基于分类研判与 K-Means++的玻璃制品的成分分析与鉴别

刘雅旭

廊坊师范学院

【摘要】古代玻璃在风化过程中，由于内部元素与环境元素进行大量交换，导致其成分比例发生变化，从而影响对其类别的正确判断，本文通过对已知数据进行统计性分析进行相关的分类研判并采用无监督学习的 K-means++算法对玻璃制品成分进行分析与鉴别。首先进行数据预处理，根据题目要求，删除编号为 15 和 17 的两条错误数据，将颜色为空值的部分根据风化程度规律设定为黑色处理，将其他空值进行填“0”处理，进行接下来的计算首先针对四个定类变量使用卡方检验进行分析，根据显著性 P 值是否小于 0.05 判断出玻璃类型和颜色与表面风化存在显著性差异，与纹饰不存在显著性差异，在此基础上进行效应量化分析，其中包含 phi、Crammer's V、列联系数与 lambda。

【关键词】玻璃成分鉴别；卡方检验；加权平均预测；系统聚类；K-Means++

【收稿日期】2022 年 11 月 25 日 【出刊日期】2022 年 12 月 28 日 【DOI】10.12208/j.jccr.20220017

Composition analysis and identification of glass products based on classification and K-means + +

Liu Yaxu

Langfang Normal College

【**Abstract**】In the weathering process of ancient glass, due to a large amount of exchange between internal elements and environmental elements, the composition proportion of ancient glass changed, which affected the correct judgment of its category. This paper conducts relevant classification research and judgment through statistical analysis of known data, and uses K-means++algorithm without supervision to analyze and identify the composition of glass products. First, conduct data preprocessing, delete the two wrong data numbered 15 and 17 according to the requirements of the title, set the part with blank color as black according to the weathering degree law, fill in "0" for other blank values, and conduct the next calculation. First, use chi square test to analyze the four categorical variables, According to whether the significance P value is less than 0.05, it can be judged that there is a significant difference between glass type and color and surface weathering, and there is no significant difference between glass type and color and ornamentation. On this basis, quantitative analysis of effects is conducted, including phi, Crammer's V, contingency coefficient and lambda

【**Keywords**】glass composition identification; Chi square test; Weighted average forecast; System clustering; K-Means

1 研究背景

玻璃的主要原料是石英砂，其主要的化学成分是二氧化硅（SiO2）。在玻璃的煅烧过程中所添加的助熔剂不相同，它们的主要化学成分也不相同。例如在铅钡玻璃的烧制过程当中加入了铅矿石作为添加的助熔剂，其中氧化铅（PbO）、氧化钡（BaO）的含量比较高，这通常被认为是我们国家自己所发明的玻璃品种，楚文化中的玻璃就主要是铅钡玻璃。钾玻璃是用含钾量比较高的物质比如草木灰作为添加的助熔剂而烧制成的，其主要流行于我国的岭南和东南亚跟印度等地区[1]。

2 建模与求解

2.1 问题一的建模与求解

首先我们要对玻璃表面的风化情况和玻璃类型，纹饰和颜色的差异性分别进行分析，并且结合玻璃的类型来分析化学成分含量的变化规律且预测风化前

的化学成分的含量，共需要解决三个小问题，问题一的建模分析流程图。

（1）数据的预处理

①首先我们做数据预处理的工作，根据本题的题目要求：将成分比例累加和介于 85%~105%之间的数据视为有效数据，我们根据分析编号 15 和编号 17 的总成分小于了 5%，因此在此后的计算中不需考虑编号 15 跟编号 17 两组错误的数据，我们将其做剔除处理[2]。

②附件表单 1 中的颜色列中的数据之中，我们可以发现有四个空值，通过观察数据的变化情况可以知道其颜色的深浅程度和风化程度呈现正相关的变化，因此我们将这四个空值进行填补，填补为“黑色”。

给出了相应主要成分占的比例，空白处来表示未检测到该成分，而非缺失值，所以我们把未检测到的数据做补“0”处理，来方便接下来的计算。

2.2 针对表面风化情况进行卡方检验

首先我们用 Excel 当中的 VLOOKUP 函数把表单 1 和表单 2 中数据进行合并，以便方便接下来的统计，经过观察分析数据可知纹饰、类型、颜色、表面风化都为定类变量，对多组定类变量间的差异性分析我们可以用卡方检验。卡方检验主要是来比较定类变量跟定类变量间的差异性分析。我们通过统计样本的实际观测值和理论推断值之中的偏离程度，实际观测值跟理论推断值之间的偏离程度就可以决定卡方值大小，假如卡方值越大，则二者的偏差程度就越大；反之，二者的偏差越小；如果两个值完全相等，卡方值为 0，表示理论值完全符合[3]。变量 X：表面风化；变量 Y：纹饰，类型，颜色，用 SPSS 软件做交互分析，得出如下由上表的卡方检验分析结果可得出：表面风化和纹饰，显著性 P 值为 0.057*，接受原假设，所以不存在显著性差异；表面风化和类型，显著性 P 值为 0.003***，拒绝原假设，因此存在显著性差异；表面风化和颜色，显著性 P 值为 0.045**，拒绝原假设，所以存在显著性差异。以此为基础做效应量化分析，包括了 phi、Crammer's V、列联系数、lambda ，作为分析其表面风化跟其余三个指标的相关程度，量化分析指标的解释如下：

①phi 系数：phi 相关系数的大小，表示了两样本间的关联程度。若 phi 系数小于 0.3，表示相关较弱；若其 phi 系数大于 0.6，则表示相关较强。

②Cramer's V：与 phi 系数作用类似，但 Cramer's V 系数作用的范围比较广。

③列联系数：简称为 C 系数。

④lambda：用于反映自变量对于因变量的预测效果。

根据效应量化分析的结果我们可得到：纹饰 Cramer’s V 值为 0.326，因此纹饰跟其表面风化的差异程度是中等程度的差异；同理可得玻璃类型的 Cramer’s V 值是 0.316，其是中等程度差异；颜色的 Cramer’s V 值为 0.341， 其差异程度是中等程度差异。

由 PHI 值我们可分析并得出：纹饰、颜色、玻璃类型值当中纹饰的 PHI 值都小于 0.3，这可以说明其和表面风化的相关性较弱，颜色跟玻璃类型的 PHI 值在 0.3 和 0.6 之间，可看出其相关程度是中等。

根据效应量化分析的结果我们可得到：纹饰 Cramer’s V 值为 0.326，因此纹饰跟其表面风化的差异程度是中等程度的差异；同理可得玻璃类型的 Cramer’s V 值是 0.316，其是中等程度差异；颜色的 Cramer’s V 值为 0.341，其差异程度是中等程度差异[4]。

Q1 1.5* IQR, Q3 1.5* IQR ，则区间外的值即为异常值。

可以直观的看出氧化镁（MgO）和氧化铝（Al2O3）的显著性 P 值小于 0.05，因此选择这两个指标作为合适的变量，同理可知，其余不同类型的玻璃风化程度计算过程相同，高钾类玻璃风化后中，氧化镁和氧化铝呈现显著性差异，因此选择二者进行区分，进行亚类划分，其他分类情况也是如此，最后进行灵敏度分析，对数据进行扰动。使用 Matlab 进行编程，设置随机出范围，将结果重新带回到 K-Means 算法种进行计算与原始聚类中心点坐标的欧氏距离，重新进行判断，与真实值进行比较，计算出模型的准确率，设定灵敏度变化分别为 1%、2%、5%、10%、20%、30%，最终得出当扰动范围为 30%时，模型的准确率为 91%。可以看出高钾类玻璃在风化前后增加随机扰动，扰动值小于 30%的时候，模型的准确率依然为 100%，说明聚类效果良好，模型的鲁棒性较强；针对铅钡类玻璃未风化时，仅当扰动效果为 30%的时候，模型的准确度为 91%，其余情况正确率为 100%，因此在一定程度上反映了我们模型的准确性与稳定性。

问题三需要我们对表单 3 中未知玻璃文物的化

学成分进行分析，并预测其所属的类型，并进行敏感性分析，我们在问题二的基础上再次使用 K-Means 模型，针对指标的差异性选择合适的化学成分指标作为分类距离判断的依据，从而对表单 3 中的数据进行预测问题

3 不同类别之间的化学成分关联关系的差异性判断

在 K-Means++聚类结果的基础上选择一些合适的指标进行分类结果差异性的判断，本文我们选择 CHI、DBI、轮廓系数进行比较分析，三个指标的说明如下：

（1）CHI 指数：CHI 指标是数据集的分离度与紧密度的比值，以各类中心点与数据集的中心点的距离平方和来度量数据集的分离度，以类内各点与其类中心的距离平方和来度量数据的紧密度。聚类效果越好，类间差距应该越大，类内差距越小，即类自身越紧密，类间越分散，CH 指标值越大聚类效果越好[5]。

（2）DBI 指数：Davies-Bouldin 准测基于簇内和簇间距离的比率，最优聚类解决方案具有最小的 Davies-Bouldin 指数值（DBI）（越小越好）

（3）轮廓系数：轮廓系数评价指标说明：轮廓系数的取值在[-1,1]之间，越趋近于 1 代表内聚度和分离度都相对较优，即聚类效果越好。（当簇内只有一点时，我们定义轮廓系数 s（i）为 0）由上表分析可知，高钾类玻璃风化后分类数目增加，铅钡的玻璃风化后分类数目减少。高钾类玻璃 CHI 指数风化后明显增加，铅钡类玻璃风化后相对减少，通过 CHI 指数可以明显的判断出不同玻璃类型分化前后的聚类效果；针对 DBI 指数，高钾类玻璃。DBI 指数增加千倍玻璃风化后减少；针对轮廓系数，高钾玻璃风化后更远离+1，而铅钡玻璃风化后更接近于+1，说明铅钡玻璃风化后化学成分关联关系较强，而高钾玻璃风化后化学成分关联关系相对减弱。在高钾玻璃中，二氧化硅（SiO2）风化前后均属于单独一类， 氧化钙（CaO）和氧化铜（CuO）在风化后共同属于一类，氧化铝（Al2O3）在风化后属于单独一类，其余化学成分的所属类别均无变化；在铅钡玻璃中，二氧化硅（SiO2）、氧化铝（Al2O3）、氧化铅（PbO）和氧化钡（BaO）单独属于一类，其余化学成分属于一类，风化后二氧化硅（SiO2）、氧化铅（PbO）和氧化钡（BaO）属于单独一类，其余化学成分属于一类[6]。

参考文献

[1] 崔汝麦,司守奎 .数据可视化处理的数学模型[J].海军航空工程学院学报,2010,25(01):105-108+112.

[2] 刘利红,陈亚峰.基于加权平均算法和 Middle Plane 网格的保险杠开模收缩率预测方法[J]. 塑料科技,2021,49(07):133-136

[3] 韩存鸽, 刘长勇. 一种改进的 K-Means 算法[J]. 闽江学院学报, 2019, 40(5): 1-5.

[4] 王菲菲, 李秦, 张梦佳. k-means 聚类算法的改进研究[J]. 甘肃科技纵横, 2017, 46(3): 68-70+83.

[5] 尹化荣, 陈莉, 张永新, 等. 基于 RUSBoost 和积矩系数的神经网络优化算法[J]. 计算机应用研究, 2018, 35(9): 2592-2596.

[6] 尹絮童. RUSBoost 算法在不平衡数据集上的应用[D]: [硕士学位论文]. 大连: 大连理工大学, 2018.