基于梯度提升决策树和多元回归的古代玻璃制品分类与成分分析模型

摘要

本文针对古代玻璃文物风化状态检测和文物分类问题，以WIP和PbO/SiO_2为重要指标研究与风化状态的统计关系，并建立岭回归化学成分预测模型，之后基于决策树和聚类分析建立分类模型，最后建立多元线性回归模型对化学成分关联性进行分析。

针对问题一，首先基于*Spearman*相关系数和卡方检测对玻璃理化性质和风化状态进行相关性和差异性分析，发现玻璃类型与风化状态存在高度相关关系，得到高钾类更有可能不被风化，铅钡类更有可能被风化的结论。由于化学成分较多，选取K_2O和BaO等重要指标，并引入风化指数WIP和 PbO/SiO_2作为风化状态判断重要依据，带入**支持向量机**中，得到基于指标 **WIP** 的高钾玻璃风化统计模型和基于PbO/SiO_2和BaO的铅钡玻璃风化统计模型，准确率达到**85.7%**，同时从横向纵向两个角度对玻璃风化与化学成分之间的统计规律进行分析，得到PbO/SiO_2和风化状态存在极大关系，证明指标选取的合理性。针对定类变量，基于*onehot*编码构建以四个理化特征为自变量，化学成分含量为因变量的**岭回归化学成分预测模型**，并与前文分析得到的风化前后化学成分变化关系进行对比，证明模型的预测结果与前文得到的结论基本符合。

针对问题二，建立**决策树**分类模型，得到以PbO为决策节点的决策树分类结构，得出以$PbO \le 5.46\%$划分的分类规律，并由此证明PbO为重要特征的合理性。采用聚类分析对高钾和铅钡玻璃进行亚类划分，由肘部图确定**高钾玻璃分为 3 个亚类，铅钡玻璃分为 5 个亚类**的结论，再利用决策树模型对高钾和铅钡玻璃分类，得到以SiO_2、P_2O_5为指标的高钾玻璃亚分类模型和以SiO_2、PbO、BaO为指标的铅钡玻璃亚分类模型。并查找资料给出玻璃类型与化学成分关系表，将模型所分亚类的化学成分特征与关系表对比，找到与各亚类对应的玻璃类型，证明选取重要指标的模型分类方式具有合理性。通过，改变聚类数量得到聚类数大于 3 时模型的敏感性较低的结论。最后，对决策树模型的性能进行评价，模型分类准确率为**71.4%**。

针对问题三，为提高分类准确率并解决决策树过拟合的问题，对两类玻璃分别构建基于 **GBDT** 的亚分类预测模型，并在表 11 中给出编号 A1~A8 的文物类型和亚类预测结果。从性能和模型参数两方面对模型进行敏感性分析性能角度通过使得各化学成分含量上下浮动，得到以化学成分变异百分比为自变量，模型预测准确率为因变量的敏感性分析图，得到模型对于 10%以内的扰动具有很好的容错性；参数敏感性角度，通过改变基学习器个数和决策树深度，得到高钾玻璃和铅钡玻璃分别在学习器个数大于 6 和 4，树深大于 3 时敏感性较低。最后模型分类准确率达到**95%**以上。

针对问题四，两类玻璃分别通过相关性分析得到 14 种化学成分之间的相关系数，对每一种化学成分选取与之相关性最大的几个元素，建立**多元线性回归模型**，得到化学成分之间的关联性。对 14 个关联组的拟合优度进行排序，结合相关性选优和拟合优度选优两次筛选，选取其中最大的 5 组，进行不同类之间化学元素关联性的差异性分析。并对不同类间SiO_2和BaO含量的差异性结合机理进行分析，得到铅钡玻璃中SiO_2含量受PbO影响最为显著（$r=-0.76$），并对其影响机理进行分析，高钾玻璃SiO_2含量主要受K_2O影响（$r=-0.86$）,受PbO影响较小（$r=-0.32$），证明了前文得到的成分变化规律的正确性。铅钡玻璃中BaO与CuO相关性较强，而高钾玻璃中相关性较弱。

关键词：机理分析 风化指数 岭回归 GBDT 决策树 聚类分析 多元线性回归