基于支持向量机的玻璃制品成分分析鉴别与研究

邵周健　白群星　张天池　罗　泽

安徽新华学院，中国·安徽　合肥　230088

【摘　要】中国古代玻璃作为其中文化交流重要的载体，针对古代玻璃制品化学成分与其他特征之间关联性的问题，本文结合聚类与决策树算法，建立逻辑回归预测模型，得出文物样品表面有无风化与化学成分的关系及不同类型玻璃的分类规律，并预测其风化前的化学成分含量，最后分析不同类别的玻璃文物样品之间的化学成分关联与差异性.

【关键词】K-means 算法；SVM 算法；决策树算法；逻辑回归预测模型

【基金项目】安徽新华学院2022年校级自然科学研究项目（项目编号：2022zr015）

引言

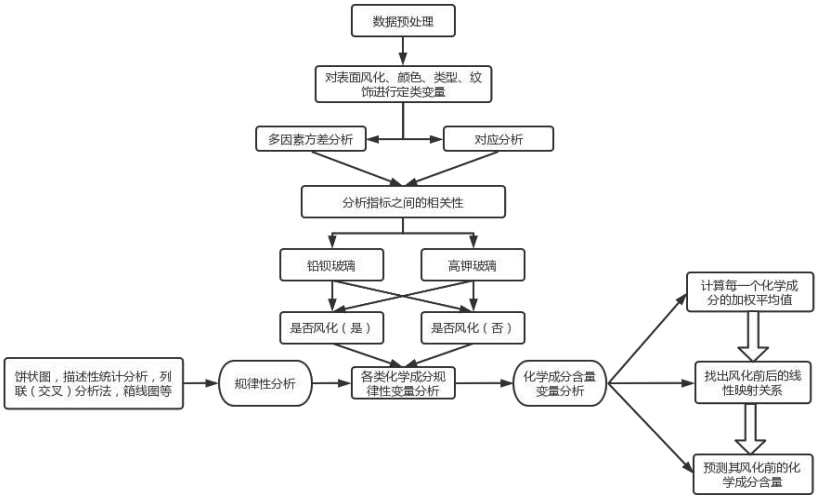
丝绸之路作为古代中国与西方最主要的文化与经济交流通道，一直以来受到人们广泛的关注，中国古代玻璃作为其中文化交流重要的载体，吸引着大量的科研人员对其化学成分进行研究与讨论.本文主要研究玻璃制品成分分析与鉴别关系，对于古代玻璃制品的成分分析与鉴别需要利用收集到的数据 建立适当的数学模型，对玻璃文物表面风化与玻璃类型等其他因素是否存在关系以及分析文物表面的化学成分从而得出统计规律，最终根据风化点来预测风化前化学成分及含量，以及对于不同的类型进行亚类划分，最后对于分类结果进行分析.对不同类别的玻璃文物建立数学模型分析其关联性以及差异性。

1　模型的建立

鉴于掩埋环境对古代玻璃的影响而造成不同程度的风化，并且其成分比例也会发生变化，影响考古的结果[1]。本文首先对古代玻璃表面风化、颜色、类型、纹饰进行分类定义变量，用多因素方差分析和对应分析方法分析指标之间的相关性，对玻璃风化前后化学成分规律变化进行分析，流程分析图如图1所示.本文先进行数据的预处理工作，将每个文物的化学成分含量分别求和将小于85%和大于105%的数据进行剔除，其中缺值对其当做无效数据除去，经过分析发现空值为未能检测到此化学元素，故将其赋值为0，最终得出的预处理结果见附录链接中表1。

为了给出玻璃文物风化与其他因素的关系，以及玻璃类型与化学成分的统计规律.本文将玻璃类型，玻璃纹饰[2]（分成ABC）以及颜色作为风化特征再利用统计软件分析得出玻璃文物表面风化与玻璃类型的关系得到如下结论：（1）B 纹饰非常容易风化， C 较容易被风化，A 的抗风化能力最强；（2）铅钡类型玻璃更容易被风化；（3）玻璃颜色为黑色的更加容易被风化，绿色和深蓝色最不容易被风化，饼状图见附录链接中图2-图5。

图1 流程图



运用皮尔逊相关系数求解，然后根据斯皮尔曼系数来判定某化学成分是否与风化相关性高，从而确定风化化学成

$$r_{x,y}=\frac{\sum_{i=1}^{n}(x_i-\mu_x)(y_i-\mu_y)}{(n-1)\sigma_x\sigma_y} \tag{1}$$

分.总体和样本皮尔逊系数的绝对值小于或等于1.皮尔逊相关系数两个变量的位置和尺度的变化并不会引起该系数的改变,对于样本皮尔逊相关系数:

其中各个参数为.

$$\mu_x=\frac{\sum x}{n},\mu_y=\frac{\sum y}{n},\sigma_x=\sqrt{\frac{\sum_{i=1}^{n}(x_i-\mu_x)^2}{n-1}},\sigma_y=\sqrt{\frac{\sum_{i=1}^{n}(y_i-\mu_y)^2}{n-1}}$$

将相关数据代入上述公式，进行数据的相关性分析，得出结果如附录的表3所示，进行热成像，然后求解风化系数见附录链接中图6、图7。

综合两种方法的结果可以得出：氧化钾（K_2O），氧化钙（CaO），氧化铁（Fe_2O_3），氧化铅（PbO），五氧化二磷（P_2O_5）与风化情况有较大关系.为了对高钾玻璃和铅

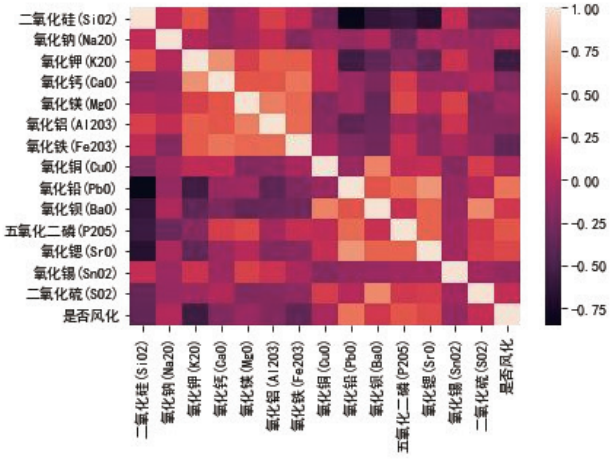


图6热成像分析

钡玻璃的分类规律进行刻画本文建立了分类模型.本文采用K-means 聚类法，运用 SPSS 软件对预处理整理好的数据进行聚类分析.运用K-means 聚类法对风化的古代玻璃文物进行聚类，寻找对高钾玻璃以及铅钡玻璃最合适的分类个数.为了简化类内平均距离和类间平均距离我们采用欧拉距离公式，具体公式如下：

$$J(c,\mu)=\sum_{i=1}^{M}\left\|x_i-\mu_{c_i}\right\|^2 \tag{2}$$

其中 x_i 代表 i 第个样本， c_i 是 x_i 所属的簇， μ_{c_i} 代表簇对应的中心点， M 是样本总数.在得到类别个数后我们开始迭代设置聚类簇数为2 利用如下迭代公式进行类心迭代.

将迭代结果写入Python 中利用pandas 库以sklearn 库进行分析得到结论如下结论：玻璃类型为高钾且纹饰为 B 时文物更容易风化，当纹饰为 A或C 时不容易风化，当玻璃类型为铅钡时纹饰为 A或C 更容易风化.为了验证上述结论是否合理我们参考统计学方法将聚类数据和附件表单中原始数据进行对比得出符合系数为 0.726，证明本文得出的结论是准确且合理的.对亚类划分使用K-means 聚类分析法，利用Python 进行分析求出高钾亚类数量，之后再对数据进行分析得出其他亚类数量，将结果绘制出图 8（见附录链接）。

根据图表分析我们发现函数的拐点在簇族数为3的时候所以得出结论：三个亚类是一组比较合适.之后对上述的数据进行处理与分析得到的分类结果如表4所示（见附录链接）.

用支持向量机，决策树和逻辑回归分别训练模型预测出结果，再使用投票法确定文物的类型，三种模型投票的方式可以在数据较少的情况下减少欠拟合以及噪声的影响.信息熵是度量样本集合纯度最常用的一种指标.假设当前样本集合D中第k类样本所占比例为 $p_k(k=1,2,\cdots,n)$，则D 的信息熵定义为

$$Entropy(D)=-\sum_{k=1}^{n}p_k\log p_k \tag{3}$$

其中， $Entropy(D)$ 越小，D的纯度越高.（信息熵反映的是数据集中的不纯度）

计算信息熵时约定： $Entropy(D)\geq 0$ ， $Entropy(D)$ 的最小值为0，最大值为 $p=0.5$.对于上表中的数据，该数据集中有两个类别，即是高钾或铅钡，所以 $k=2$ ，由此可以计算该数据集的信息熵.

假设特征 a 有 N 个可能取值 $\{a_1,a_2,\cdots a_N\}$ ，若使用特征 a 来对样本集 D 进行数据划分，则会产生 N 个分支结点，其中第n个分支结点包含了样本集 D 中所有在特征 a 上取值 a_n 为样本，记为 D_n .此时，样本集 D 的信息熵就是以特征 a 划分后的 N 个样本信息熵的加权和.得出以下公式：

$$newEntropy(D)=\frac{\sum_{n=1}^{N}Entropy(D_n)}{\sum_{n=1}^{N}Entropy(D)} \tag{4}$$

一般而言，信息增益越大，即以某特征划分样本后信息熵减少了，也就是说纯度提升了.ID3决策树学习算法就是以信息增益为准则来选择划分特征的.本文使用信息增益来计算一下其中一种特征划分后的结果.以此类推，从理论上讲，该特征的信息增益最大，先以它为划分特征最好.通过代码得出各模型预测结果和验证表，如表5及表6（见附录链接）所示。

2 结语

本文对于玻璃文物的化学成分与其他因素关系模型不仅可以用来分析亚类划分以及分类规律还可以推广到医学检验以及成分分析等研究问题中.在医学检验中有很好实践性以及准确性，可以对不同的化学元素提供灵敏的反应.同时在于成分分析领域也有不错的应用价值与推广空间.且通过建立鉴别所属类别模型得到不同类别玻璃文物所属类别以及之间关系.这种根据本文中数学模型得出结论的方法可以推广到现实中很多的问题中，例如：鉴别泥土中土壤成分，分析不同气体成分等.在解决现实问题领域有着很好的推广价值。

参考文献：

[1]刘焕军基于多核支持向量机集成的智能玻璃制品检测算法[J].计算机测量与控制,2011,19(2):4.

[2]千福熹，赵虹霞，李青会等.湖北省出土战国玻璃制品的科技分析与研究[J].江汉考古,2010(2):11.

附录1

http://note.youdao.com/s/CnlUZ2zX

作者简介：

邵周健(1996--)，女，硕士，助教，研究方向：数学。