古代玻璃制品的成分分析与鉴别

摘　要

古代玻璃制品成分种类繁多，而且成分组成复杂多样，也因技术受限，考古工作者无法很好地鉴别出土文物。在此背景下，本文着重研究古代玻璃制品属性和成分的分析与鉴别。

针对问题一，本文先进行数据预处理，筛选掉无效数据并按属性分类，根据整理的数据对文物表面风化与其类型、纹饰和颜色的关系进行**卡方检验**，得出 p 值分别为 0.009<0.05、0.307>0.05 和 0.084>0.05，说明文物表面风化与其类型密切相关。接着，对数据进行描述性统计，得出各化学成分含量的均值、方差等数据，通过折线图发现已风化高钾玻璃文物二氧化硅含量高等统计规律。最后，根据所得各化学成分含量均值求出风化化学成分含量，据此可预测风化文物风化前化学成分含量。

针对问题二，本文分别从化学成分含量、纹饰、颜色的角度来区分两种玻璃，发现铅钡玻璃文物的氧化铅(PbO)、氧化钡(BaO)含量占比都较高，且缺少 B 纹饰，但含有的颜色类型比高钾玻璃文物的更高。为了进行亚类划分，本文先根据标准差分别对高钾、铅钡玻璃提取出关键化学成分，然后用 **Kmeans** 进行不同类数的聚类，最终经过分析，**将高钾玻璃细分为 2 类，将铅钡玻璃细分为 3 类**。

针对问题三，为缓解小数据集的过拟合问题，本文首先对划分后的数据集进行**主成分分析**特征降维，然后利用提取出的主成分进行分类；本文提出并采用了**简单启发式均值分类模型**进行分类，该模型较为简单，适用于小数据集的分类，在题目所给的检测点数据中该模型取得了 **100%**的分类正确率，模型较为可靠。

针对问题四，本文采用散点图和**斯皮尔曼相关系数**的方法分析各类别玻璃文物化学成分的关系。运用 Python 的 Seaborn 库进行图形绘制和 SPSS 求解相关系数，剔除不满足显著性检验的数据之后可以得出高钾玻璃文物的氧化锡和二氧化硫均与其他化学成分无关联等成分关联关系和两种玻璃之间，高钾玻璃中二氧化硫与其他化学成分无关联，而铅钡玻璃中的二氧化硫与大多数其他化学成分呈正或负相关等成分差异性。

关键词：卡方检验　Kmeans　主成分分析　启发式均值分类 斯皮尔曼相关系数

一、 问题重述

玻璃的主要原料是石英砂（主要化学成分是二氧化硅），但由于在炼制时所添加的助熔剂不同，玻璃的主要化学成分不同。根据主要化学成分的不同，可以将玻璃分为不同的类别。古代玻璃极易受埋藏环境的影响而风化，导致其成分比例发生变化，这样会影响考古工作者对其类别的判断。现有一批我国古代玻璃制品的相关数据，考古工作者已经通过一些检测手段将这批样品分为高钾玻璃和铅钡玻璃两种类型。现需要对附件中的相关数据进行分析建模来求解以下问题：

问题 1：首先，需要通过对数据分析来发现玻璃文物的表面风化与其玻璃类型、纹饰、颜色的关系；其次，需要针对不同的玻璃类型，分析文物的表面有风化、表面无风化的化学成分含量的统计规律；最后，根据前面的统计规律，预测风化点数据在风化前的化学成分含量。

问题 2：首先，根据附件的数据来分析高钾玻璃和铅钡玻璃的分类规律；其次，针对每个类别，通过选择合适的化学成分区别来进一步对其分类进而得到亚类划分结果，需要给出具体的划分方法，并分析分类结果的合理性和敏感性。

问题 3：对附件表单 3 中未知类别的玻璃文物的化学成分进行分析，进而给这些文物分类并分析分类结果的敏感性。

问题 4：分析每个类别中玻璃文化样品的化学成分之间的关联关系，比较不同类别化学成分关联关系的差异性。

二、 问题分析

2.1 问题一的分析

附件表单 1 提供了玻璃文物类型、纹饰等对应数据信息，问题一需要我们分析风化与类型、纹饰和颜色三者之间的关系，并且根据玻璃类型，分类统计已风化文物的化学成分含量的规律和未风化文物的化学成分含量的规律，最后估计已风化文物风化前的化学成分含量。

首先，计算文物风化、类型等出现频数，运用 SPSSAU 对玻璃文物的表面风化与其玻璃类型、纹饰和颜色的关系进行卡方检验，依据卡方检验得到的结果分析关系。接着，对原有数据进行筛选与分类，运用 SPSS 对各类玻璃类型的化学成分含量进行描述性统计分析，得到其统计规律。最后，根据描述性统计得到的样本总体均值算出

风化文物与未风化文物的化学成分含量的差异，以此预估文物风化前的化学成分含量。

2.2 问题二的分析

问题二需要我们依据已给出的高钾玻璃与铅钡玻璃的各化学成分含量以及其颜色和纹饰，分析两种玻璃的分类规律，并对这两种类别再次进行细分类，分析细分类结果的合理性与敏感性。

首先，由于风化前后对玻璃的化学成分含量影响较大，需要将风化前后的样本区分开来，找出风化前后的共性点，最后比较两种玻璃的化学含量占比区别。除此之外，可以通过绘制玻璃类型与颜色、纹饰的交叉图来直观感受两种玻璃类型的表面区别。

对于亚类划分部分，需要先筛选出主要的化学成分，如在不同样本中差异较大的化学成分。在筛选得到主要化学成分之后，可以通过聚类的方式将样本集聚成不同数量的类别，并对其进行分析，最终找出比较合适的分类。

2.3 问题三的分析

问题三承接着问题二中的分析文物分类规律，需要我们鉴别未知的玻璃文物，并对所得结果进行敏感性分析。

首先，我们根据玻璃文物的类型和表面风化进行数据集的划分（划分为高钾-风化、高钾-未风化、铅钡-风化、铅钡-未风化四类），然后利用主成分分析法对化学成分特征进行提取，起到特征降维的作用，最后利用提取出的主成分和简单启发式均值分类模型对待分类玻璃文物进行分类。

2.4 问题四的分析

根据附件表单 1 我们可以将玻璃文物样品划分为多种类别进行其化学成分之间关联性与差异性的分析。但由于样本数量较少，为保证数据结果的可靠性，本文取高钾玻璃与铅钡玻璃两种类别进行分析。

首先，将原有的数据进行预处理，筛选掉无效数据并进行分类。接着，运用 Python 的 Seaborn 库绘制各类别化学成分散点图矩阵，初步判断各化学成分之间的相关性，并利用 SPSS 进行斯皮尔曼相关系数的计算，筛选掉不满足显著性检验的数据。最后，整理汇总化学成分相关系数表，结合散点图分析关联性与差异性。

三、 模型假设

1. 假设没有检测出来的元素，默认占比为 0，即无此成分。
2. 假设所给样本特征能够代表总体特征。
3. 假设颜色为空的数据单元为无颜色。

四、 符号说明

符号	说明	单位
H_0	原假设	
χ^2	卡方值	
W_{ij}	文物风化前化学成分含量	%
F_{ij}	文物化学成分含量	%
M_j	文物风化化学成分含量均值	%
d_i	等级差	
r_s	斯皮尔曼相关系数	

五、 模型的建立与求解

5.1 问题一模型的建立与求解

问题一的总体分析思路如图 5.1.1 所示，其中，数据预处理所得表格可见附录。

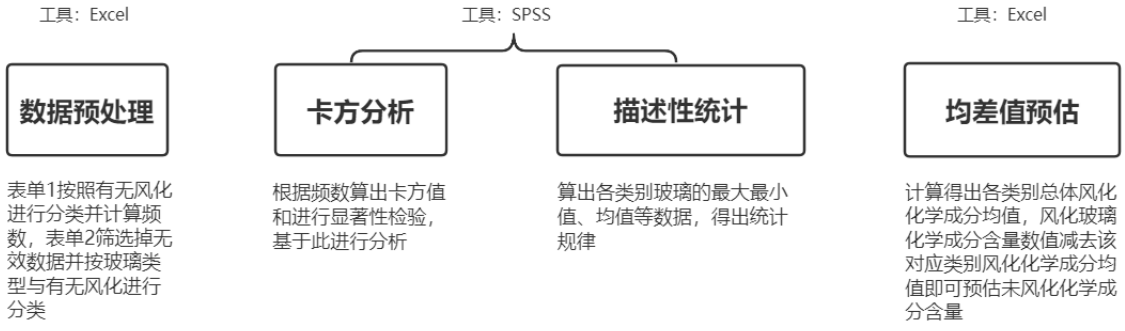


图 5.1.1 问题一分析思路图

5.1.1 卡方检验模型的建立

（1） 卡方检验

由于附件表单 1 为定性数据，分析玻璃文物表面风化与其玻璃类型、纹饰和颜色

的关系可运用卡方分析，卡方分析是用来研究两个定类变量间是否独立即是否存在某种关联性的最常用的方法[1]。

●提出原假设H_0：玻璃类型、纹饰和颜色与文物表面风化的关系均相同，不存在差异。

●将总体取值范围分成k个互不相交的小区间。

●把落入某个区间的的样本值的个数记作f_i。

●当原假设为真时，可算出总体值落入某个区间的概率为p_i，当样本量为n时，np_i即为即为落入某区间的期望值。

●计算卡方统计量

$$\chi^2 = \sum_{i=1}^{k} (f_i - np_i)^2 / np_i \tag{1}$$

●显著性检验：根据显著性水平和自由度查找卡方分布临界值表，与所求卡方值对比，若实际卡方值大于理论卡方值，则拒绝原假设。

（2） 均差值预估公式

附件表单 2 给出了每个玻璃文物个体的化学成分含量，要想估计风化后文物所变化的化学成分含量，需要从总体数据中提取出来，以此预估每个风化文物个体风化前的化学成分含量。

$$W_{ij} = F_{ij} - M_j \tag{2}$$

其中，W_{ij}是第i个已风化高钾（或铅钡）玻璃文物的第j个化学成分风化前的含量，F_{ij}是第i个已风化高钾（或铅钡）玻璃文物的第j个化学成分含量，M_j是高钾（或铅钡）玻璃文物第j个因风化所产生的化学成分含量均值。

5.1.2 文物属性及化学成分分析

（1） 卡方分析

通过卡方检验分析玻璃文物表面风化与其玻璃类型、纹饰和颜色的关系，我们运用 SPSSAU 工具得出卡方值，判断其是否存在一定关联度，表 5.1.1 给出了卡方检验的结果。

表 5.1.1 卡方检验结果表

题目	名称	表面风化(%)	总计	X^2	p

		无风化	风化			
类型	铅钡	12(50.00)	28(82.35)	40(68.97)	6.880	0.009**
	高钾	12(50.00)	6(17.65)	18(31.03)		
总计		24	34	58		
颜色	无	0(0.00)	4(11.76)	4(6.90)	9.432	0.307
	浅绿	2(8.33)	1(2.94)	3(5.17)		
	浅蓝	8(33.33)	12(35.29)	20(34.48)		
	深绿	3(12.50)	4(11.76)	7(12.07)		
	深蓝	2(8.33)	0(0.00)	2(3.45)		
	紫	2(8.33)	2(5.88)	4(6.90)		
	绿	1(4.17)	0(0.00)	1(1.72)		
	蓝绿	6(25.00)	9(26.47)	15(25.86)		
	黑	0(0.00)	2(5.88)	2(3.45)		
总计		24	34	58		
纹饰	A	11(45.83)	11(32.35)	22(37.93)	4.957	0.084
	B	0(0.00)	6(17.65)	6(10.34)		
	C	13(54.17)	17(50.00)	30(51.72)		
总计		24	34	58		

注：* p<0.05　** p<0.01

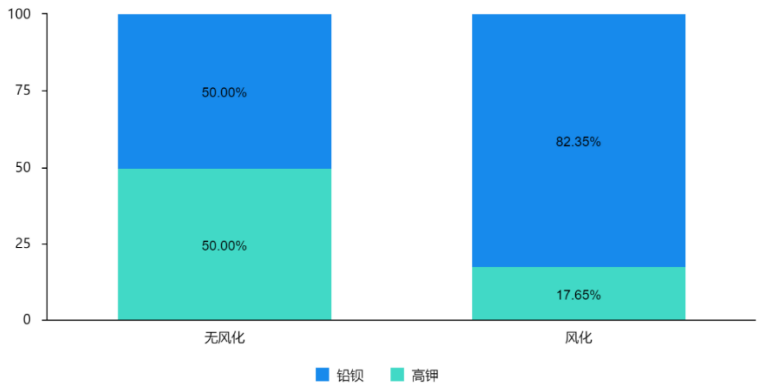


图 5.1.2 表面风化和类型的堆积柱形图

从表中可以看出，不同文物表面风化对于颜色和纹饰未表现出显著性(p>0.05)，意

味着文物表面是否风化与颜色和纹饰之间的关系不大。而文物表面风化对于类型项呈现出显著性($p<0.05$)，意味着文物表面是否风化与玻璃文物类型密切相关。

表面风化对于类型呈现出 0.01 水平显著性($chi=6.880, p=0.009<0.01$)，通过图 5.1.2 柱形图对比可知，已风化的铅钡玻璃文物的比例 82.35%，明显高于无风化的比例 50.00%。无风化高钾玻璃文物的比例 50.00%，明显高于风化的选择比例 17.65%。

总结可知，文物表面是否风化与玻璃类型相关程度大，铅钡玻璃文物易风化，高钾玻璃文物不易风化。

（2） 统计规律分析

运用 Excel 将表单 2 中的化学成分分为高钾玻璃风化，高钾玻璃未风化，铅钡玻璃风化，铅钡玻璃未风化四种，分别使用 SPSS 软件求其各化学成分含量的均值、方差等数据（具体数据见支撑材料），本文采用均值折线图再结合方差分析统计规律。

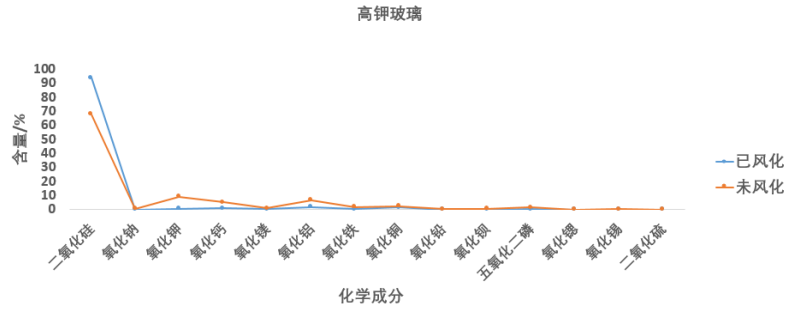


图 5.1.3 高钾玻璃化学成分含量均值折线图

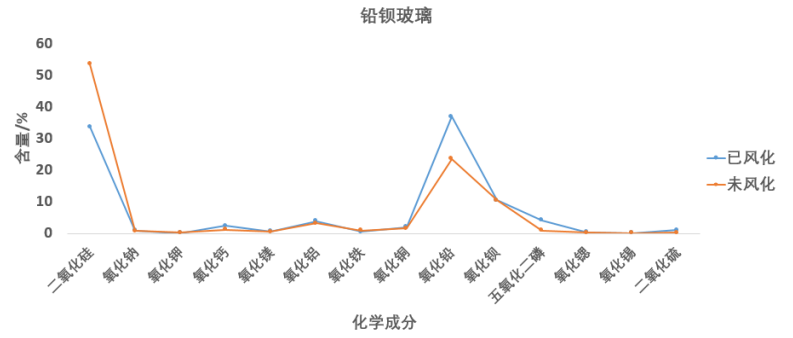


图 5.1.4 铅钡玻璃化学成分含量均值折线图

对于高钾玻璃，风化文物二氧化硅含量明显高于未风化文物，且其方差更低，即二氧化硅含量水平高且稳定；十四种化学成分均存在于未风化文物中，而风化后未有氧化钠、氧化铅、氧化钡、氧化锶、氧化锡和二氧化硫。

对于铅钡玻璃，风化文物二氧化硅含量明显低于未风化文物，而氧化铅则相反。

高钾玻璃与铅钡玻璃风化时与未风化时二氧化硅含量变化趋势相反。

（3） 风化前化学含量预测

由所得每种玻璃化学成分含量的均值可以计算出风化化学成分含量，如表 5.1.2 和表 5.1.3 所示（具体请见支撑材料）。

表 5.1.2 高钾玻璃风化化学成分含量表

高钾玻璃化学成分	已风化	未风化	风化化学成分含量
二氧化硅	93.9633	67.98	25.98
氧化钠	0	0.695	-0.7
氧化钾	0.54333	9.331	-8.79
氧化钙	0.87	5.333	-4.46
…	…	…	…

表 5.1.3 铅钡玻璃风化化学成分含量表

铅钡玻璃化学成分	已风化	未风化	风化化学成分含量
二氧化硅	33.61	53.44	-19.8
氧化钠	0.953	0.772	0.181
氧化钾	0.143	0.258	-0.12
氧化钙	2.346	1.232	1.114
…	…	…	…

根据所建均差值预估公式，可以预测出每个风化文物个体风化前的化学成分含量，如表 5.1.4 和表 5.1.5 所示（具体请见支撑材料）。

表 5.1.4 高钾玻璃风化前化学成分含量

类型	表面风化	文物采样点	二氧化硅(SiO2)	氧化钾(K2O)	氧化钙(CaO)	…
高钾	风化	07	66.65083	8.7875	5.5325	…
高钾	风化	09	69.04083	9.3775	5.0825	…
高钾	风化	10	70.79083	9.7075	4.6725	…
高钾	风化	12	68.31083	9.7975	5.1825	…
高钾	风化	22	66.37083	9.5275	6.1225	…

高钾	风化	27	66.74083	8.7875	5.4025	…
…	…	…	…	…	…	…

表 5.1.5 铅钡玻璃风化前化学成分含量

类型	表面风化	文物采样点	二氧化硅(SiO2)	氧化钾(K2O)	氧化钙(CaO)	…
铅钡	风化	02	56.10912	1.165684	1.225983	…
铅钡	风化	08	39.96912	0.115684	0.365983	…
铅钡	风化	08 严重风化点	24.43912	0.115684	2.075983	…
铅钡	风化	11	53.41912	0.325684	2.395983	…
铅钡	风化	19	49.46912	0.115684	1.815983	…
铅钡	风化	26	39.61912	0.115684	0.325983	…
…	…	…	…	…	…	…

可以看出，风化前文物均存在化学成分，与未风化文物原始数据基本相符，且主要化学成分例如二氧化硅等成分含量与原始数据中未风化文物的化学成分含量水平基本一致，预测效果较好。

5.2 问题二模型的建立与求解

5.2.1 玻璃文物分类规律分析

首先，本文从化学成分含量占比的角度来分析高钾玻璃、铅钡玻璃的分类规律，由于风化前后玻璃的化学成分变化较大，本文将风化前后的数据进行区分，图为风化前后高钾玻璃、铅钡玻璃平均化学含量占比的柱形图。从图 5.2.1 中可以明显看出，铅钡玻璃文物的氧化铅(PbO)、氧化钡(BaO)含量占比都较高，被称为铅钡玻璃可能与该特点有一定的关系。

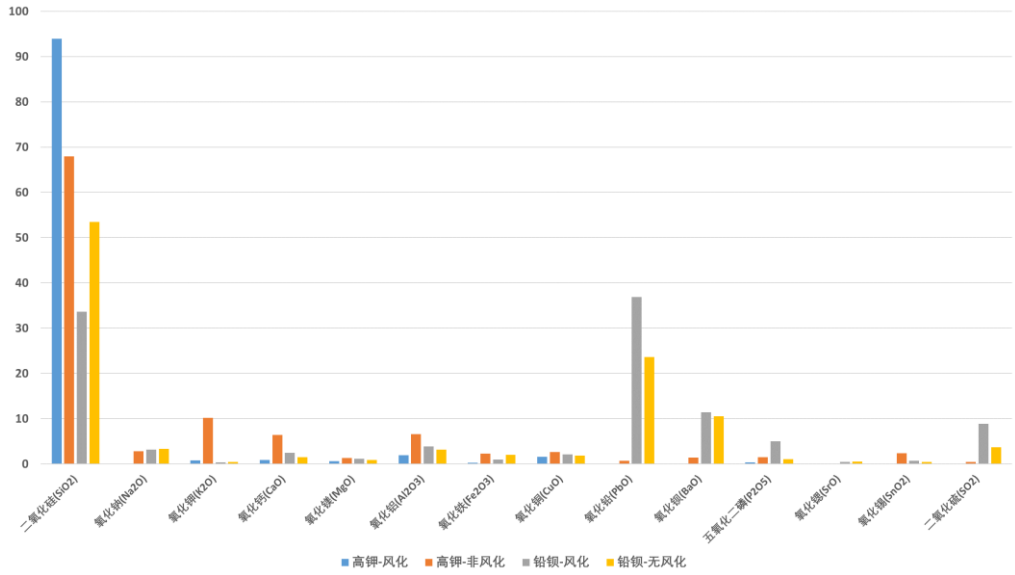


图 5.2.1 化学成分含量占比

除了从具体的化学成分来分析之外，如果可以通过表面的信息来简单判断玻璃的类别，这将会更加高效。由于所给数据的样本量较少，假设题目所提供数据的统计规律可以代表大样本数据的统计规律。通过图 5.2.2 可知，铅钡玻璃文物中没有纹饰为 B 类纹饰的，即若给定一个文物，其纹饰为 B 类纹饰，可以推测其为高钾玻璃；通过图 5.2.3 可知，高钾玻璃文物和铅钡玻璃文物的颜色分布有区别，如铅钡玻璃文物中含有无色的类型，则若遇到无色的玻璃，可大致判断其为铅钡玻璃。

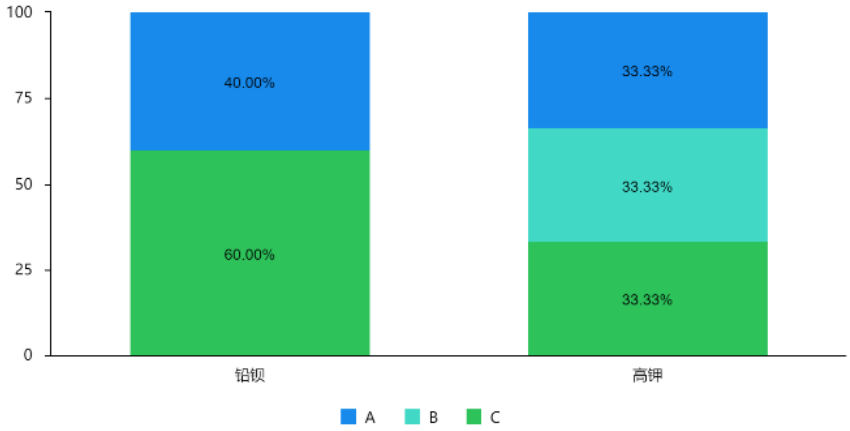


图 5.2.2 玻璃类型和纹饰的交叉图

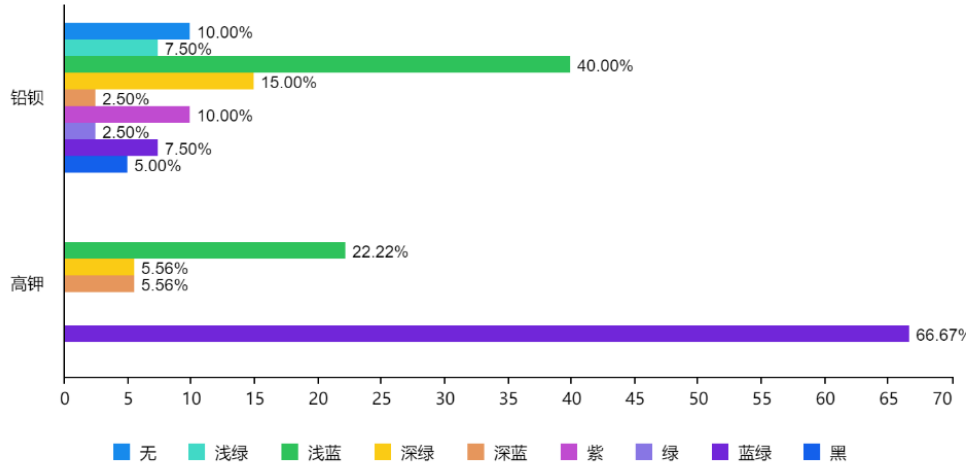


图 5.2.3 玻璃类型和颜色的交叉图

5.2.2 基于 Kmeans 聚类的亚类划分

本文通过计算每个类别每种化学成分的标准差，因为标准差较小的化学成分数据列数据相差不大，作为分类的依据效果较差，因此本文选择出标准差较大（标准差>1）的化学成分来作为亚类划分的特征输入，主要化学成分选择表 5.2.1 主要化学成分选择中被标白的化学成分。

表 5.2.1 主要化学成分选择

高钾		铅钡	
二氧化硅(SiO_2)	14.0591	二氧化硅(SiO_2)	18.4552
氧化钾(K_2O)	5.1582	氧化铅(PbO)	14.7940
氧化钙(CaO)	3.2149	氧化钡(BaO)	8.2459
氧化铝(Al_2O_3)	2.9900	五氧化二磷(P_2O_5)	3.8691
氧化铁(Fe_2O_3)	1.5219	二氧化硫(SO_2)	3.1065
氧化铜(CuO)	1.3919	氧化铝(Al_2O_3)	2.9781
五氧化二磷(P_2O_5)	1.2445	氧化铜(CuO)	2.4451
氧化钠(Na_2O)	1.0580	氧化钠(Na_2O)	1.7942
氧化钡(BaO)	0.8179	氧化钙(CaO)	1.6178
氧化镁(MgO)	0.6917	氧化铁(Fe_2O_3)	0.9387
氧化锡(SnO_2)	0.5406	氧化镁(MgO)	0.6236

氧化铅(PbO)	0.4997	氧化钾(K_2O)	0.2727
二氧化硫(SO_2)	0.1527	氧化锶(SrO)	0.2608
氧化锶(SrO)	0.0426	氧化锡(SnO_2)	0.2108

在选择出主要化学成分之后，使用主要化学成分分别对高钾玻璃数据和铅钡玻璃数据来进行多次 Kmeans 聚类，分别对高钾玻璃、铅钡玻璃聚为 2 类、3 类、4 类来进行分析。

最终决定将高钾玻璃划分为亚类 1-高钾 A 类玻璃、亚类 2-高钾 B 类玻璃，划分详情如附录，划分方法如图 5.2.4 高钾玻璃亚类划分，高钾 B 类平均二氧化硅含量占比较高，接近 90%，高钾 A 类平均二氧化硅含量占比则较低；将铅钡玻璃划分为亚类 1-铅钡 A 类玻璃、亚类 2-铅钡 B 类玻璃、亚类 3-铅钡 C 类玻璃，划分详情如附录一，划分方法如图 5.2.5 铅钡玻璃亚类划分，铅钡 A 类平均二氧化硅含量占比较高，接近 60%，铅钡 B 类玻璃和铅钡 C 类玻璃的二氧化硅含量接近，但铅钡 C 类玻璃的氯化钡含量占比较高。之所以确定将高钾玻璃分为两类，铅钡玻璃分为三类，是因为分为更多的类，得到的折线趋势比较接近，不容易将不同亚类区分出来，其他分类折线图如附录 1。

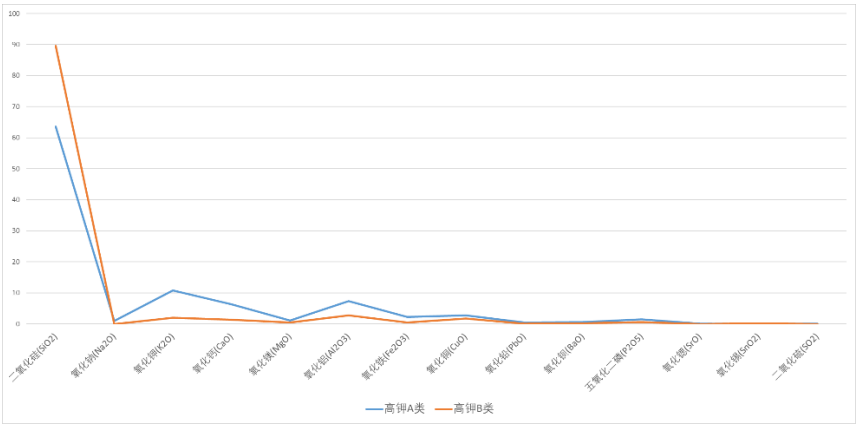


图 5.2.4 高钾玻璃亚类划分

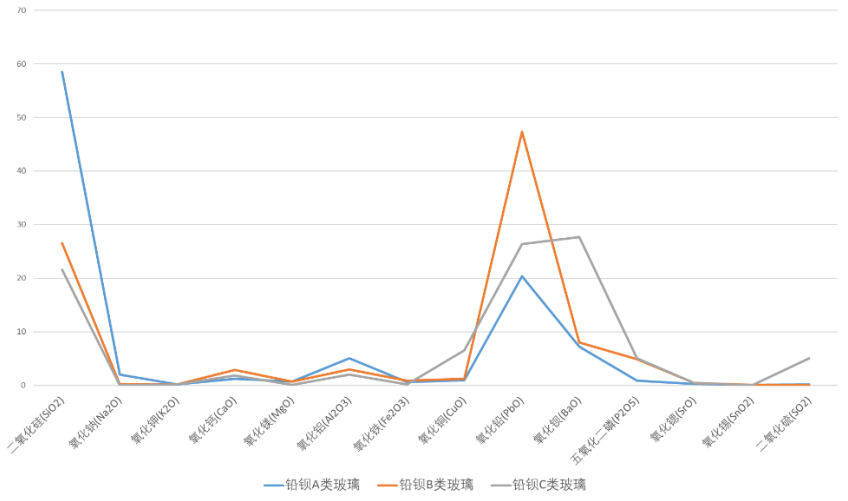


图 5.2.5 铅钡玻璃亚类划分

5.3 问题三模型的建立与求解

针对问题三，我们需要利用附件.xlsx 中的表单 2 的有效数据对未知类型玻璃文物进行类型鉴定。对此，问题三可以定义为一个分类问题，可以采用机器学习方法对其进行建模和求解。但我们也注意到，有效数据集大小仅为 67，一般的机器学习算法和工具都集中在“大数据”和具有大量数据集的场景中，而本题可利用的数据集比较有限，一般的算法处理起来比较棘手，容易出现过拟合问题。

为了解决小数据集容易给机器学习算法带来过拟合的问题，我们决定不使用机器学习算法进行分类，转而建立了一个简单的分类模型，并对原特征进行降维处理，使用较少的特征进行分类，以缓解小数据集带来的过拟合问题。

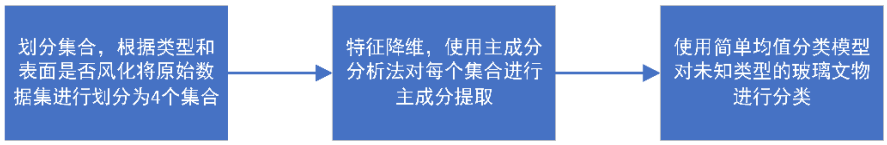


图 5.3.1 问题三求解流程图

5.3.1 划分集合

我们考虑到玻璃文物的类型和表面是否风化对玻璃文物的化学成分影响较大，故在提取主成分前，我们根据类型和表面是否风化将数据分为了高钾-风化、高钾-未风化、铅钡-风化和铅钡-未风化四类，然后利用 PCA 分别对这四类数据集合进行特征提取、降维。

5.3.2 主成分分析特征降维

特征选择对于建立模型尤为重要。尽管正则化之类的方法有助于减少特征，但是如果特征数远远大于样本数，那么过拟合的问题仍然会持续存在。为了缓解小数据集、多特征项带来的过拟合的问题，我们首先通过主成分分析法对数据特征进行降维，将原本的 14 个化学成分（特征）提取为少数几个可以反映出原始数据大部分信息（累积方差解释率超过 80%）的主成分特征，然后以提取出的主成分特征进行下一步的简单均值分类。下表展示了高钾-风化类的 PCA 线性组合系数表，其被提取为了 2 个主成分特征，特征数减少了 12 个。（由于表格所占篇幅较大，我们会在附录 2 中放入其他 3 类的 PCA 线性组合系数表）

表 5.3.2 高钾-风化类的 PCA 线性组合系数表

名称	成分	
	成分 1	成分 2
二氧化硅(SiO2)	-0.475	-0.093
氧化钾(K2O)	-0.386	0.284
氧化钙(CaO)	0.402	0.287
氧化镁(MgO)	0.294	0.406
氧化铝(Al2O3)	0.396	0.342
氧化铁(Fe2O3)	-0.182	0.452
氧化铜(CuO)	0.208	-0.493
五氧化二磷(P2O5)	0.379	-0.316

5.3.3 简单启发式均值分类模型

使用更简单的模型，是因为它们不太容易过拟合，比如正则化线性模型，弹性网络分类器等等。然而线性模型虽然简单，但是在特征维数较多的小数据集上的分类准确率往往不尽人意。

对此，我们一开始想到使用较为简单的 K-NN 模型来进行分类，其原理就是给定一个已知标签类别的训练数据集，输入没有标签的新数据后，在训练数据集中找到与新数据最邻近的 k 个实例，如果这 k 个实例的多数属于某个类别，那么新数据就属于这个类别[2]。

但由于通过类型和表面是否风化对数据进行划分后的每个类别的数据数量不一致，这在一定程度上会影响 K-NN 分类出来结果的准确性。于是，我们借鉴 K-NN 分类的思想，提出了利用每个类型的特征（这里的特征指的是 PCA 提取出的主成分特征）均值进行分类的模型，并在其中加入了一些启发式规则来提升分类的准确率，称之为简单启发式均值分类模型。下图展示了简单启发式均值分类模型的流程图：

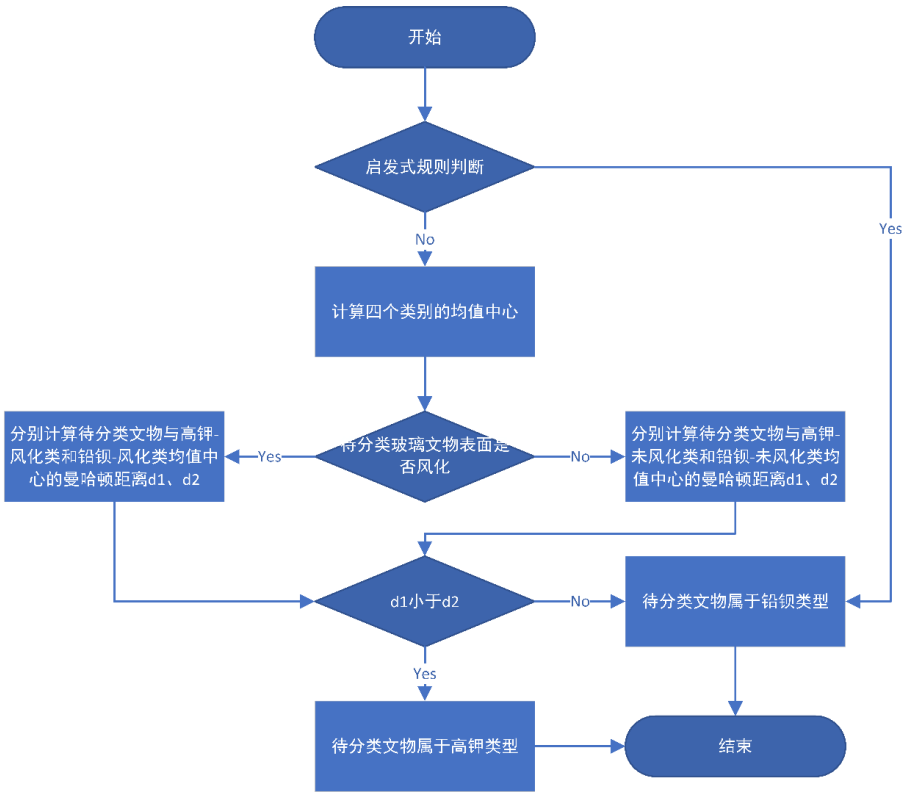


图 5.3.3 简单启发式均值分类模型流程图

使用简单启发式均值分类模型的详细步骤如下：

***Step*0**：启发式规则判断

由问题二的子问题 1 分析可得，铅钡玻璃文物的氧化铅含量较高，且高钾玻璃文物的氧化铅含量极少，故我们有一定理由相信具有较大氧化铅含量的文物属于铅钡文物。设题目所给的有效检测点数据中的铅钡文物检测点的最低氧化铅含量为$MinPbO$，如果待分类玻璃文物的氧化铅含量大于等于$MinPbO$，则我们直接认为其属于铅钡文物，程序结束；否则，进入 Step1。

***Step*1**：求出每个类别的均值中心向量M_i

设R_{ij}是第 i 个类别中的第 j 个数据（向量），则每个类别的均值中心向量的计算公式可以表达如下：

$$M_i = \frac{\sum_{j=1}^{n} R_{ij}}{n} \tag{3}$$

下图展示了二维平面上的均值中心，蓝色点为原始数据，红色点为蓝色点的均值中心：

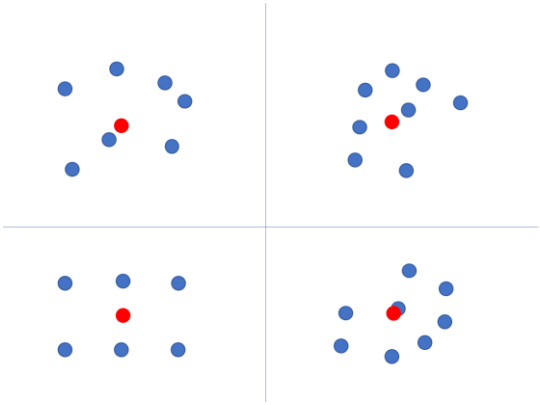


图 5.3.4 二维平面上的均值中心示意图

Step2：分别计算待分类数据与其对应类型数据的均值中心的距离

设待分类数据的主成分特征数据向量为P，为避免欧氏距离给小于 1 的正数据带来额外的缩小，我们采用曼哈顿距离对两个向量进行距离的评估，则待分类数据与对应类型数据的均值中心的距离计算公式如下：

$$d = \sum_{k=1}^{m} |P_k - M_k| \tag{4}$$

Step3：选出与待分类数据距离最小的均值中心所属的类别作为待分类数据的类别，程序结束

例如：假设高钾-风化数据的均值中心为（1，2），高钾-未风化数据的均值中心为（0，4），铅钡-风化的均值中心为（7，3），铅钡-未风化的均值中心为（2，1），待分类玻璃文物的数据向量为（2，3），已知其表面风化，假设其通过了启发式规则判断。

1) 由于已知待分类玻璃文物的表面风化，故其可能的分类属于高钾-风化和铅钡-风化类的其中一个

2) 分别计算待分类玻璃文物与高钾-风化类和铅钡-风化类均值中心的曼哈顿距离为：

$$d_{\text{高钾}-\text{风化}} = |2-1| + |3-2| = 2 \tag{5}$$

$$d_{\text{铅钡}-\text{风化}} = |2-7| + |3-3| = 5 \tag{6}$$

3) 由于$d_{高钾-风化}=d_1=2<d_{铅钡-风化}=d_2=5$，故我们认为待分类文物属于高钾类型

5.3.4 分类结果

下表展示了所有待分类玻璃文物的文物编号、表面风化和使用简单启发式均值分类模型预测的类型信息。

表 5.3.1 待分类玻璃文物的分类结果表

文物编号	表面风化	预测类型	氧化铅含量	D1	D2
A1	无风化	高钾	0.00	**14.45**	49.91
A2	风化	铅钡	34.3	38.36	**21.80**
A3	无风化	铅钡	39.58	41.46	**37.84**
A4	无风化	铅钡	24.28	43.35	**16.13**
A5	风化	铅钡	12.23	**24.62**	48.57
A6	风化	高钾	0	**52.52**	79.24
A7	风化	高钾	0	**48.00**	76.84
A8	无风化	铅钡	21.24	57.73	**27.09**

从上表可以看出，所有待分类玻璃文物均可通过启发式规则判断直接确定其所属类型。如果不考虑启发式规则判断，大部分分类结果不会改变，仅有文物编号为 A5 的待分类玻璃文物会因为 D1<D2 而被判定为高钾类型。

5.3.5 敏感性分析

为了验证简单启发式均值分类模型的敏感性，我们对题目所给的 67 个有效检测点数据进行了类型的预测，并绘制了相应的混淆矩阵（如下图所示），其中 0 代表高钾类型，1 代表铅钡类型，所有有效检测点数据均预测正确，说明我们提出的简单启发式均值分类模型在小数据集上具有较好的分类效果。

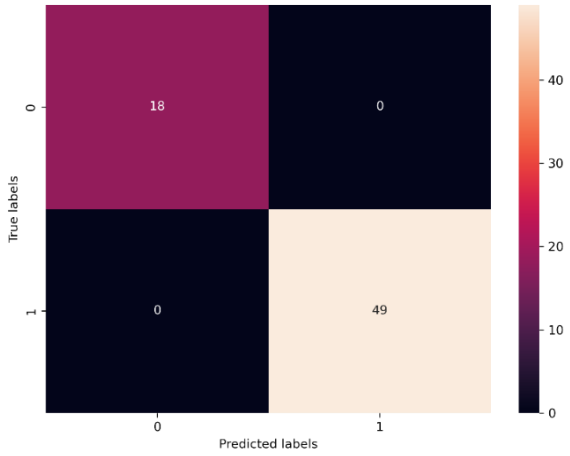


图 5.3.5 原始数据预测-混淆矩阵

5.4 问题四模型的建立与求解

由于表单 2 中的数据量较少，为确保数据分析时维持其可靠性的样本量足够大，故本文只针对高钾玻璃和铅钡玻璃两个大类进行分析，而未在此大类基础上再分小类类进行分析。两种玻璃的化学成分关联与差异分析思路如图 5.4.1 所示。

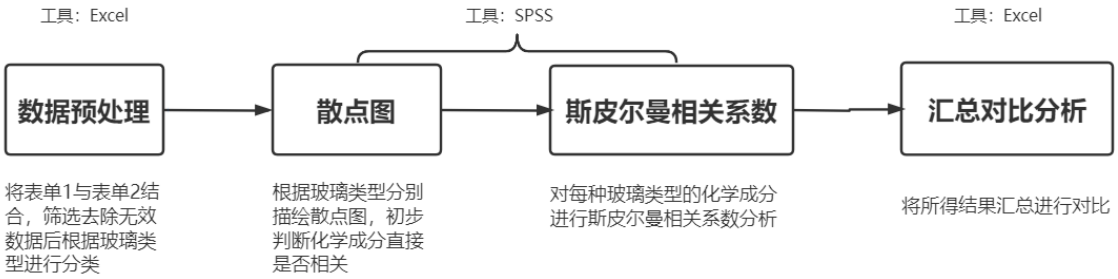


图 5.4.1 问题四分析思路

5.4.1 斯皮尔曼相关系数模型的建立

（1） 数据预处理

目前所给表单数据中存在化学成分比例累加和未满足 85%～105%之间的数据，本文将表单二中未满足条件的文物采样点删除，并将表单 1 和表单 2 合并在一起，确定各文物采样点属于哪种玻璃类型，根据玻璃类型进行筛选分类得到表 5.4.1（表为部分数据，详情请见支撑材料）。

表 5.4.1 文物化学成分含量表（部分）

类型	文物采样点	二氧化硅	氧化钠	氧化钾	氧化钙	氧化镁	…

		(SiO_2)	(Na_2O)	(K_2O)	(CaO)	(MgO)	
高钾	01	69.33	0	9.99	6.32	0.87	…
高钾	03 部位 1	87.05	0	5.19	2.01	0	…
高钾	03 部位 2	61.71	0	12.37	5.87	1.11	…
高钾	04	65.88	0	9.67	7.12	1.56	…
高钾	05	61.58	0	10.95	7.35	1.77	…
高钾	06 部位 1	67.65	0	7.37	0	1.98	…
…	…	…	…	…	…	…	…

（2） 绘制散点图

得到高钾玻璃与铅钡玻璃两类的各化学成分含量之后，运用 Python 的 Seaborn 库绘制每种玻璃各化学成分之间的散点图矩阵，用以初步判断化学成分之间是否相关联。（铅钡玻璃化学成分散点图见附录 6）

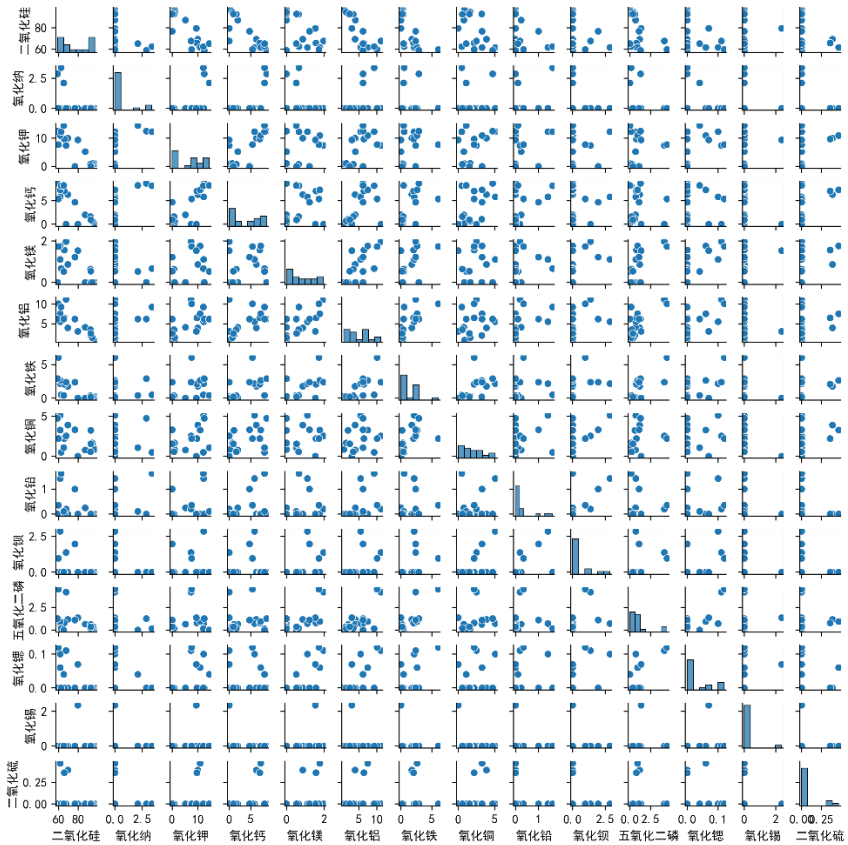


图 5.4.2 高钾玻璃化学成分散点图

（3） 建立斯皮尔曼相关系数模型

根据表 5.4.1 所得的文物化学成分含量，我们对每种玻璃的所有化学成分，进行了正态分布假设性检验，发现各个指标服从正态分布的效果并不理想，对此，本文建立斯皮尔曼相关系数模型进行化学成分间的相关性分析。

●确定等级序列

一个数的等级，就是将它所在的一列按照从小到大排序后，这个数所在的位置。例如表 5.4.2 所示。

表 5.4.2 等级序列示例表

类型	二氧化硅(SiO_2)	氧化钾(K_2O)	二氧化硅等级	氧化钾等级
高钾	69.33	9.99	3	3
高钾	87.05	5.19	4	1
高钾	61.71	12.37	1	4
高钾	65.88	9.67	2	2

●根据等级序列得出等级差的平方d_i^2

表 5.4.3 等级差计算示例表

类型	二氧化硅(SiO_2)	氧化钾(K_2O)	二氧化硅等级	氧化钾等级	等级差d_i^2
高钾	69.33	9.99	3	3	0
高钾	87.05	5.19	4	1	9
高钾	61.71	12.37	1	4	9
高钾	65.88	9.67	2	2	0

●每一类玻璃文物中的每种化学成分两两对比得出斯皮尔曼相关系数r_s

$$r_s = 1 - 6\sum_{i=1}^{n} d_i^2 \Big/ n(n^2-1) \tag{7}$$

其中，d_i为每一类玻璃文物中的化学成分两两之间的等级差，n为每种化学成分的含量样本数。

●斯皮尔曼假设检验

由于每种化学成分中的样本数量均小于 30，故可依据样本数和显著性水平进行查表检验，若计算得出的相关系数r_s大于表中值则证明该值有效。

5.4.2 化学成分关系分析

依据所整理的数据，运用 SPSS 求解斯皮尔曼相关系数（见支撑材料），再进行归纳总结得出两种玻璃类型各化学成分之间的相关系数热力图。其中部分数据因其所得显著性（双尾）大于 0.05，即根据显著性检验，该部分数据不存在显著的相关关系，用 0 表示。

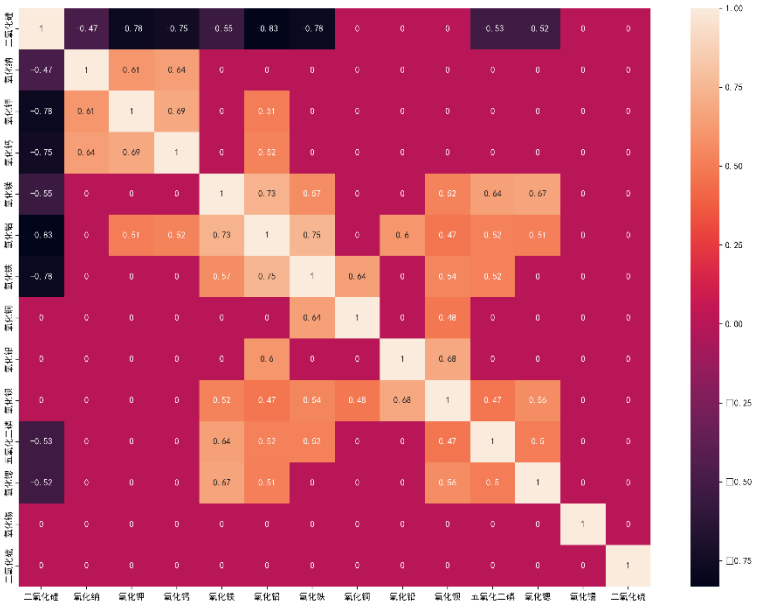


图 5.4.3 高钾玻璃化学成分相关系数热力图

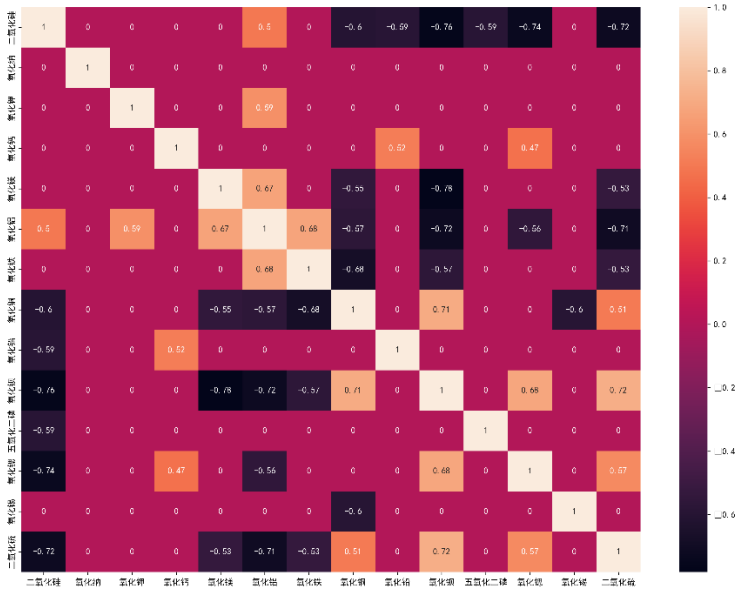


图 5.4.4 铅钡玻璃化学成分相关系数热力图

（1） 高钾玻璃文物样品化学成分之间关联关系分析

结合各化学成分的散点图和求解所得的斯皮尔曼相关系数显著性检验的p值可知二氧化硅与氧化铜、氧化铅等无关联，氧化钠与氧化镁、氧化铝等无关联等等。

从以上图表分析中我们可以总结出该种玻璃化学成分关联的规律：二氧化硅与其他化学成分呈现负相关或者无关的关系，氧化锡和二氧化硫均与其他化学成分无关联，其余化学成分均与其他化学成分呈现正相关的关系。

（2） 铅钡玻璃文物样品化学成分之间关联关系分析

结合各化学成分的散点图和求解所得的斯皮尔曼相关系数显著性检验的p值可知二氧化硅与氧化钠、氧化钾等无关联，氧化钠与氧化钙、氧化镁等无关联等等。

从以上图表分析中我们可以总结出该种玻璃化学成分关联的规律：氧化钾、氧化钙、五氧化二磷和氧化锡均与其他化学成分呈现正相关或者无关的关系，氧化钠与其他化学成分无关联，其余化学成分均与其他化学成分呈现正负相关和无关。

（3） 高钾玻璃与铅钡玻璃之间的化学成分关联关系差异性

由两种玻璃化学成分的相关系数表可以看出，两种玻璃之间的化学成分关联关系差异性较大。例如，高钾玻璃中二氧化硫与其他化学成分无关联，而铅钡玻璃中的二氧化硫与大多数其他化学成分呈正或负相关；高钾玻璃中除了二氧化硅，其余化学成分呈正相关或者无关联，而铅钡玻璃中正相关与负相关出现频次基本一致等等。

六、 模型的评价、改进与推广

6.1 问题一模型

6.1.1 模型优点

1. 卡方检验模型适用于该问题求解，能够较好地判断两个定性变量之间的关系
2. 均值差预估公式提取总体特征，再运用至每个个体中可得到较为准确的数据

6.1.2 模型缺点

1. 数据量少，对信息的挖掘不够深入

6.2 问题二模型

6.2.1 模型优点

1. 根据标准差筛选出主要的化学成分，可以提升亚类划分的可解释性
2. 采用 Kmeans 聚类算法进行聚类，原理接单，聚类效果中上
3. Kmeans 算法的时间和空间复杂度较低，求解效率高

6.2.2 模型缺点

1. Kmeans 聚类算法需要提供簇心数 k，聚类个数不可动态改变，需要多次调用 Kmeans 算法才能求得不同簇心数下亚类划分的效果
2. Kmeans 算法对离群点非常敏感，且结果不一定是全局最优

6.2.3 模型改进

1. 可以采用 ISODATA 聚类算法，该算法的簇心数会随着聚类过程而动态改变（合并、分裂），该算法可迭代出合适的簇心数及其坐标，不需要多次调用

6.3 问题三模型

6.3.1 模型优点

1. 根据玻璃文物的类型和表面风化划分数据集，去除其对化学成分的影响，提升分类结果的准确性
2. 使用主成分分析对特征进行提取，使用数量较少的主成分进行分类，达到特

征降维的作用，有利于缓解小数据集下的过拟合问题

3. 采用简单的启发式均值分类模型，启发式规则可以有效提升分类准确率，简单的均值分类可以缓解过拟合问题

6.3.2 模型缺点

1. 只适用于求解本题背景下的小数据集的分类问题
2. 基于均值的分类，容易受到异常值的影响

6.3.3 模型改进

1. 在数据集足够大的情况下，可以转而采用 XGboost、随机森林等机器学习模型，甚至可以采用多层感知机、CNN、RNN 等深度学习模型，充分利用数据，提高分类准确率

6.4 问题四模型

6.4.1 模型优点

1. 使用数形结合的方法进行分析更加合理
2. 合理运用斯皮尔曼相关系数可精确得出两两之间的关系

6.4.2 模型缺点

1. 无法剔除不合理数据

七、 参考文献

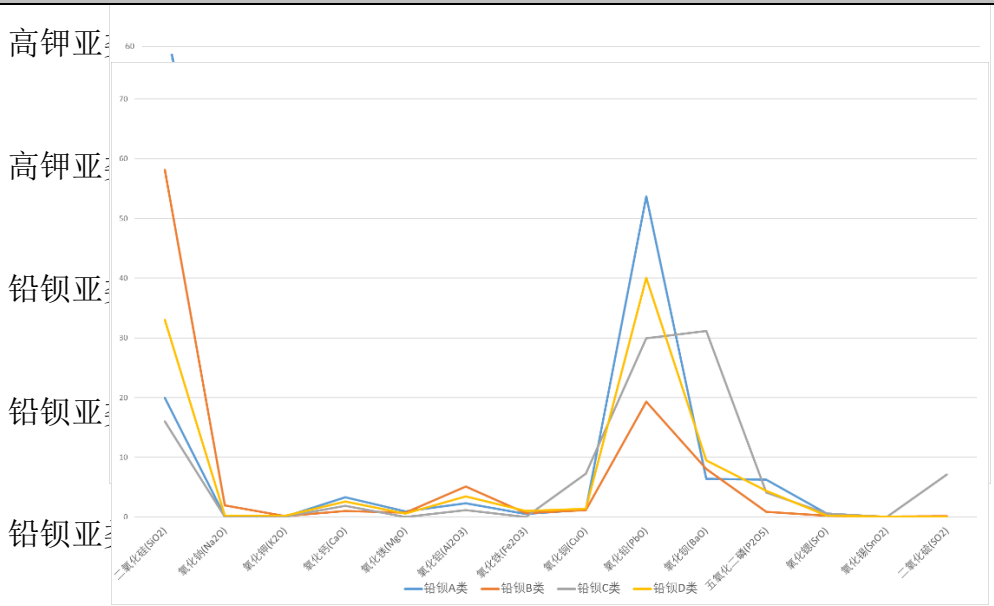
[1] 王金桃, 周利锋, 高尔生. 第六讲 卡方检验[J]. 实验动物与比较医学, 2000(4):251-254.

[2] 耿丽娟,李星毅.用于大数据分类的 KNN 算法研究[J].计算机应用研究,2014,31(05):1342-1344+1373.

附录

附录 1

介绍：玻璃文物亚类划分详情



附录 2

介绍：PCA 线性组合系数表

高钾-未风化类的 PCA 线性组合系数表

名称	成分			
	成分 1	成分 2	成分 3	成分 4
二氧化硅(SiO2)	-0.154	-0.476	0.101	-0.185
氧化钠(Na2O)	-0.260	0.315	0.229	0.343
氧化钾(K2O)	-0.155	0.300	-0.149	0.380
氧化钙(CaO)	-0.167	0.476	-0.088	-0.080
氧化镁(MgO)	0.379	-0.108	-0.225	0.094

氧化铝(Al2O3)	0.311	0.229	0.107	0.288
氧化铁(Fe2O3)	0.386	0.215	-0.120	-0.071
氧化铜(CuO)	0.158	0.277	-0.059	-0.468
氧化铅(PbO)	0.023	0.139	0.566	-0.101
氧化钡(BaO)	0.308	0.014	0.392	-0.313
五氧化二磷(P2O5)	0.429	-0.070	-0.055	0.188
氧化锶(SrO)	0.398	-0.051	0.029	0.282
氧化锡(SnO2)	-0.064	-0.368	-0.064	0.310
二氧化硫(SO2)	-0.032	0.092	-0.588	-0.249

铅钡-风化类的 PCA 线性组合系数表

名称	成分					
	成分 1	成分 2	成分 3	成分 4	成分 5	成分 6
二氧化硅(SiO2)	0.441	-0.217	-0.038	-0.051	-0.085	0.026
氧化钠(Na2O)	0.202	-0.343	-0.086	-0.173	-0.117	0.500
氧化钾(K2O)	0.210	0.097	0.247	-0.230	0.735	0.045
氧化钙(CaO)	-0.057	0.438	0.242	-0.155	-0.227	-0.098
氧化镁(MgO)	0.260	0.336	0.133	-0.114	-0.072	0.413
氧化铝(Al2O3)	0.396	0.076	0.258	0.234	-0.026	0.052
氧化铁(Fe2O3)	0.191	0.367	0.102	-0.281	-0.028	-0.218
氧化铜(CuO)	-0.262	-0.184	0.295	-0.230	-0.358	0.032

氧化铅(PbO)	-0.275	0.262	-0.443	0.128	0.228	-0.134
氧化钡(BaO)	-0.323	-0.266	0.416	-0.011	0.011	-0.055
五氧化二磷(P2O5)	-0.175	0.400	0.113	-0.120	-0.246	0.203
氧化锶(SrO)	-0.253	0.186	-0.073	0.371	0.104	0.639
氧化锡(SnO2)	0.202	0.067	0.287	0.720	-0.136	-0.190
二氧化硫(SO2)	-0.270	-0.086	0.469	0.052	0.332	0.091

铅钡-未风化类的 PCA 线性组合系数表

名称	成分				
	成分 1	成分 2	成分 3	成分 4	成分 5
二氧化硅(SiO2)	-0.357	-0.371	0.134	-0.019	0.062
氧化钠(Na2O)	-0.069	-0.071	-0.326	0.411	0.016
氧化钾(K2O)	0.304	0.073	0.222	0.154	0.396
氧化钙(CaO)	0.459	-0.146	-0.025	-0.224	0.069
氧化镁(MgO)	0.165	-0.269	0.082	0.504	0.066
氧化铝(Al2O3)	0.218	0.080	0.382	0.489	0.160
氧化铁(Fe2O3)	0.263	-0.088	0.325	-0.200	-0.402
氧化铜(CuO)	-0.072	0.527	-0.112	-0.057	-0.053
氧化铅(PbO)	0.371	-0.086	-0.390	-0.157	-0.016
氧化钡(BaO)	0.080	0.547	-0.052	0.111	-0.060
五氧化二磷(P2O5)	0.004	0.328	0.477	-0.046	-0.128

氧化锶(SrO)	0.173	0.190	-0.413	0.226	0.024
氧化锡(SnO2)	0.476	-0.081	0.005	-0.190	0.064
二氧化硫(SO2)	-0.122	0.098	0.039	-0.307	0.785

附录 3

介绍：斯皮尔曼相关系数热力图和化学成分散点图绘制（Python 语言实现）

```
import matplotlib.pyplot as plt
import seaborn as sns
# 设置兼容中文
plt.rcParams['font.family'] = ['sans-serif']
plt.rcParams['font.sans-serif'] = ['SimHei']

x_ticks = ['二氧化硅', '氧化纳', '氧化钾','氧化钙','氧化镁','氧化铝','氧化铁','氧化铜','氧化铅','氧化钡','五氧化二磷','氧化锶','氧化锡','二氧化硫']
y_ticks = x_ticks  # 自定义横纵轴
confusion_matrix_result = [[i for i in range(len(x_ticks))] for _ in range(len(x_ticks))]
lst = [
    [1,-0.475,-0.781,-0.752,-0.546,-0.831,-0.783,0,0,0,-0.53,-0.524,0,0],
    [1,0.606,0.64,0,0,0,0,0,0,0,0,0,0],
    [1,0.69,0,0.506,0,0,0,0,0,0,0,0],
    [1,0,0.521,0,0,0,0,0,0,0,0],
    [1,0.735,0.569,0,0,0.521,0.643,0.666,0,0],
    [1,0.748,0,0.602,0.473,0.518,0.511,0,0],
    [1,0.643,0,0.539,0.517,0,0,0],
    [1,0,0.483,0,0,0,0],
    [1,0.682,0,0,0,0],
    [1,0.471,0.565,0,0],
    [1,0.5,0,0],
    [1,0,0],
    [1,0],
    [1],
]
```

```
    for i in range(len(x_ticks)):
        j = i
        while j < len(x_ticks):
            confusion_matrix_result[i][j] = lst[i][j-i]
            confusion_matrix_result[j][i] = confusion_matrix_result[i][j
]
            j += 1

    # 利用热力图对于结果进行可视化
    plt.figure(figsize=(16, 12))
    sns.heatmap(confusion_matrix_result, annot=True,xticklabels=x_ticks,
 yticklabels=y_ticks)
    plt.savefig(r'C:/Users/xxxx/Desktop/ee.png',dpi=300)
    plt.show()

    x_ticks = ['二氧化硅', '氧化纳', '氧化钾','氧化钙','氧化镁','氧化铝','氧
化铁','氧化铜','氧化铅','氧化钡','五氧化二磷','氧化锶','氧化锡','二氧化硫']
    y_ticks = x_ticks  # 自定义横纵轴
    confusion_matrix_result = [[i for i in range(len(x_ticks))] for _ in
 range(len(x_ticks))]
    lst = [
        [1,0,0,0,0,0.498,0,-0.604,-0.589,-0.761,-0.593,-0.737,0,-0.725],
        [1,0,0,0,0,0,0,0,0,0,0,0,0],
        [1,0,0,0.591,0,0,0,0,0,0,0,0],
        [1,0,0,0,0,0.519,0,0,0.469,0,0],
        [1,0.669,0,-0.545,0,-0.781,0,0,0,-0.534],
        [1,0.676,-0.566,0,-0.721,0,-0.555,0,-0.708],
        [1,-0.682,0,-0.57,0,0,0,-0.534],
        [1,0,0.706,0,0,-0.599,0.506],
        [1,0,0,0,0,0],
        [1,0,0.682,0,0.717],
        [1,0,0,0],
        [1,0,0.568],
        [1,0],
        [1]
    ]
    for i in range(len(x_ticks)):
        j = i
        while j < len(x_ticks):
            confusion_matrix_result[i][j] = lst[i][j-i]
            confusion_matrix_result[j][i] = confusion_matrix_result[i][j
]
            j += 1
```

```
64.
65.    # 利用热力图对于结果进行可视化
66.    plt.figure(figsize=(16, 12))
67.    sns.heatmap(confusion_matrix_result, annot=True,xticklabels=x_ticks,
 yticklabels=y_ticks)
68.    plt.savefig(r'C:/Users/xxxx/Desktop/ee.png',dpi=300)
69.    plt.show()
70.
71.    import numpy as np
72.    # 设置兼容中文
73.    plt.rcParams['font.family'] = ['sans-serif']
74.    plt.rcParams['font.sans-serif'] = ['SimHei']
75.
76.    x_ticks = ['二氧化硅', '氧化纳', '氧化钾','氧化钙','氧化镁','氧化铝','氧
化铁','氧化铜','氧化铅','氧化钡','五氧化二磷','氧化锶','氧化锡','二氧化硫']
77.    y_ticks = x_ticks  # 自定义横纵轴
78.    data = pd.read_excel(
79.        r'xxxxx\data\data.xlsx',
80.        sheet_name="高钾")
81.    x = []
82.    columns = data.columns
83.    for i in range(data.shape[0]):
84.        lst = []
85.        for column in columns[1:15]:
86.            lst.append(data.iloc[i][column])
87.        x.append(lst)
88.    pd_data = pd.DataFrame(np.array(x))
89.    pd_data.columns = x_ticks
90.    sns.pairplot(pd_data,height = 0.7)
91.    plt.savefig(r'C:/Users/xxxxxxxx/Desktop/ee.png',dpi=300)
92.
93.    x_ticks = ['二氧化硅', '氧化纳', '氧化钾','氧化钙','氧化镁','氧化铝','氧
化铁','氧化铜','氧化铅','氧化钡','五氧化二磷','氧化锶','氧化锡','二氧化硫']
94.    y_ticks = x_ticks  # 自定义横纵轴
95.    data = pd.read_excel(
96.        r'xxxxx\data\data.xlsx',
97.        sheet_name="铅钡")
98.    x = []
99.    columns = data.columns
100.   for i in range(data.shape[0]):
101.       lst = []
102.       for column in columns[1:15]:
103.           lst.append(data.iloc[i][column])
```

```
        x.append(lst)
    pd_data = pd.DataFrame(np.array(x))
    pd_data.columns = x_ticks
    sns.pairplot(pd_data,height = 0.7)
    plt.savefig(r'C:/Users/xxxx/Desktop/ee.png',dpi=300)
```

附录 4

介绍：Sklearn-Kmeans 聚类模型（Python 语言实现）

```
from sklearn.cluster import KMeans
import pandas as pd
data = pd.read_excel(
    r'xxxx\data\data.xlsx',
    sheet_name="高钾")
x = []
columns = data.columns
for i in range(data.shape[0]):
    lst = []
    for column in columns[1:15]:
        if column == "二氧化硅(SiO2)" or column == "氧化钾
(K2O)" or column == "氧化钙(CaO)" or column == "氧化铝
(Al2O3)" or column == "氧化铁(Fe2O3)" or column == "氧化铜
(CuO)" or column == "五氧化二磷(P2O5)" or column == "氧化钠(Na2O)":
            lst.append(data.iloc[i][column])
    x.append(lst)
# print(x)
y_pred = KMeans(n_clusters=2, random_state=520).fit_predict(x)
print(y_pred)
print(len(y_pred))

data = pd.read_excel(
    r'xxxxx\data\data.xlsx',
    sheet_name="铅钡")
x = []
columns = data.columns
for i in range(data.shape[0]):
    lst = []
    for column in columns[1:15]:
        if column == "二氧化硅(SiO2)" or column == "氧化铅
(PbO)" or column == "氧化钡(BaO)" or column == "五氧化二磷
(P2O5)" or column == "二氧化硫(SO2)" or column == "氧化铝
```

```
(Al2O3)" or column == "氧化铜(CuO)" or column == "氧化钠
(Na2O)" or column == "氧化钙(CaO)":
                lst.append(data.iloc[i][column])
        x.append(lst)
    # print(x)
    y_pred = KMeans(n_clusters=2, random_state=520).fit_predict(x)
    print(y_pred)
    print(len(y_pred))
```

附录 5

介绍：数据处理（Python 语言实现）

```
import pandas as pd
data = pd.read_excel(
    r'xxxx\src\data\xxxx.xlsx',
    sheet_name="表单 1")

dic1,dic2,dic3,dic4 = {},{},{},{}
for i in range(data.shape[0]):
    if data.iloc[i]['纹饰'] not in dic1:
        dic1[data.iloc[i]['纹饰']] = len(dic1)
    data.loc[i,'纹饰'] = dic1[data.iloc[i]['纹饰']]

    if data.iloc[i]['类型'] not in dic2:
        dic2[data.iloc[i]['类型']] = len(dic2)
    data.loc[i,'类型'] = dic2[data.iloc[i]['类型']]

    if data.iloc[i]['颜色'] not in dic3:
        dic3[data.iloc[i]['颜色']] = len(dic3)
    data.loc[i,'颜色'] = dic3[data.iloc[i]['颜色']]

    if data.iloc[i]['表面风化'] not in dic4:
        dic4[data.iloc[i]['表面风化']] = len(dic4)
    data.loc[i,'表面风化'] = dic4[data.iloc[i]['表面风化']]
print(dic1)
print(dic2)
print(dic3)
print(dic4)
print(data)
```

```
28.    data.to_excel(r'xxxxx\data\export.xlsx')
```

附录 6

介绍：铅钡玻璃化学成分散点图