基于随机森林分类算法的古代玻璃种类鉴别

资微君　许淑琪　黄　希
东莞理工学院 广东省东莞市 523000

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化学成分数据进行分析，探究其成分变化规律及成分之间的联系，进而可以对未知类别的古代玻璃进行准确的分类。随机森林分类模型具有准确率极高、可以处理离散型数据和连续型数据的特点，适用于本文数据量较小的特点，因此本文决定使用随机森林算法解决该问题。本文基于随机森林算法，建立分类预测的模型，对未知类别的古代玻璃进行类型的鉴别。最后通过控制变量法改变权重最大的节点特征分析分类结果的敏感性。通过数据结果分析，预测值和真实值高度重合，最终求解真实可靠。同时敏感性分析结果表示分类结果稳定敏感度不高。本文对于玻璃文物的分类具有一定的应用价值，同时该模型充分联系实际，具有很好的通用性和推广性。

关键词：随机森林算法；决策树；SPSS；敏感性分析；古代玻璃

Identification of ancient glass species based on random forest classification algorithm

Weijun Zi, Shuqi Xu, Xi Huang
Dongguan University of Technology　Dongguan Guangdong　523000

Abstract: Ancient glass comes in various types and is susceptible to weathering due to environmental factors. Therefore, it is necessary to analyze the chemical composition data of ancient glass, explore the patterns of compositional changes, and examine the relationships between different components. This analysis can enable accurate classification of unknown types of ancient glass. The random forest classification model is characterized by high accuracy and the ability to handle both discrete and continuous data, making it suitable for the relatively small dataset in this study. Therefore, this paper decides to use the random forest algorithm to address this problem. Based on the random forest algorithm, this paper establishes a classification prediction model to identify the types of unknown ancient glass. Finally, by controlling variables and changing the weights of the most significant node features, the sensitivity of the classification results is analyzed. Through data analysis, the predicted values closely align with the true values, indicating the reliability of the results. At the same time, the sensitivity analysis results demonstrate that the classification results are stable and not highly sensitive. This paper has practical applications in the classification of glass artifacts and the developed model exhibits good generality and scalability.

Keywords: random forest algorithm; Decision tree; SPSS；　sensitivity analysis; Ancient glass

引言

早期的玻璃通过丝绸之路被制成饰品传入我国，我国古代玻璃充分学习其技术后在本土就地取材制作玻璃，因此我国古代玻璃在外观上与外来玻璃相似，但其化学成分却大不相同。在炼制玻璃时由于其主要成分石英砂熔点很高，通常添加助熔剂来降低熔化温度同时添加石灰石作为稳定剂，助熔剂的不同会导致玻璃的化学成分不同。

古代玻璃容易受其埋藏的地理环境的影响而风化。风化时，由于内部元素与环境元素发生了交换，造成古代玻璃成分比例发生较大的变化，从而影响对其的类别判断。

现有考古工作者整理了一批我国古代玻璃制品的化学成分及类型的相关数据。根据上述数据本文提出了一种基于随机森林分类算法的古代玻璃种类鉴别方法以实现对未知类别的古代玻璃文物所属类型的检测。本文主要基于随机森林算法，建立分类预测的模型，通过已知分类的玻璃数据集，将其风化情况以及 14 种玻璃的化学成分作为特征值，随机选取百分之七十的数据作为训练集训练随机森林模型，选取剩下的百分之三十的数据作为测试集对模型进行检验，再将未知分类的数据代入模型，即可得到其各个采样点的玻璃类型，最后通过控制变量法改变权重最大的节点特征分析分类结果的敏感性。

一、基于随机森林分类算法[1]的古代玻璃种类鉴别方法

1.1 随机森林分类算法简介[2]

本文所提出的随机森林算法是基于决策树思想的一种集成机器学习算法。随机森林器通过每棵决策树生成的训练集进行训练 ，每棵树依赖于一个由训练确定的参数所组成的随机向量，森林中生成的数的参数随机向量也是

独立同分布的；分类时随机森林输出每棵树结果的组合。

分裂属性个数(mtry)和决策树的棵树（ntree)是随机森林算法最重要的两个参数,其中在随机森林算法中随机选取 mtry 个不同的输入属性，且算法仅在此范围内选取最有效的结点分裂属性。分裂属性集中属性个数 mtry 是对随机森林分类效果影响较为敏感的参数，可以自行设置它的值。

1.2 建模思路

利用随机森林分类算法，建立数学模型，通过 SPSS PRO 软件，求解得到各化学成分之间的关系，再通过该规律鉴别附件 2 中的文物类型。最后，通过控制变量法对分类结果的敏感性进行分析。

1.3 本文研究所用的实验数据

1）已知分类的玻璃数据集：铅钡玻璃分为 3 类，分别是铅钡玻璃 1、铅钡玻璃 2、铅钡玻璃 3，而高钾玻璃只有一类。

附件 1：58 个文物不同部位的化学成分（如表 1 所示）；附件 2：8 个已知化学成分和风化情况，但未知玻璃类型的文物（如表 2 所示）

表 1 附件 1 中部分数据

文物采样点	二氧化硅	氧化钠	氧化钾	氧化钙	氧化镁	氧化铝	氧化铁
01	69.33	——	9.99	6.32	0.87	3.93	1.74
02	36.28	——	1.05	2.34	1.18	5.73	1.86
03 部位 1	87.05	——	5.19	2.01	——	4.06	——
03 部位 2	61.71	——	12.37	5.87	1.11	5.5	2.16

表 2 附件 2 中部分数据

文物编号	表面风化	二氧化硅	氧化钠	氧化钾	氧化钙	氧化镁	氧化铝
A1	无风化	78.45	——	——	6.08	1.86	7.23
A2	风化	37.75	——	——	7.63	——	2.33
A3	无风化	31.95	——	1.36	7.19	0.81	2.93
A4	无风化	35.47	——	0.79	2.89	1.05	7.07

3）由于各成分比例的累加和介于 85%~105%之间的数据为有效数据。通过 EXCEL 软件计算附件 1 中各文物样本的成分比例累加和。结果显示，57 个文物样本采集点中，有 55 个样本成分比例累加和介于 85%~105%之间为有效数据，剩余两个样本不在范围内视为无效数据，分别是样本 15 和 17，剔除这两个样本，不予考虑。

1.4 模型求解

下面是随机森林算法[3]流程的步骤：（如图 1 所示）

1 利用 Booktstrap 法重新采集原本数据样本集 X，随机生成 K 个训练样本集 X1*,X2*,. . . ,XK*；

2 通过每个生成的训练集，生成对应的决策树 T1,T2,. . . ,Tk,在每个中间节点上选择 mtry 个属性中最佳分裂方式的属性作为当前节点的分裂属性在此节点上进行分裂；

3 每棵决策树都完整生长并不对其进行剪枝；

4 将每棵决策树对原始数据样本集 X 进行测试分类；

5 通过投票的方式，将 K 棵决策树输出最多的类别作为原始数据样本集 X 的所属类别。

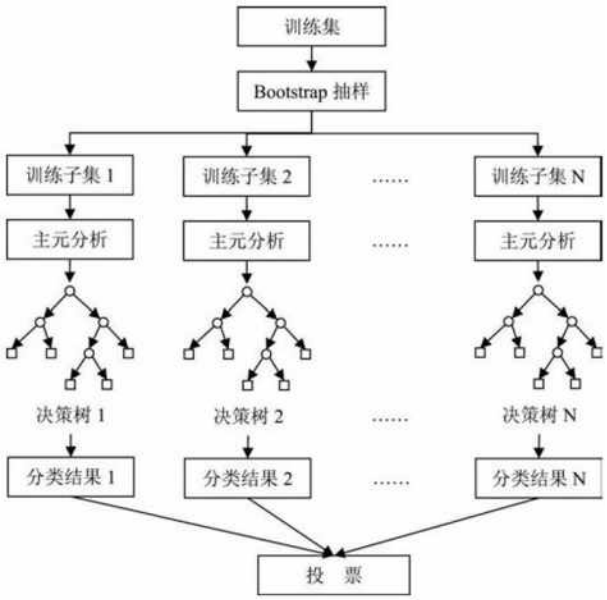

图 1 随机森林算法流程的步骤

模型建立的核心是森林中决策树的构建，每棵决策树都任其最大限度地生长，不对其进行剪枝。随机森林分类算法中不同的决策个数对模型泛化性能也存在一定的影响。决策个数的多少直接影响随机森林分类算法的运算速度和分类效果。因此，本文决定选择 50 棵决策树，防止随机森林算法的速度下降和提高模型的分类准确率。随机森林算法的优势在于它算法简洁，训练速度快，效果稳定，网络只需要训练一次就可以获得理想结果。因此，应该提前确定和设置合理的决策树棵树。取玻璃分类准确率的平

均值作为当前决策树棵树下随机森林分类的准确率。某次运行的结果如下图 2 所示。

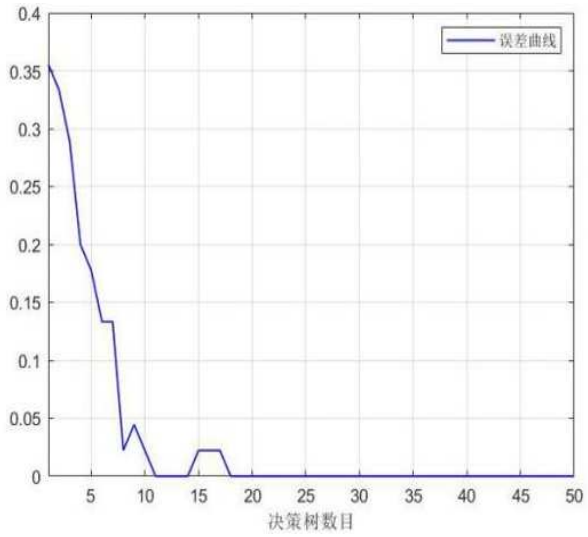

图 2 决策树运行结果

从图中可以看出，随着决策树棵树的增加，随机森林的分类误差率逐渐稳定，综合考虑随机森林中包含的决策树的棵树与建模的速度，选取随机森林分类算法中包含 50 棵决策树是比较理想的。

二、结果分析

2.1 未知类别玻璃鉴别结果

本问将附件 1 中的所有数据随机分成两部分，分别占 70%和 30%。70%的数据用于随机森林算法的学习训练得出一套规律，30%的数据用于随机森林算法检验其前面得出的模型是否正确。通过数据结果分析，训练集和测试机正确率都是 100%，预测值和真实值高度重合，证明其随机算法得到的规律正确，可用于鉴别附件 2 中未知类别的玻璃文物所属类型。结果如表 3 所示。

表 3 未知类别玻璃的类型

样品编号	A1	A2	A3	A4
类型	高钾玻璃	铅钡玻璃	铅钡玻璃	铅钡玻璃
样品编号	A5	A6	A7	A8
类型	铅钡玻璃	高钾玻璃	高钾玻璃	铅钡玻璃

2.2 敏感性分析[4]

敏感性分析是研究与分析一个系统或模型的状态或输出变化，对系统参数或周围条件变化的敏感程度的方法。

在高钾玻璃和铅钡玻璃中二氧化硅，氧化铅，氧化钾的含量相对较高，影响程度比其它化学成分大，具有代表性。因此本问将二氧化硅，氧化铅，氧化钾等三个指标作为变量，通过控制变量法，分析其对结果的敏感性。

将三个指标的含量分别增大为原来的 1.1 倍，1.1 倍，1.5 倍，将改变后的数据代入随机森林分类算法的分类预测模型中，利用 SPSS PRO 软件求解。

通过下面三张预测结果表发现，当上述三个指标的含量发生变化后，预测的结果没有任何改变，只有预测结果概率发生了略微的变化，说明用该模型得到的结果是稳定的，敏感性不高。预测结果见表 4，5，6。

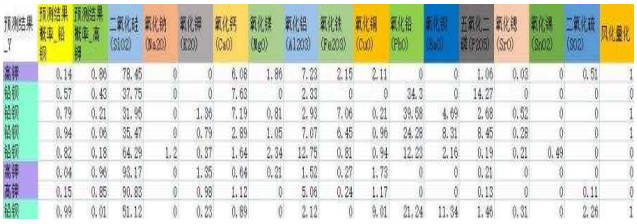

表 4 预测结果(1.1 倍二氧化硅)

表 5 预测结果(1.1 倍氧化铅)

表 6 预测结果(1.5 倍氧化钾)

三、结束语

本文针对鉴别未知类别玻类型的问题，提出了一种基于随机森林分类算法的古代玻璃种类鉴别的方法。本文基于随机森林算法，建立分类预测的模型，最终成功预测出 6 种未知类别的玻璃类型。通过数据结果分析，训练集和测试机正确率都是 100%，预测值和真实值高度重合，证明其随机算法得到的规律正确。

敏感性分析结果表明：当上述三个指标的含量发生变化后，预测的结果没有任何改变，只有预测结果概率发生

了略微的变化，说明用该模型得到的结果是稳定的。

虽然随机森林分类模型得到的结果正确率极高且模型稳定，但是本文工作仍存在以下不足：随机森林分类模型适用于数据量较小的问题中，在数据量较大时可能无法满足任务的实时性要求，故在未来的研究中可以通过增大数据集加强模型训练，减少误差产生，增大模型实用性。

参考文献：

[1]白秀显.基于决策树算法的高校招生数据挖掘与可视化系统的设计与实现[D].兰州大学,2019.

[2]马玉彬,刘仕友,曹丹平等. 基于随机森林算法的智能岩性识别方法[C]//.2022 年中国石油物探学术年会论文集（下 ）.[出版者不详],2022:530-532.DOI:10.26914/c.cnkihy.2022.039595.

[3]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.

[4]季昀,段杭,孟亚运等.基于随机响应面法的结构可靠度敏感性分析理论及其工程应用[J/OL].水力发电:1-8[2023-01-30].http://kns.cnki.net/kcms/detail/11.1845.TV.20221222.1144.001.html

[5]冯百龄. 中国出土古代玻璃珠数据库建设与应用[D].西北大学,2021.DOI:10.27405/d.cnki.gxbdu.2021.001676.