，即$x_i=[a_{i1},a_{i2},\cdots,a_{i14}]$。

Step 3：将数据带入$y_i(\omega\cdot x_i+\mathrm{b})-1\geq 0$验证是否线性可分，当满足上述不等式时，说明可以找到一个线性的分类超平面，使得两类样本完全分开，即是$\omega^T x+b=0$，ω为超平面法向量，x为样本输入数据，b为偏置项。

Step 4：下面要找到给定训练集中样本的ω和b的最优解，若不满足不等式$y_i(\omega\cdot x_i+\mathrm{b})-1\geq 0$，即是线性不可分的情况，这个时候，需要在条件中加入松驰因子ξ_i，使其变为$y_i(\omega\cdot x_i+\mathrm{b})-1+\xi_i\geq 0$。为了权衡样本的拟合和测试样本的预测能力，可以引入一个惩罚因子c，得到

$$\begin{cases}\min\dfrac{1}{2}\|\boldsymbol{\omega}\|^2+\mathrm{c}\sum\limits_{\mathrm{i}=1}^{\mathrm{n}}\xi_i\\ y_i(\omega\cdot x_i+\mathrm{b})-1\geq 0\\ i=1,2,\cdots\cdots,\mathrm{n}\end{cases}\tag{1}$$

另外，为了将不可分问题转化为可分问题，本文引入一个非线性映射函数$\boldsymbol{\Phi}$，把文物样本的特征向量映射到某个高维空间后，在这个高维空间中对转化后的线性可分问题求解。此时，用核函数$K(x_i,\ x)$代替原来空间里的内积函数。

Step 5：寻找最优超平面的过程就是求解下列凸二次规划问题：

$$\begin{cases} \min \frac{1}{2}\|\boldsymbol{\omega}\|^2 + \mathrm{c}\sum_{\mathrm{i}=1}^{\mathrm{n}} \xi_i \\ y_i[\boldsymbol{\omega}^T \Phi(x_i) + b] \geq -\xi_{\mathrm{i}} \\ i = 1, 2, \cdots\cdots, \mathrm{n} \end{cases} \tag{2}$$

$\frac{1}{2}\|\boldsymbol{\omega}\|^2$为复杂度，其值越大，表示模型越复杂。

（二）向量机模型的求解及敏感性分析

本文运用 matlab 对数据进行 SVM 训练和预测。首先，将数据划分成风化和无风化两类，观察发现，分类后的风化类型数据存在样本不均衡的问题。对此，在训练之前本文采用过采样（随机）方法对该类型数据进行预处理。由结果可知，该预测模型测试集预测结果准确率、训练集测试结果准确率均在 96%以上，模型预测效果良好，可以进行预测，将风化和无风化数据分别代入训练好的模型进行类型预测。

为了进一步检验模型的适用性，对数据进行扰动处理，将变量值在范围(-120%,+120%)中随机变化，再代入模型进行类型预测，发现尽管变量值在范围(-120%,+120%)内扰动，扰动后的分类准确率仍然在 98%以上，说明该模型合适，稳定性较强，具有良好的普适性。

亚分类模型的构建与求解

K-means 聚类是著名的划分聚类算法，K-means 聚类算法基本过程是先生成 K 个初始聚类中心，并把样本集中的样本按照最小距离原则分配到最近的聚类核心中去。然后计算每个聚类中的样本均值作为新的聚类中心。重复该过程直到聚类中心不再发生变化。

接着针对不同类别进行差异性分析，包括均值±标准差的结果、F 检验结果、显著性 P 值，针对已风化的铅钡玻璃，其差异性分析结果可以较为明显的看出二氧化硫、氧化钡、氧化铅、氧化铜的显著性 P 值小于 0.05，该化学成分既是可以作为亚分类的依据。同理，其余 3 种玻璃亚分类过程相似。

因此，在某一玻璃类别中，若某些化学成分呈现显著性差异，即可选择其作为亚分类的依据，不同类别玻璃文物亚分类划分合适化学成分的选取情况如表 1 所示：

表 1　不同类别玻璃文物亚分类划分依据

类别	聚类个数	合适的化学成分
高钾已风化	2	氧化镁、氧化铝
高钾未风化	2	氧化钾、氧化钙
铅钡已风化	2	二氧化硫、氧化钡、氧化铅、氧化铜
铅钡未风化	2	氧化锶、氧化铅

三、结束语

本文将统计学原理以及机器学习相结合，整体框架较为完整，且在最后的扰动分析中验证了模型具有较好的鲁棒性，准确率高。对玻璃化学成分间的关系分析较为透彻，大量分类、聚类的思路方法可广泛运用于其他分类聚类问题或分析某些指标间的所蕴含的关系和规律，实用性强可行性高。

参考文献：

[1]张丽艳,李洪,陈树彬,李忠镝,阮芪秩,薛天锋,钱敏,凡思军.玻璃的成分和性质的模拟方法[J].硅酸盐学报,2022,50(08).

[2]刘志翔,邓朋飞,潘世濠,罗文迪,张石林,利滔明.防火玻璃的分类及研究现状[J].广州化工,2021,49(15):16-18.

[3]江枫,吴浪,雷杰,张海洋,康泽,王宾,姚颖.硼硅酸盐玻璃成分对硫酸钡热稳定性的影响[J].玻璃,2020,47(02).

[4]韩涛,姚维.基于 K 近邻与支持向量机协同训练的风机叶片结冰早期检测[J].实验室研究与探索,2021,40(09).

[5]凌雪,周弈辰,马健,任萌,库尔班·热合曼,王建新.巴里坤东黑沟遗址出土串饰玻璃珠材质与制作工艺的初步分析[J].硅酸盐通报,2018,37(03).

[6]李沫. 战国秦汉时期费昂斯制品的制备及铅钡玻璃研究[D]. 北京:北京化工大学,2014.