Universe Scientific Publishing
Modern Chemical Industry, 现代化工 (6)2023,5
ISSN: 2661-3670(Print) 2661-3689(Online)

基于 K-means 聚类与随机森林分类的古代玻璃文物的成分分析与分类鉴别

彭伟坚[1] 崔茂然[2] 朱梅连[3]

1. 广东海洋大学 机械与能源工程学院 广东阳江 529500；
2. 广东海洋大学 计算机科学与工程学院 广东阳江 529500；
3. 广东海洋大学 商学院 广东阳江 529500

摘 要：玻璃经丝绸之路传入我国，我国古代玻璃吸收其技术同时也因地制宜地改变其化学成分。古代玻璃极易受环境影响而风化，通过肉眼辨别玻璃产地较难，但可通过其检测化学成分判断。本文采用 K-均值聚类的方法，对铅钡玻璃与高钾玻璃进行亚类划分，判定各个类别的玻璃因划分为四个亚类，接着建立随机森林分类模型，对古代玻璃类型进行预测，并通过计算 Kappa 值来对随机森林算法敏感性进行评价。

关键词：古代玻璃文物；K-means 聚类；随机森林分类；机器学习

Composition Analysis and Classification of Ancient Glass Relics Based on K-means Clustering and Random Forest Classification

Weijian Peng[1], Maoran Cui[2], Meilian Zhu[3]

1. College of Mechanical and Energy Engineering ,Guangdong Ocean University, Yangjiang Guangdong 529500, China;
2. College of Computer Science and Engineering,Guangdong Ocean University, Yangjiang Guangdong 529500, China;
3.College of Business, Guangdong Ocean University, Yangjiang Guangdong 529500, China

Abstract: Glass was introduced into China through the Silk Road. China's ancient glass absorbed its technology and changed its chemical composition according to local conditions. Ancient glass is very vulnerable to weathering due to environmental impact. It is difficult to identify the origin of glass by the naked eye, but it can be determined by its chemical composition. In this paper, the K-means clustering method is used to classify the lead-barium glass and high-potassium glass into four subclasses, and then the random forest classification model is established to predict the ancient glass types, and the sensitivity of the random forest algorithm is evaluated by calculating the Kappa value

Key words: Ancient glass relics;K-means clustering; random forest classification; machine learning

引言

丝绸之路是古代中西方文化交流的通道，而玻璃——作为早期贸易往来的宝贵商品，更是见证了早期全球经济文化技术的交流。早期的玻璃是在西亚和埃及地区传入我国，匠人们吸收其技术后在本土就地取材制作。为此我国古代玻璃制品虽与外来的外观相似，但化学成分却大不相同[1]。

1 数据准备

1.1 数据来源

分别对铅钡玻璃和高钾玻璃不同的部位进行采样分析，得到各个采样点化学成分含量，并分析其表面有无风化得到本文使用的数据

1.2 原始数据预处理

在数据建模前，一般都需要对给予的附件数据进行处理与分析，即数据预处理。而

数据预处理的好坏很大程度上就决定了后面结果的好坏。

表 1：异常数据

文物采样点	各成分比例累加和
15	79.47%
17	71.89%

2 玻璃类别亚类划分-基于 K-means 算法

2.1 铅钡玻璃

K-均值算法：首先，根据对数据分析与理解，确定簇个数为 4（即计划将数据划分为 4 个类别）。其次，再随机确定 4 个在数据边界范围之内的初始点作为质心。

而对于其余化学成分变量数据与 k 个质心的距离[3]，我们采用欧氏距离公式来计算：

$$d(\mathrm{x},\mathrm{C}_i)=\sqrt{\sum_{j=1}^{m}(x_i-C_{ij})^2} \quad (5)$$

其中，x 为化学成分变量，Ci 为第 i 个聚类中心，m 为变量的维度，xj，Cij 为 x 和 Ci 的第 j 个属性值。

计算得到距离后，应选择最小距离的质心[4]，并将其分配给该质心所对应的簇，直到所有数据全都分配给 4 个簇，同时计算每个簇中的数据对象的平均值，使其作为新的聚类中心。

以此为一个循环，进行下一次迭代，重新分配每个数据到新的质心，直到所有数据的分配结果不再发生改变为止[5]。

通过 K-means 聚类可以得到：仅有氧化钾(K2O)、五氧化二磷(P2O5)、氧化锡(SnO2)三个变量的四个类别的平均值 ± 标准差的显著性 P 值大于 0.05，水平上不呈现显著性，不能拒绝原假设，说明三个变量在聚类分析划分的类别之间不存在显著性差异，分类无效。而其余的变量显著性 P 值都小于 0.05，水平上呈现显著性，变量在聚类分析划分的类别之间存在显著性差异，分类有效。

综上所述，实际设置为 4 个类别时，分类效果较好。分类结果如下图所示

表 2：铅钡玻璃分类结果

类别	文物采样点
1（n=10）	11、23 未风化点、28 未风化点、29 未风化点、42 未风化点 1、42 未风化点 2、44 未风化点、48、49 未风化点、53 未风化点
2（n=16）	02、19、25 未风化点、34、36、38、41、43 部位 2、49、50、50 未风化点、51 部位 1、52、56、57、58
3（n=4）	08、08 严重风化点、26、26 严重风化点
4（n=6）	03、05、08、17、23、28

2.2 高钾玻璃

K-均值聚类算法：

相同地，将高钾类型文物采样点先分成 4 个类别，以此确定簇个数为 4。然后继续计算出除质心外的其余化学成分变量数据与 4 个质心的欧式距离。

通过对上述分析，通过 K-means 聚类可以得到：由上表可知，仅有氧化铅(PbO)、氧化钡(BaO)、氧化锡(SnO_2)三个变量的四个类别的平均值 ± 标准差的显著性 P 值大于 0.05，水平上不呈现显著性，说明这三个变量在聚类分析划分的类别之间不存在显著性差异，分类无效。而其余的变量显著性 P 值都小于 0.05，水平上呈现显著性，变量在聚类分析划分的类别之间存在显著性差异，分类有效。

表 3：分类结果

类别	文物采样点
1（n=8）	03 部位 1、07、09、10、12、21、22、27
2（n=5）	01、03 部位 2、04、05、18
3（n=3）	13、14、16、
4（n=2）	06 部位 1、06 部位 2

3 玻璃分类预测—基于随机森林算法

3.1 随机森林算法

随机森林算法是一种基于决策树的集成学习算法[6]，且可以解决分类和回归两种问题，可以同时处理分类和数值特征。本题使用随机森林是为了处理分类问题，通过所给的数据，将未知类别的玻璃文物分到高钾或铅钡玻璃上。

随机森林算法的基本原理[7]为：首先，采用 Boot - strap 重抽样技术从原始样本中随机抽取多个数据，构造出多个样本。第二，对每个重抽样样本，采用节点的随机分裂技术，构造多棵决策树，并将数据放到每个决策树中，而每个决策树都会输出一个结果。最后，将多棵决策树组合在一起，就可以通过对决策树的判断结果进行投票，最终得到随机森林的预测结果。

而对于随机森林决策树上的特征（即为化学成分），我们要对其进行权重分析，进而选取出最优的特征，以此来保证提升决策树之间的多样性，提升分类性能[8]。

为了更加直观地看出每个特征在随机森林中的每棵决策树上做了多少贡献，一般使用特征重要性来进行评估——基于基尼指数[9]：

基尼指数计算方法及公式：n代表 n 个化学成分，P_n代表类别 n 样本权重[10]

$$Gini = \sum_{n=1}^{N} P_n(1-P_n) = 1 - \sum_{n=1}^{N} P_n^2 \quad (7)$$

根据上式进行计算分析，得到下图：

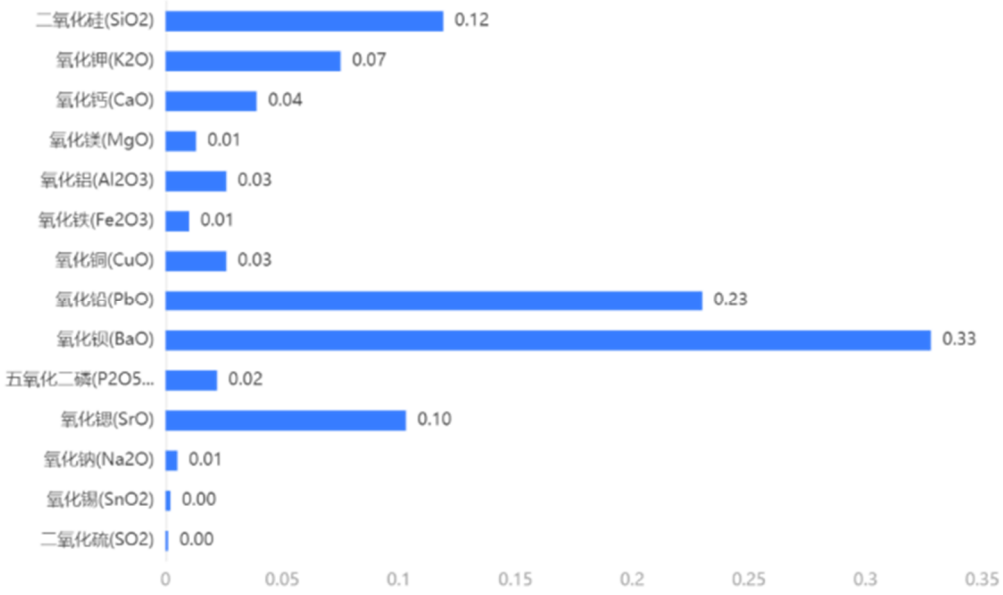

图 1：特征重要性

通过随机森林算法对附件表 3 的预测结果如下表所示：

表 4：分类预测结果

文物编号	A1	A2	A3	A4	A5	A6	A7	A8
类别	高钾	铅钡	铅钡	铅钡	铅钡	高钾	高钾	铅钡

这与我们根据文物化学成分含量来对文物玻璃类别鉴定的结果完全一致，这表明通过随机森林分类预测的结果符合自然规律，可解释性强，可信度高。

3.2 分类结果的敏感性

对于分类结果的敏感度分析，我们选择采用混淆矩阵进行分类的精度评估。混淆矩阵能够说明不同类型地物的分类结果与实际地物类别的相符程度,通过混淆矩阵计算得到的总体分类精度和 Kappa 系数是评价分类结果的重要指标。

而我们主要采用计算 Kappa 系数来评价分类精度的好坏：

$$Kappa = \frac{P\sum_{i=1}^{P} P_{ij} - \sum_{i=1}^{P} P_{i+}P_{+i}}{P^2 - \sum_{i=1}^{r} P_{i+}P_{+i}} \quad (8)$$

其中，P_{ij} 表示矩阵中的元素，i，j=1,2，…，n。

我们使未知玻璃类别的文物的化学成分含量上下浮动 5%，通过计算变化前后的 Kappa 值来判断模型的稳定性，计算结果如下图所示

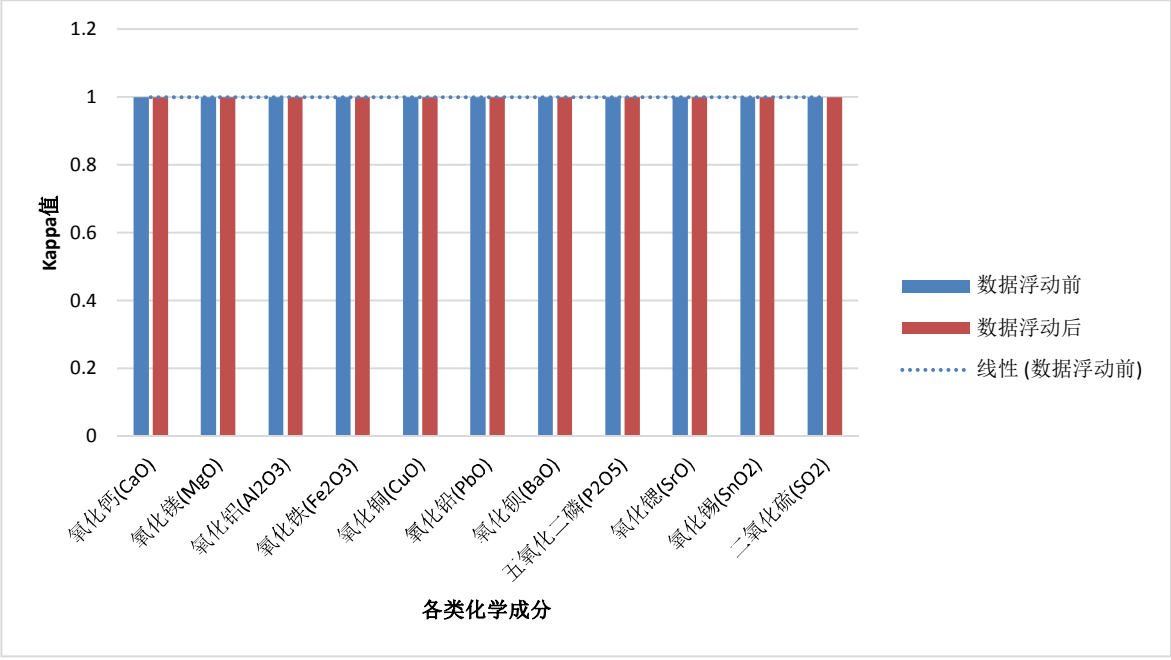

图 2：数据浮动 5%前后随机森林 Kappa 值

由上图可知，通过比对各个化学成分含量浮动 5%前后随机森林的 Kappa 值，我们发现数值前后并无变化，且 Kappa 值等于 1，这表示，模型的分类稳定性很强，敏感性很弱。

参考文献

[1]王承遇，陶瑛. 硅酸盐玻璃的风化[J]. 硅酸盐学报，2003,31(1):8.

[2]陈兴. 蚌埠市 O_3 污染特征及对气象要素的敏感度分析[D]. 南京信息工程大学,2021.DOI:10.27248/d.cnki.gnjqc.2021.001035.

[3]白雨佳,李靖,高升.基于最优 K 均值聚类算法的负荷大数据任务均衡调度研究[J].电力电容器与无功补偿,2022,43(06):85-91. DOI:10.14044/j.1674-1757.pcrpc.2022.06.013.

[4]崔利娜,侯立旺.一种快速稳健的 K 均值聚类算法[J].产业科技创新,2022,4(04):48-49.

[5]赵佳媛,卢中玉,刘鑫宇,徐鹏,孟宇龙.一种基于改进随机森林的知识检索优化算法[J/OL].软件导刊:1-6[2023-02-01].http://kns.cnki.net/kcms/detail/42.1671.TP.20230201.1015.003.html

[6]张瑜,宋建,吴忠庆.机器学习在矿物结构搜索及性质预测方面的应用[J/OL].矿物岩石地球化学通报:1-18[2023-02-01].DOI:10. 19658/j.issn.1007-2802.2023.42.005.

[7]刘悦,马舒畅,杨正伟,邹欣欣,施思齐.面向材料领域机器学习的数据质量治理[J/OL].硅酸盐学报:1-12[2023-02-01].DOI:10.14062/j.issn.0454-5648.20220991.

[8]李菊花,秦顺利,王洁,梁成钢,陈依伟,胡可.随机森林算法在吉木萨尔页岩油藏中的应用[J/OL].长江大学学报(自然科学版):1-8[2023-02-01].DOI:10.16772/j.cnki.1673-1409.20230106.001.

[9]余剑锋,何云良,吴华华,魏骁雄,陈博,钟震远.基于随机森林算法的配电网停电研判方案设计[J].微型电脑应用,2022,38(12):76-79.

[10]李明,褚恬恬.基于贝叶斯优化的随机森林算法在地下空间开发适宜性评价中的应用[J].吉林建筑大学学报,2022,39(06):15-20.

作者简介：

彭伟坚（2002-），男，本科，无职称，从事机器学习，嵌入式开发等。

崔茂然（2002-），男，本科，无职称，从事机器学习等。

朱梅连（2001-），女，本科，无职称，从事经济学研究。