基于分类算法的古代玻璃制品成分分析与鉴别模型*

周祖灏　李文鑫　黄培峰　林宁**

南宁学院信息工程学院，南宁 530200

摘　要　为帮助考古工作者能够更好地分析和鉴别古代玻璃制品的成分，本文使用 Spearman 相关系数、卡方检验、岭回归方法、肘部法等一系列的方法和聚类分析 K-Means 算法、决策树-遗传算法、LGBM 分类算法等分类算法，建立了决策树机器分类模型、亚分类决策树模型，充分利用了 Spsspro 软件实现对古代玻璃制品成分的分析问题进行分析求解。

关键字　Spearman 相关系数，卡方检验，聚类分析，决策树，MannWhitney 检验

Composition Analysis and Identification Model of Ancient Glass Products Based on Classification Algorithm

Zhou Zuhao　Li Wenxin　Huang Peifeng　Lin Ning

School of Information Engineering
Nanning University

Nanning 530200 China

Abstract—In order to assist archaeologists to better analyze and identify the composition of ancient glass products, this paper uses a series of methods such as Spearman correlation coefficient, Chi-squared test, ridge regression method, elbow method and clustering analysis K-Means algorithm, decision tree genetic algorithm, LGBM classification algorithm and other classification algorithms to establish a decision tree machine classification model and a sub classification decision Tree model. Then, it utilized Spsspro software to fully analyze and solve the composition of ancient glass products.

Key words—Spearman correlation coefficient，Chi-square test，Cluster analysis，Decision tree，MannWhitney test

1　引　言

玻璃的主要原料是石英砂，主要化学成分是二氧化硅。古代在炼制时需要添加助熔剂，常用的助熔剂有草木灰、天然泡碱、硝石和铅矿石等，并添加石灰石作为稳定剂，石灰石煅烧以后转化为氧化钙（CaO）。添加的助熔剂不同，其主要化学成分也不同。古代玻璃极易受埋藏环境的影响而风化，在风化过程中，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。根据现有的我国古代玻璃制品的相关数据、化学成分和检测手段，考古工作者将文物样品分为高钾玻璃和铅钡玻璃两类。

对于玻璃成分的分析与分类问题，本文做了下面几个假设：

（1） 假设玻璃风化与未风化的玻璃类型以及它们的化学元素含量的成分变化时不考虑不同环境和风化的时间长短的影响。

（2） 假设在对不同类型的玻璃进行亚分类讨论的时候不考虑玻璃的颜色与纹路之间对分类结果所造成的影响。

（3） 假设在选择指标时，就仅考虑显著性 p 值造成的影响，不考虑其它因素造成的影响。

本文根据一些已有的数据和分类方法，分析求解以下问题：

问题一：玻璃文物的表面风化与玻璃类型、纹饰和颜色存在什么关系；然后再根据玻璃的类型分析出文物样品表面有无风化化学成分含量的统计规律，预测出风化的化学成分含量。

问题二：分析高钾和铅钡玻璃的分类规律，再对每个分类选择合适的化学成分进行亚分类。

问题三：分析未知类别的玻璃的化学成分去鉴定其类型。

问题四：对不同类别的玻璃样品分析他们化学成分之间的关系，比较不同类别之间化学成分关联关系的差异性。

*基金资助：基金资助：本文得到南宁学院 2019 年度教授培育工程项目（2019JSGC12）资助。
**通讯作者：林宁，副教授，bgy_2009@163.com

2 模型的建立与求解

2.1 问题一的模型汇总与求解

因为 Spearman 方法可以同时对多个数据进行相关性分析，所以第一步先采取这个方法去对它们四者的关系大概分析一下，相关系数越接近 1 或-1 的相关性越大[1]。

表 1 相关系数热力

颜色	0.541	-0.112	-0.481	1
纹饰	-0.37	0.128	1	-0.481
表面风化	0.316	1	0.128	-0.112
类型	1	0.316	-0.37	0.541
	类型	表面风化	纹饰	颜色

由表 1 可以看出类型与颜色呈现强正相关性，且显著；表面风化与纹饰呈现弱正相关性，且不显著，同理可得出类型与纹饰、类型与表面风化等的相关性和显著性水平。卡方检验可以统计样本的实际观测值与理论推断值之间的关系，可以得出风化与未风化对于玻璃文物的纹路，颜色与材料之间的相关性，观测值与理论值之间的偏离程度可以决定卡方检验值的大小，如果卡方检验值越大则它们的偏离程度越大反之越小，如果它们相等时就表明理论值完全符合[2]，我们先确定唯一变量风化变化，对其他三者进行卡方检验，然后将数据继续在 spsspro 选择则差异性分析里面的卡方检验进行分析得出结果如表 2 所示。

表 2 表面风化对纹饰、类型、颜色的显著性的卡方检验分析结果

题目	名称	表面风化		总计	X^2	校正 X^2	P
		无风化	风化				
纹饰	C	13	15	26	5.747	5.747	0.056**
	A	11	9	20			
	B	0	6	6			
合计		24	30	54			
类型	高钾	12	6	18	5.4	4.134	0.20**
	铅钡	12	24	36			
合计		24	30	54			
颜色	蓝绿	6	9	15	6.287	6.287	0.507
	浅蓝	8	12	20			
	紫	2	2	4			
	深绿	3	4	7			
	深蓝	2	0	2			
	浅绿	2	1	3			
	黑	0	2	2			
	绿	1	0	1			
合计		24	30	54			

注：***、**、*分别代表 1%、5%、10%的显著性水平

我们假设风化与纹饰，颜色，玻璃类型之间不存在相关性，由表 1 中我们可以得出结论基于风化程度与纹饰的 p 值为 0.056，在水平上呈现的显著性不存在接受假设，风化程度与纹饰它们俩之间不存在显著差异。风化程度与类型的 p 值为 0.020 在水平上呈现出显著性，所以拒绝假设，风化程度与纹饰之间存在着显著差异。风化程度与颜色的 p 值为 0.507，则在水平上呈现不明显，接受假设，它们俩之间显著性不存在。然后我们以玻璃材料为唯一变量在对其余二者进行卡方检验得出结果如表 3 所示。

表 3 类型对纹饰和颜色的显著性的卡方检验分析结果

题目	名称	类型		总计	X^2	校正 X^2	P
		高钾	铅钡				
纹饰	C	6	22	28	13.886	13.886	0.001***
	A	6	14	20			
	B	6	0	6			
合计		18	36	54			
颜色	蓝绿	12	3	15	22.693	22.693	0.002***
	浅蓝	4	16	20			
	紫	0	4	4			
	深绿	1	6	7			
	深蓝	1	1	2			
	浅绿	0	3	3			
	黑	0	2	2			
	绿	0	1	1			
合计		18	36	54			

注：***、**、*分别代表 1%、5%、10%的显著性水平

同上我们一样假设玻璃材料与纹饰和颜色之间不存在相关性，由表 3 中我们可得基于类型和纹饰的 P 值为 0.001***，在水平上呈现显著性，所以拒绝原假设，玻璃材料和纹饰之间存在着显著差异，同理可得玻璃材料与颜色之间也存在着显著差异，最后我们对剩下的两个再一次进行卡方检验得出表 4。

表 4 纹饰对颜色的显著性的卡方检验分析结果

题目	名称	纹饰			总计	X^2	校正 X^2	P
		C	A	B				
颜色	蓝绿	3	6	6	15	38.11	38.11	0.001***
	浅蓝	10	10	0	20			
	紫	4	0	0	4			
	深绿	7	0	0	7			
	深蓝	0	2	0	2			
	浅绿	3	0	0	3			
	黑	0	2	0	2			
	绿	1	0	0	1			
合计		28	20	6	54			

注：***、**、*分别代表 1%、5%、10%的显著性水平

这次的假设同上，由图可知颜色与纹饰之间存在着显著的差异。在对于不同的玻璃类型进行有无风化的化学成分含量统计时，我们将所有的元素进行分类汇总方便我们后面统计规律。

然后，将铅钡材料以及高钾材料的玻璃文物进行风化前与风化后相对重要的化学成分频率分布直方图来进行对比分析，将数据导入 spsspro 然后在数据分析中选择描述性分析里面的分类汇总，将类型和表面风化改成定类拖入分组，将剩下的的元素放入汇总得到如图 1 所示的结果。

由频率分布直方图中的信息可以直观的看出高钾材料的玻璃主要成分的含量在风化后呈现下降的趋势，铅钡材料的玻璃主要成分的含量在风化后呈现上升的

趋势。首先，我们对附件中的颜色缺失值进行中位数填充，将缺失的数据用 0 来填充，然后尝试去用最简单的线性回归模型去把所有元素作为 y 值去程序中跑出多个线性回归模型，将数据导出后通过对模型的解读发现存在两个元素分别是氧化纳和氧化铁在水平上不呈现显著性，不能拒绝回归系数为 0 的假设，所以模型无效不能使用线性回归模型。接着，尝试寻找其他的预测模型去解决，最终选择岭回归方法。即将另外俩个不符合线性回归的元素提取出来对数据进行了处理，再用岭回归模型进行分析求解得出它们的方程关系式[3]，然后在程序中将表单中已风化的玻璃文物样品筛选出来，将它们的属性（颜色、纹饰、类型）采用 01 化将所有回归方程关系式写入程序方便计算，导入已经 01 化的数据放进程序来预测出玻璃文物样品在未风化前的化学成分含量。

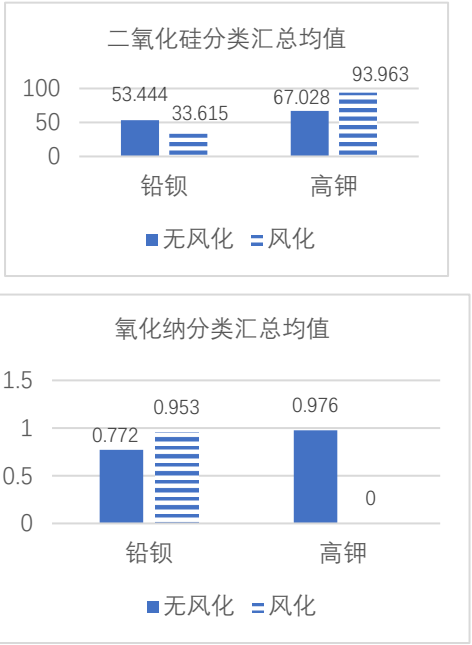

图 1　风化前后元素含量直方图（部分）

2．2　问题二的模型汇总与求解

因为玻璃的分类有多种化学元素的影响，不同变量之间的相互影响因素太多，使得不容易去找出它们之间的分类规律，于是我们用一些数学工具和代码建立各种不同元素的决策变量之间的关系式与模型，将复杂的问题简单化，然后采取决策树模型去解决这个问题。决策树的遗传算法可以加大结果的可信程度，因为它是串集开始搜索的，不是从某一个解开始的，这样就能避免的传统算法的缺陷，减少陷入局部最优解的风险[4]。因为是从串集开始搜索的，所覆盖的面积大而广，利于全局则优。为此，将所有数据导入 spsspro 选择决策树遗传算法训练出的训练占比为 1 后的结果发现，我们只需要通过氧化铅的含量（如图 2 所示）就可以判断出玻璃类型。当含量超过 5.46 是为铅钡玻璃，低于 5.46 时为高钾玻璃，训练的准确率和召回率都是 1，说明这个模型对于这个数据的处理非常合适，拟合程度非常高的。

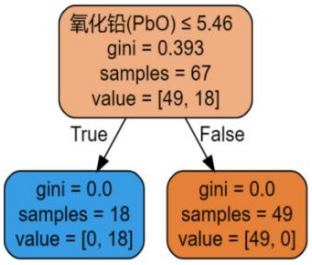

图 2　高钾玻璃与铅钡玻璃分类的决策树

在解决对于玻璃文物做亚分类划分的问题时，把附件里面的表格拆分成高钾玻璃和铅钡玻璃两个数据集方便我们去划分不同类型玻璃的成分。接着为了使样本划分的更精细，误差更小，我们就使用了肘部法思想。因为在做机器学习中聚类 clustering 的时候，聚类数量 K 值的选择其实是核心问题我们就经常使用肘部法则，最小化点到聚类中心距离，并进行程序设计[5]，计算 K 值从 2-10 的情况，随机数为 4 并绘画出高钾的 elbow 图（如图 3 所示）。

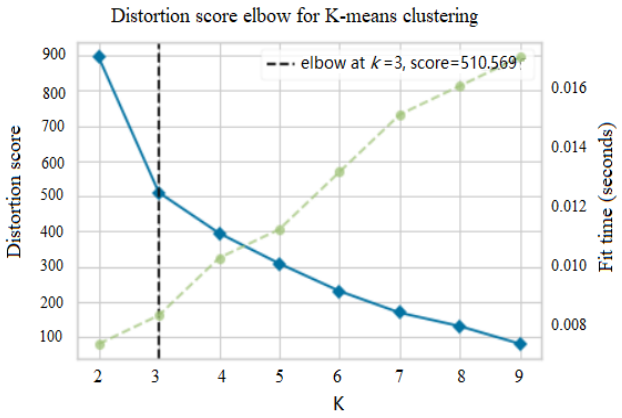

图 3　高钾玻璃的 K 值变化图

由于有随机数的存在图中的 k 值在 3 和 4 之间摆动，多次运行后观察 elbow 图，最后选择选择 4 作为 k 值放入聚类分析 K-Means 算法里面运行出的数据保存到表格中，然后将表格上传到 spsspro 中继续选择决策树-遗传算法在训练系数 1 的情况下进行训练最后得出高钾亚分类的决策树结构如图 4 所示。

训练后发现，决策树的训练集中准确率、召回率和精确率都是 1，说明模型的亚分类划分合理性是非常强的。又根据相同操作去运行程序求解铅钡类的文物的亚类划分，运行得出的 k 值 elbow 图如图 5 所示。

因为有随机数的存在，所以可能每次绘出来的图都不太一样，但是 k 值是固定在 5 的。同理，将 5 代

入程序中进行聚类分析后把得出来的结果写入到 2.2 铅钡表格中，也将它上传到 spsspro 中进行决策树-遗传算法的分析，最后得出铅钡的亚分类决策树图（如图 6 所示）。

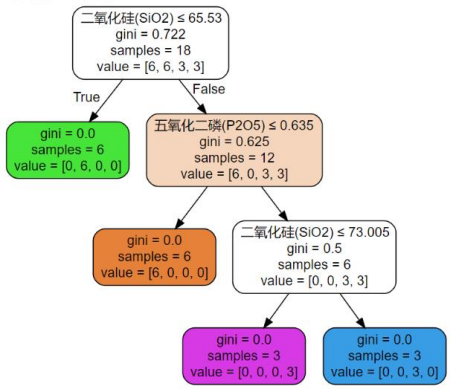

图 4　高钾玻璃亚类划分的决策树

训练后铅钡的训练集指数也都是 1，这是一种期望的理想情况，说明决策树图中对铅钡的亚分类也是非常合理的。原因是我们采用的是机器学习模型去解决问题，所以在分析肘部法则的 K 值时图中已经表达出了敏感性。因为机器学习在训练的时候，预测模型已经固定了，所以不用在对敏感性进行分析了，对此问题 2 已经完全解决了。

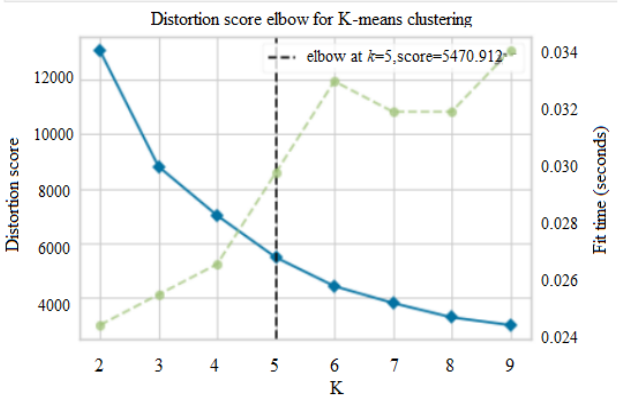

图 5　铅钡玻璃的 K 值变化图

2．3　问题三的模型汇总与求解

首先，将附件所给的表放入程序中读出将空值进行填补方便我们后续对数据进行处理，然后将表面风化这个属性进行另外的 01 编码。

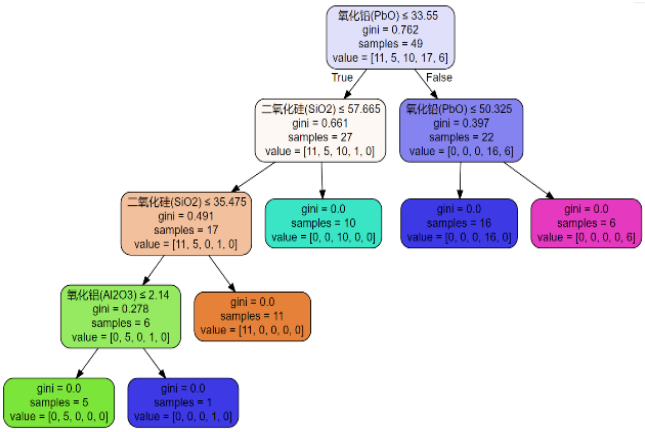

图 6　铅钡玻璃亚类划分的决策树

将所有的数据处理好后，将原来得到的数据放入程序内让程序去深度学习，让机器去判断未知类型的文物材料类型。在选用分类算法时为了避免 XGBoost 分类算法消耗空间大和训练时间久的缺陷情况出现，选择使用基于 Histogram 的决策树算法 LGBM 分类算法，它可以以更小的内存占用和更小的计算代价去高效的运行[6]，将数据集导入后直接调用 LGBM 算法运行结果如图 7 所示。

```
# lgbm分类
model = lgb.LGBMClassifier()
model.fit(x_train, y_train)
print('lgbm分类')
print(classification_report(model.predict(x_test),y_test))
```

```
lgbm分类
              precision    recall  f1-score   support

          铅钡       1.00      1.00      1.00        17
          高钾       1.00      1.00      1.00         4

    accuracy                           1.00        21
   macro avg       1.00      1.00      1.00        21
weighted avg       1.00      1.00      1.00        21
```

图 7　LGBM 算法的训练结果

由图中可知，这次训练的精度、召回率和 F1 的值都为 1，说明这次机器训练被完全正确的进行了分类。将处理好的表 3 的数据集继续放入程序里面让它进行分类后得出的结果如图 8 所示。

```
array(['高钾', '铅钡', '铅钡', '铅钡', '铅钡', '高钾', '高钾', '铅钡'], dtype=object)
```

图 8　LGBM 算法的分类结果

在分析敏感性时因为采用的是机器学习模型，所以只需要在训练的时候分析敏感性即可。因为预测时的模型已经固定了，我们继续沿用上面的 LGBM 算法，用树节点和树深度两个变量来训练。为了获得更加精确的数据，将树节点以 1 到 40 的范围以 1 为迭代来运行，画出以 x 轴为树目 y 轴为准确率的敏感度分析图

（如图 9 所示）。

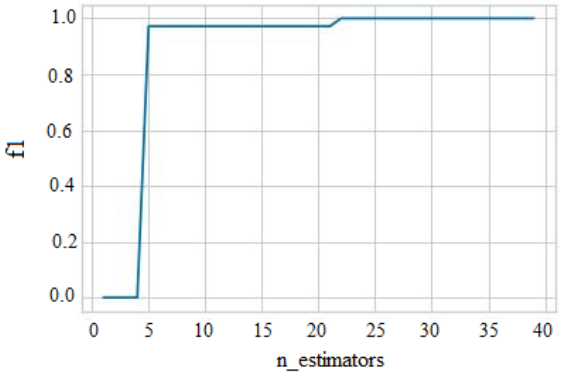

图 9 树节点的敏感性分析

然后我们取图中稳定点 25 为参数，范围用 1 到 50 以 1 为迭代去算树深度的敏感度并绘图（如图 10 所示）。从图中可以看出数值一直是 1，则说明这个算法分类的敏感度很低。

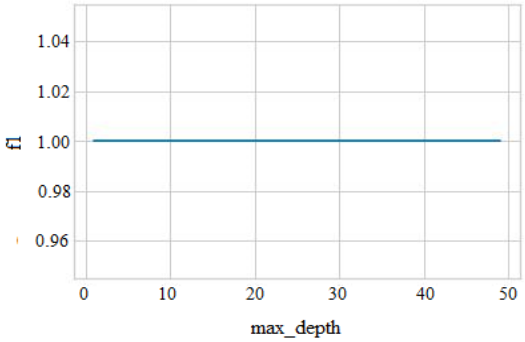

图 10 树深度的敏感性分析

2.4 问题四的模型汇总与求解

为了方便求解问题 4 中的第 1 个小问题，我们用程序将表单中的高钾和铅钡属性的数据分离出来保存为高钾表格和铅钡表格。因为存在多种元素，所以继续沿用问题 1 中的方法，通过 Spearman 相关检验来分析不同种元素之间的相关性。将高钾表格上传到 spsspro 中选择相关性分析中的 Spearman 相关系数分析并且同上操作得铅钡的 Spearman 相关系数分析。

根据 Spearman 相关系数的特性系数越接近 1 或-1 相关性越大，可以由图中的信息可以看出不同材料玻璃文物之间化学成分的相关性[1]。因为标中的数据不符合正态性检验，所以使用独立样本 MannWhitney 检验，利用两个样本之间观察值的平均秩次来推断俩样本分别代表总体中位数有无差异，并且可以对其的置信区间和总体中位数进行计算[7]，为了方便数据的操作，将运行程序到第四问题的第二步骤会分离出一个整理好如表 5 所示的表格，将其表格上传到 spsspro 上面对其进行数据分析。为了验证数据不满足正态性检验，我们对其进行验证。假设满足正态分布猜想，由分析表中可以得出其所有元素的 p 值都小于 0.05，所以成显著性。这否定了假设，对其不同元素的正态分布图如图 11 所示。

表 5 第四问题的第二步骤会分离出的整理好的结果

变量名	样本量	平均值	标准差	偏度	峰度	S-W 检验	K-S 检验
二氧化硅（SiO2）	67	49.022	24.316	0.183	-0.818	0.963(0.042**)	0.102(0.457)
氧化纳(Na2O)	67	0.786	1.653	2.446	6.106	0.554(0.000**)	0.429(0.000**)
氧化钾（K2O）	67	1.847	3.879	2.113	3.008	0.526(0.000**)	0.402(0.000**)
氧化钙(CaO)	67	2.532	2.325	1.038	0.214	0.879(0.000**)	0.154(0.076**)
氧化镁(MgO)	67	0.683	0.65	0.616	-0.099	0.883(0.000**)	0.226(0.002**)
氧化铝(AI2O3)	67	4.041	3.067	1.626	2.782	0.844(0.000**)	0.141(0.125)
氧化铁Fe2O3	67	0.849	1.179	2.087	5.541	0.741(0.000**)	0.236(0.001**)
氧化铜(CuO)	67	1.945	2.242	2.176	5.447	0.75(0.000**)	0.213(0.004**)
氧化铅(PbO)	67	24.463	19.513	0.263	-0.951	0.929(0.000**)	0.148(0.097**)
氧化钡(BaO)	67	7.779	8.425	1.659	2.736	0.806(0.000**)	0.189(0.015**)
五氧化二磷(P2O5)	67	2.684	3.554	1.479	1.389	0.764(0.000**)	0.262(0.000**)
氧化锶(SrO)	67	0.262	0.267	1.051	0.808	0.875(0.000**)	0.164(0.049**)
氧化锡(SnO2)	67	0.078	0.337	5.648	34.673	0.249(0.000**)	0.502(0.000**)
二氧化硫(SO2)	67	0.603	2.698	5.284	27.846	0.233(0.000**)	0.469(0.000**)

注：***、**、*分别代表 1%、5%、10%的显著性水平

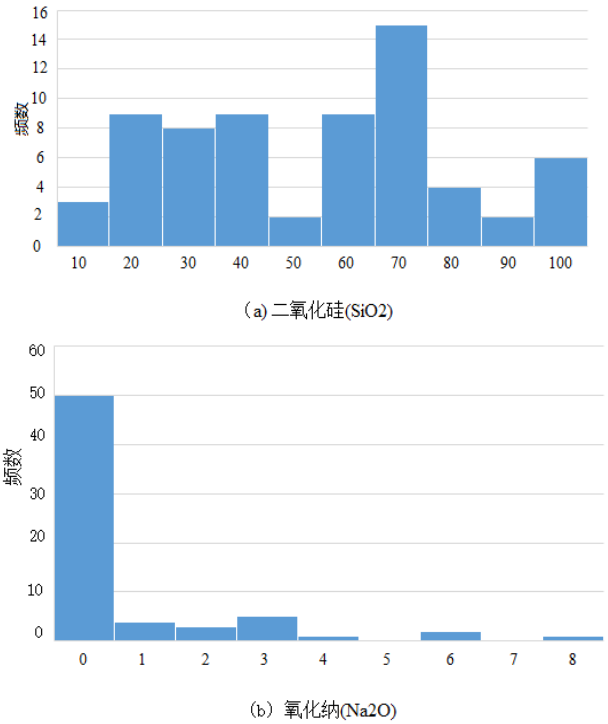

图 11 部分元素的正态分布图

表 6　MannWhitney U 部分检验分析结果表

变量名	变量值	样本值	中位数	标准差	统计量	P	中位数值差值	Cohen D值
二氧化硅(SiO2)	高钾	18	73.01	14.47	826	0.00***	37.225	2.14
	铅钡	49	35.78	18.65				
	合计	67	51.3	24.32				
氧化纳(Na2O)	高钾	18	0	1.089	389.5	0.341	0	0.267
	铅钡	49	0	1.813				
	合计	67	0	1.653				
氧化钾(K2O)	高钾	18	7.525	5.308	763.5	0.00***	7.525	0.816
	铅钡	49	0	0.276				
	合计	67	0.2	3.879				
氧化钙(CaO)	高钾	18	3.36	3.308	552	0.116	1.88	0.816
	铅钡	49	1.48	1.635				
	合计	67	1.66	2.325				

注：***、**、*分别代表 1%、5%、10%的显著性水平

由于数据不满足正态分布，所以可以进行独立样本 MannWhitney 检验，在 spsspro 选择差异性分析中的独立样本 MannWhitney 检验，得出它们的分析结果如表 6 所示。

3　结束语

本文使用卡方检验、岭回归方法、肘部法等一系列的方法和聚类分析 K-Means 算法、决策树-遗传算法、LGBM 分类算法等，建立了可用于对古代玻璃制品成分进行分析与鉴别的决策树机器分类模型、亚分类决策树模型，将古代玻璃制品分为高钾玻璃和铅钡玻璃两种类型。可以利用 Spsspro 软件，根据一些已知的数据分析把对古代玻璃制品成分分析分为四类进行分析处理。部分检验分析结果说明了本文所提出模型的可用性和有效性。

参 考 文 献

[1] Piantadosi J,Howlett P, Boland J.Matching the grade correlation coefficient using a copula with maximum disorder[J].Journal of Industrial and Management Optimization, 2007,3(2):305–312

[2] 王金桃, 周利锋, 高尔生. 第六讲 卡方检验[J]. 实验动物与比较医学, 2000(4):251-254.

[3] 万丽颖.岭回归分析及其应用[J].许昌学院学报, 2016，35(2):19-23

[4] 吴菲,黄梯云.用遗传算法构造二元决策树[J].计算机研究与发,1999, 36(11):1323-1328.

[5] 龙文佳,张晓峰,张链.基于 k-means 和肘部法则的业务流程聚类方法[J].江汉大学学报：自然科学版, 2020(1):81-90

[6] 谢勇,项薇,季孟忠,等.基于 Xgboost 和 LightGBM 算法预测住房月租金的应用分析[J].计算机应用与软件, 2019, 36(9):151-155+191.

[7] Ruxton G D .The unequal variance t-test is an underused alternative to Student's t-test and the Mann–Whitney U test[J].Behavioral Ecology, 2010, 17(4):688-690