**摘要**

古代玻璃极易受到埋藏在土里的环境影响，玻璃内部的元素与其所处环境的元素进行大量交换从而使得外表产生风化，从而影响对其类别的正确判断，误导研究结论。本文主要是针对古代不同类型玻璃样品表面风化情况以及其主要化学成分含量关系建立模型，分析该玻璃样品风化情况与其他因素之间的关系，并对相关问题进行更深入的研究。研究古代玻璃制品的外观形特征、整体化学分析等方面对考古学家历史考证有着重要的意义。

首先对于问题一，文中运用了Apriori关联规则模型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相关性分析，设置最小支持度、最小置信度找出频繁项集从而筛选出强关联规则，得到B类纹饰、铅钡玻璃对样品表面风化有推进作用。针对探究文物样品表面有无风化化学成分含量的统计规律，我们采用独立样本T检验模型，同一种类型的玻璃文物的各种化学成分风化前后的含量的差异性不同，并且同一个化学成分在不同的玻璃类型中会存在不同的差异性。同时运用均值比较的方法预测出该检测点在风化前的化学成分含量。

对于问题二，我们对数据进行预处理，经过数据清洗和数据归约，得到规范化后的数据，在此分别用：基于划分的聚类（K-means）和基于密度的聚类（DBSCAN）两种模型分别对高钾玻璃和铅钡玻璃进行亚类划分，用轮廓系数评价方法、Calinski-Harabasz评价方法和inertia指标对该模型结果综合评价，选取较好的方法得到高钾玻璃可以再划分成两个细类、铅钡玻璃可以再划分成三个细类。

对于问题三，根据未知分类的玻璃样品的化学成分分析鉴定其类型，这里建立全连接神经网络（Fully Connected Neural Network）预测模型，并以预测结果与实际结果的均方差（Mean Square Error, MSE）作为评估指标，鉴别出A1~A8八个玻璃样品分别属于类型是的高钾、铅钡、铅钡、铅钡、铅钡、高钾、高钾、铅钡。

对于问题四，探究不同类别样品不同成分之间的关联，由SPSS相关分析皮尔逊系数矩阵得出以下结论，二氧化硅SiO2与氧化钙CaO、氧化铝Al2O3、氧化铅PbO、氧化钡BaO、五氧化二磷P2O5、氧化铯SrO、二氧化硫SO2关系非常显著，与氧化纳Na2O、氧化铜CuO关系一般显著，与其余化学成分关系不显著。

**关键词：**古代玻璃文物、风化、玻璃类型、Apriori算法、独立样本T检验、K-means算法、DBSCAN算法、全连接神经网络、皮尔逊相关系数矩阵

1. **问题背景**

丝绸之路是亚洲与欧洲进行经济、政治、文化交流的一条“大动脉”，在古代中外交流中起了很重要的作用。玻璃是早期在古代丝绸之路的贸易往来的宝贵物证，古代玻璃物品和技术的交流和传播上的作用是十分有意义的。早期的玻璃在西亚和埃及地区常被制作成珠形饰品传入我国，对我国古代玻璃制品产生和制造技术的有着深刻的影响，尽管与外来的玻璃制品外观相似，但化学成分却不相同。因此从中国古代玻璃的发展中也可看到东、西文化和贸易的交流，亦可推进考古学家们考古研究之路。

玻璃的材料是非常简单、非常可得的，主要材料就是石英砂（SiO2），纯净的沙子、石英石都是最主要的原料。但如果要把这种石英砂熔融成液态，需要1700摄氏度的高温，达到这个温度就很难，但如果人工加一些材料进去作为助熔剂，就能把玻璃的熔点降低，比如加进自然泡碱（Na2CO3）、草木灰（K2CO3）、硝石或铅矿石等，都可以把二氧化硅的熔点往下降，尤其像加入铅矿石之后，都可以降到1000摄氏度以下，就可以把石英砂熔成液态。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）。

当烧制过程中，助熔剂为铅矿石作时产生的是铅钡玻璃；助熔剂为铅矿石作时产生的是钾玻璃。因此加入不同的助熔剂从而会导致产生的玻璃制品的主要成分也不同。

研究古代玻璃制品的外观形特征、整体化学分析等方面对考古学家历史考证有着重要的意义，玻璃内部各个元素含量、化学成分组成是判断产地来源及制作工艺的重要依据，反映了诸多考古信息，但古代玻璃样品通常因为埋藏在土里的环境而受到影响，玻璃内部的元素与其所处环境的元素进行大量交换从而使得外表产生风化，风化后的玻璃其成分比例发生变化使得文物的颜色、纹饰发生很大的变化，从而影响对其类别的正确判断，错误的判断结果必然误导研究结论，加上古代玻璃样品稀少、珍贵，使得研究工作受到限制。

1. **问题重述与分析**

针对问题一：

问题一包含了两个小问。

（1）根据附件表单1看出，这一批我国古代玻璃制品的相关数据：玻璃的类型有两种，分别是高钾玻璃、铅钡玻璃；玻璃的纹饰有三种，分别是A、B、C；玻璃的颜色有八种，分别是：蓝绿色、浅蓝色、深蓝色、浅绿色、绿色、深绿色、紫色、黑色。

因变量是文物是否有风化，自变量有三个分别是其玻璃类型、纹饰和颜色。此题目是找出文物表面风化分别与其三个自变量的关系，对古代玻璃文物的玻璃类型、纹饰、颜色与表面风化的关系进行关联性分析，因此可以用Apriori关联规则模型进行相关性分析。

首先设置最小支持度、最小置信度，找到最小支持度的频繁项集，生成需要的关联规则，利用最小置信度筛选出强关联规则，从而得出表面风化与古代玻璃文物属性的关系。

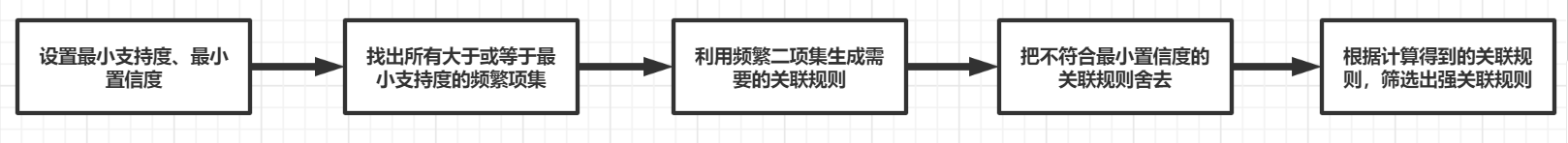


图1-2-1 Apriori关联规则模型简单阐述解题步骤

1. 根据附件表单2，这批文物样品的主要化学成分有二氧化硅(SiO2)、氧化钠(Na2O)、氧化钾(K2O)等十四种，结合玻璃的类型，分析文物样品表面有无风化化学成分含量的统计规律，此类题型分析定类数据与定量数据之间的关系，两组数据之间一组为分组资料（样品表面有风化、样品表面无风化）、一组为定量资料（各个化学成分的含量），为故可以运用独立样本T检验的方法，即可以分析出样品表面有无风化与化学成分含量的相关性，由此可以得到文物样品表面有无风化化学成分含量的统计规律。

在SPSS软件中，先建立数据文件，该模型应该有两个变量，一个是组别变量，另一个变量分别为各化学成分含量，进行正态性检验，若满足正态性，则可以使用T检验；若不满足正态性，则使用非参数的统计检验。

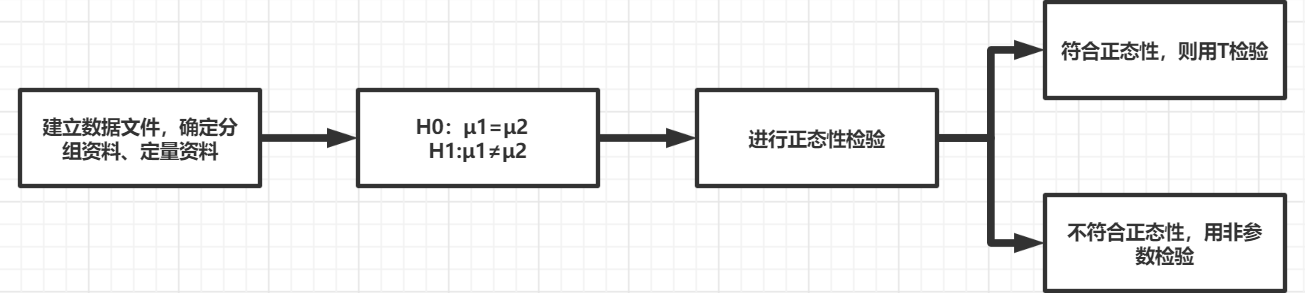


图1-2-2 独立样本T模型简单阐述解题步骤

并且根据已风化的文物样品风化点的各个化学成分的含量的检测数据，预测该样品在风化前的化学成分含量，在此问中运用的方法模型是均值比较。首先将附件表单2中的文物样品信息进行筛选，将无风化的样品筛选出去，计算有风化组的各个化学成分含量的平均值，可以认为该数据为已风化组风化前的各个化学成分的含量，此方法可以预测其风化前的化学成分含量。

针对问题二：

该问题是先分析高钾玻璃、铅钡玻璃的分类规律，选择合适的化学成分将这两类玻璃进行亚类划分，即凭合适的划分方法将高钾玻璃再细分几类，铅钡玻璃也是如此，但是这两类划分的细类中可能会含有包含关系，并不是完全独立的，是关于聚类的问题。为了使得属于同一个类的样品间的距离尽可能小，使不同类的样品间距离尽可能打，我们在此分别用：基于划分的聚类（K-means）和基于密度的聚类（DBSCAN）两种模型分别对高钾玻璃和铅钡玻璃进行亚类划分，最后取得指标较好得方法。

首先用使用PCA主成分分析方法将十四个化学成分含量进行降维，对聚类前降维好的数据进行立体可视化，看其分布情况，对数据集进行聚类模型训练，聚类评价指标分别使用轮廓系数评价方法、Calinski-Harabasz评价方法和inertia指标对该模型的聚类结果进行评价，根据指标选择最优聚类数目，最后再一次对最优聚类数目的聚类结果进行可视化，得到最优的划分方法，使得评价聚类性能准则达到最优。

针对问题三：

在附件表单3中，有A1~A8八件未知类别玻璃文物样品的信息（包括样品表面风化情况、以及各个化学成分的含量），现在要对该八件样品的风化情况对应的化学成分含量进行分析来鉴别其类型。常用的分类与预测算法有回归分析、决策树、人工神经网络、支持向量机等，由于维度有14维，关系较为复杂，所以本问采用深度学习中的人工神经网络，建立全连接神经网络（Fully Connected Neural Network）预测模型，并以预测结果与实际结果的均方差（Mean Square Error, MSE）作为评估指标。

针对问题四：

不同类别的玻璃文物样品，其化学成分可能具有一定的关联或者差异，所以此问题是看不同类别玻璃文物样品间化学成分之间的关联关系，同时也要比较不同类别类别之间化学成分关联关系的差异性。该问题建立是典型的相关性模型，所以可以选择建立皮尔逊积距相关系数矩阵来表示不同化学成分之间的相关系数。运用此模型可以得到不同类别的皮尔逊积距相关系数矩阵，其中各个元素值可以初步判断哪些化学成分是相关的，哪些化学成分是相异的，哪些化学成分是相斥的。

1. **模型假设**
2. 独立样本T检验假设：

原假设 (H0)：两个样本的均值没有显著差异： H0：μ1=μ2

备择假设 (H1)：两个样本的均值有显著差异： H1:μ1≠μ2

1. 假设不发生小概率事件
2. **数据处理**

**4.1数据分析**

根据该题目给出的附件数据表单可以看出：

1. 附件表单1：分别给出了58件玻璃文物样品的纹饰、玻璃类型、颜色、和表面有无风化的分类信息。
2. 附件表单2：分别给出了对已分类的58件玻璃文物样品随机采样的检测点相应的主要化学成分及其所占比例，但该表中有效数据仅是成分比例累加和介于 85%~105%之间的数据，因此要筛选出不合符这个条件的数据去除再进行数据分析建模，该表中样品检测点未检测出相应化学成分的值为空值，因此要对该表进行异常值处理、缺失值处理。
3. 附件表单3：分别给出了八件未知类别的玻璃文物样品对应的表面风化情况以及各相应化学成分含量的信息。该表的有效数据、数据缺失情况与表单2相同。

**4.2数据预处理**

运用Jupyter Notebook软件，对表单进行数据清洗与数据归约。

1. 缺失值处理

表单中的空值是由于再文物表面检测点中没有检测到该化学成分，所以可以看作是在该检测点中无该成分的含量，故运用添加默认值0的方法消除缺失数据。

代码如下：



图4-2-1 缺失值处理代码

1. 剔除无效数据

成分比例累加和介于 85%~105%之间的数据视为有效数据。

代码图如下：



图4-2-1 缺失值处理代码

原表单有67条数据，剔除无效数据后的表单有67条数据，剔除了文物样品8和文物样品11的数据。

1. 数据降维

运用PCA主成分分析方法将表单的14列化学成分含量降成三维数据。



图4-2-3 对高钾玻璃数据进行PCA降维代码

降维后的数据情况：

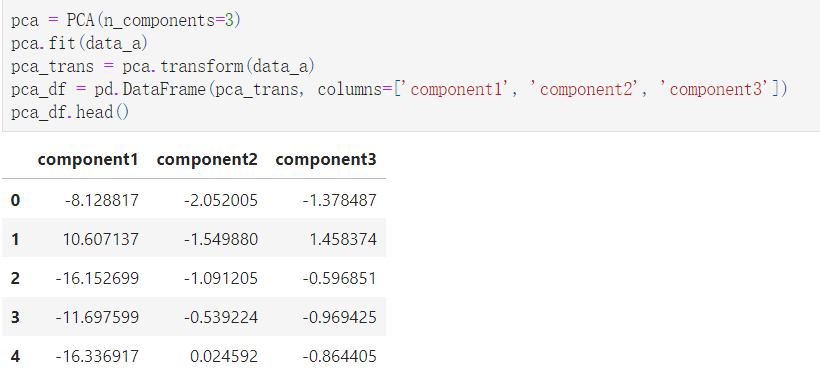


图4-2-4 对高钾玻璃数据进行PCA降维后的数据情况

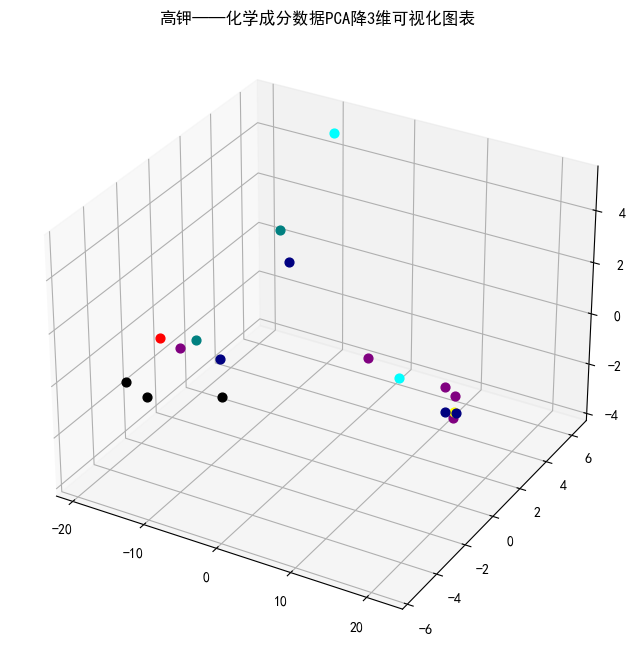


图4-2-4 对高钾玻璃数据进行PCA降维后可视化



图4-2-5 对铅钡玻璃数据进行PCA降维后可视化

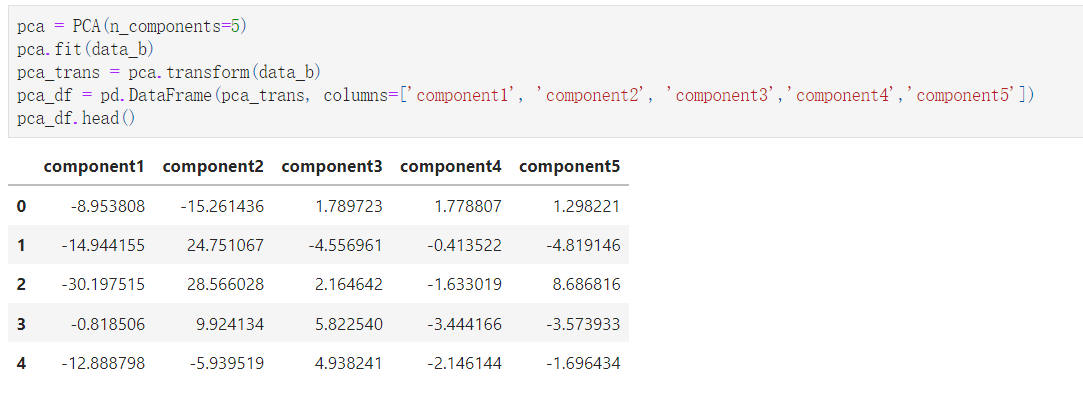


图4-2-6 对铅钡玻璃数据进行PCA降维后的数据

**五、问题一的求解**

**5.1** **问题分析**

随着时间的推移，由于古代玻璃很容易受到埋藏坏境的影响而风化，在风化的过程中，内部元素的成分比例会因为环境发生一定的改变，从而会影响人们鉴定其类别。并出现了古代玻璃的哪些属性与表面风化有关，风化前后其化学成分变化了多少等一系列问题。目前人们对于如何解决这些问题存在着一定的疑惑。也没有合适的方法进行判断。

本问要求对这些古代玻璃文物的玻璃类型、纹饰和颜色与其是否表面风化进行相关性分析，根据古代玻璃的类型这一属性，对比文物样品表面有无风化化学成分含量，得出相应的统计规律。并根据风化点来预测文物风化前各化学成分的含量。

**5.2 数据预处理**

我们得到了一些文物的分类信息和文物的主要化学成分所占比例的相关数据，但其中有些数据是存在空值和异常值的，因此，我们对这些数据进行了数据清洗，从而得到完整、正确、一致的数据信息。

**5.3 模型的建立和求解**

**5.3.1 文物的表面风化与其玻璃类型、纹饰和颜色的关系处理**

我们通过表格1得出文物的玻璃类型共高钾、铅钡两种，纹饰共A、B、C三种，颜色共蓝绿、浅蓝、浅绿、紫、深绿、深蓝、黑七种

图5-3-1 表面风化与各属性值的比率

可以看出，玻璃类型为铅钡文物容易表面风化，颜色为黑色和蓝绿色的文物表面受风化的概率也很高，而纹饰为B类型的文物表面风化的概率达到百分之百。因此，可以初步判断玻璃类型为铅钡、纹饰为B、颜色为蓝绿色和黑色的古代玻璃文物与其表面发生风化存在一定的相关性。

#### 于是本文打算对古代玻璃文物的玻璃类型、纹饰、颜色与表面风化的关系进行关联性分析。 Apriori关联规则模型进行相关性分析，对古代玻璃文物表面风化与哪些具体因素有关有更清楚的认识，使得在之后为解决根据未知玻璃文物的表面风化进行分析，推测出属于哪种类型的玻璃文物的问题打下了一定的基础。

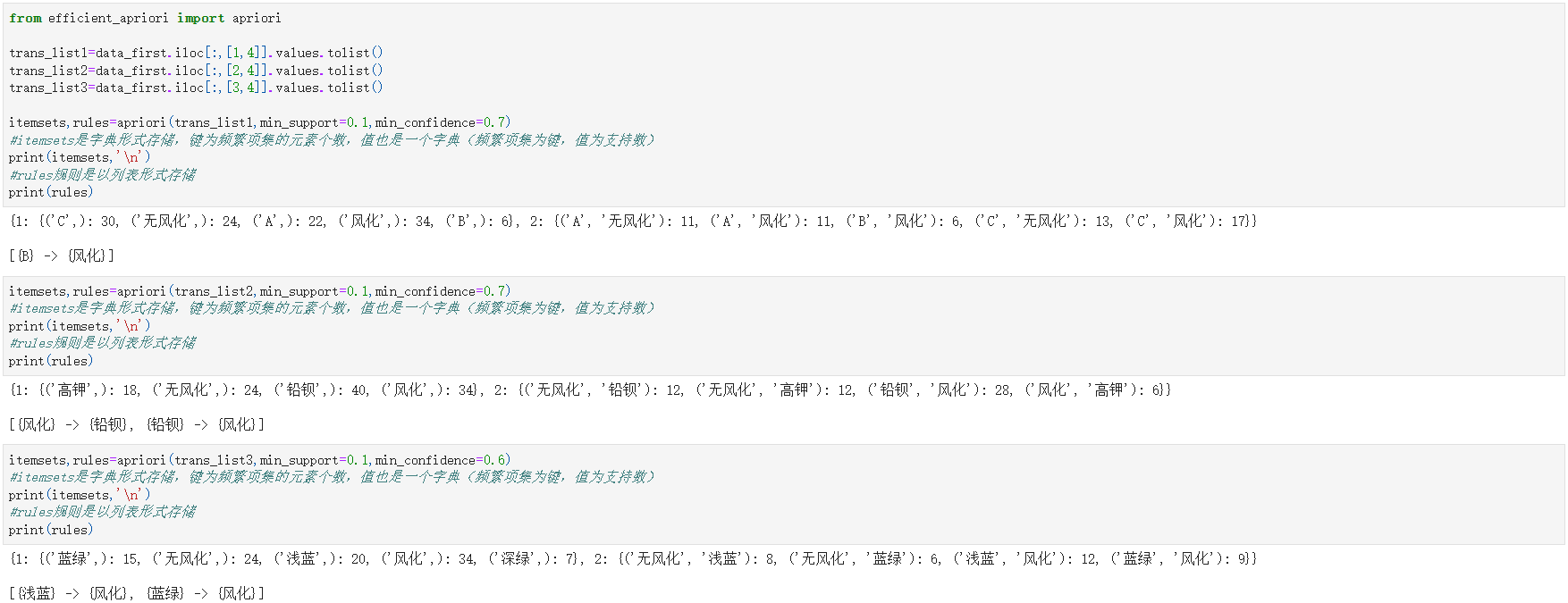


图5-3-2 Apriori算法相关代码

从代码实现的结果可以更详细的看出，在玻璃类型中，铅钡与表面风化可能具有一定的关联性；在颜色中，蓝绿色和浅蓝色与表面风化可能具有一定的关联性；而在纹饰中，B类与表面风化可能具有一定的关联性。

1. 对古代玻璃文物的各个属性与表面风化的关系使用Apriori关联算法分析

首先在我们建立的数据库中找出所有大于或等于我们指定的最小支持度的频繁项集，其中包括频繁一项集和频繁二项集，这里的最小支持度我们设置为0.1。然后利用频繁二项集生成需要的关联规则，根据设定的最小置信度进行一定地取舍，这里的最小置信度我们设置为0.7。最后根据计算得到的关联规则，筛选出强关联规则，从而得出表面风化与古代玻璃文物属性的关系。

本文利用jupyter软件帮我们进行了关联性分析，得到了以下的数据表格：

**表5-3-1  Apriori分析关联性**

|  |  |
| --- | --- |
| 关联规则 | 置信度 |
| （B类纹饰，表面风化） | 1.0 |
| （铅钡，表面风化） | 0.823529 |
| （表面风化，铅钡） | 0.700000 |
| （浅蓝色，表面风化） | 0.6 |
| （蓝绿色，表面风化） | 0.6 |

1. 结论

由该表可知，铅钡与表面风化、B类纹饰与表面风化的置信度均大于或等于最小置信度。因此，这些规则都为强关联规则，即具有很强的关联性。由此，我们可以得出B类纹饰、铅钡玻璃对表面风化有推进作用。

**5.3.2 文物样品表面有无风化化学成分含量的统计规律问题**

由于古代玻璃极易受埋藏环境的影响，从而表面会发生风化。但风化后的文物中的化学成分含量会发生变化。并且风化前与风化后的成分含量存在着一定的规律，而这些规律为今后研究古代玻璃文物风化的问题提供了参考。

图5-3-3 不同玻璃类型的表面是否风化分布情况

经过统计我们得到玻璃类型为铅钡且表面风化的文物有28个，无风化的有12个；玻璃类型为高钾且表面风化的文物有6个，无风化的有12个。由此可看出，不同玻璃类型的文物有无风化的样本个数有所不同，也可以推断不同古代玻璃样本其有无风化化学成分含量的规律也有所不同。假设各化学成分含量风化前后无显著差异。本文使用SPSS中的两独立样本T检验的方法检验差异性，可以得到：（假设ɑ=0.05）

表5-3-2 高钾样本各化学成分风化前后的差异性

|  |  |
| --- | --- |
| 化学成分 | 有无表面风化的含量是否存在差异 |
| 二氧化硅 | 是 |
| 氧化纳 | 否 |
| 氧化钾 | 是 |
| 氧化钙 | 是 |
| 氧化镁 | 是 |
| 氧化铝 | 是 |
| 氧化铁 | 是 |
| 氧化铜 | 否 |
| 氧化铅 | 是 |
| 氧化钡 | 否 |
| 五氧化二磷 | 否 |
| 氧化锶 | 是 |
| 氧化锡 | 否 |
| 二氧化硫 | 否 |

表5-3-3 铅钡样本各化学成分风化前后的差异性

|  |  |
| --- | --- |
| 化学成分 | 有无表面风化的含量是否存在差异 |
| 二氧化硅 | 是 |
| 氧化纳 | 是 |
| 氧化钾 | 是 |
| 氧化钙 | 是 |
| 氧化镁 | 否 |
| 氧化铝 | 否 |
| 氧化铁 | 否 |
| 氧化铜 | 否 |
| 氧化铅 | 是 |
| 氧化钡 | 否 |
| 五氧化二磷 | 是 |
| 氧化锶 | 是 |
| 氧化锡 | 否 |
| 二氧化硫 | 否 |

综上，可以总结出文物样品表面有无风化化学成分含量的统计规律。由以上图表可知，同一种玻璃类型的玻璃文物的各种化学成分风化前后的含量的差异性不同。大部分化学成分的含量前后在风化前后发生改变，而有些化学成分的含量在风化前后并无差异。并且同一个化学成分在不同的玻璃类型中可能会存在不同的差异性。

**5.3.3 预测文物样品表面风化前化学成分含量的问题**

对于考古学家来说，知道一件玻璃文物的成分含量可能会帮助他鉴别文物的真假。因此，知道文物的含量有着很重要的意义。而大部分的玻璃文物由于环境的因素，会出现表面风化的问题。大部分风化后的文物由于其成分含量会发生变化，从而会影响专家们的判断。故他们需要知晓文物风化前的化学成分含量，进而能够精确地判断文物类型。

我们可以对文物样本的某些部位进行采样，并根据这些采样点，可以观测到表单2中相应的风化后的主要化学成分所占比例，其中空白处表示未检测到该成分，可通过数据预处理对空值进行缺失值处理。通常将成分比例累加和介于 85%~105%之间的数据视为有效数据。并且，我们可以利用这些数据来预测文物样本风化前的化学物质的含量。

在文物采样点中，有风化、未被风化两种类型。筛选每个化学成分风化前的和风化后的数据，分为两组，计算每一组的平均值。最后进行均值比较，得出各个化学成分风化前的含量的预测结果。

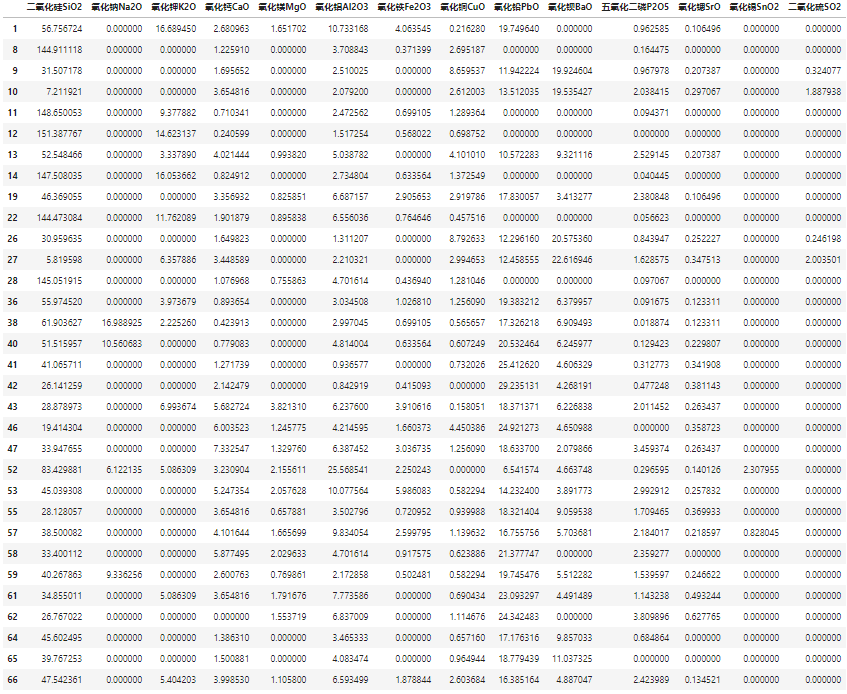


图5-3-4 预测风化点未风化时的化学成分

我们可以看出有的采样点的某些化学成分（如第1采样点的二氧化硅、氧化铜等）在风化后的含量比在风化前的含量高，说明这些化学物质含量会随着物品表面风化而增加，也进一步说明了风化会导致玻璃文物的某些化学物质的生成。

六、问题二的求解

**6.1** **问题分析**

分析高钾玻璃、铅钡玻璃的分类规律，选择合适的化学成分将这两类玻璃进行亚类划分，即凭合适的划分方法将高钾玻璃再细分几类，但是这两类划分的细类中可能会含有相互包含关系，并不是完全互斥的。根据题目可以看出可以用聚类模型，把数据对象按照数据的相似性划分成子集（更小的类）的过程；为了使得属于同一个类的样品间的距离尽可能小，使不同类的样品间距离尽可能打，我们在此分别用：基于划分的聚类（K-means）和基于密度的聚类（DBSCAN）两种模型分别对高钾玻璃和铅钡玻璃进行亚类划分，最后取得指标较好得方法。

**6.2 模型的建立和求解**

将表单附件2中的数据按玻璃类型分成两组，一组类型高钾玻璃，另一组为铅钡玻璃。利用PCA主成分分析方法将这两组数据。

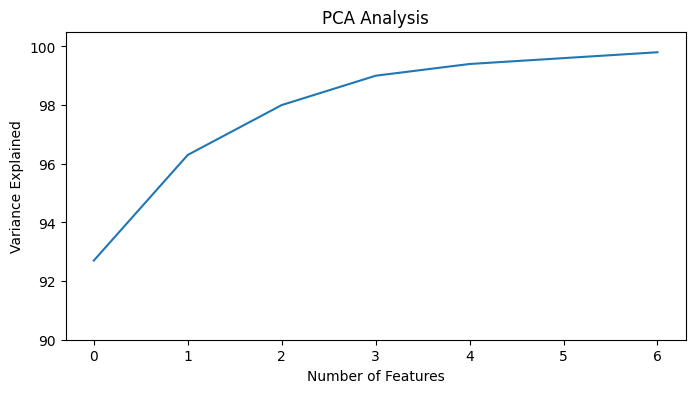


图6-2-1 PCA方法对高钾玻璃组数据降维

由图6-2-1得：排名最前的三个主成分已经可以解释98%的数据，所以选择将高钾玻璃组数据用PCA降成三维。

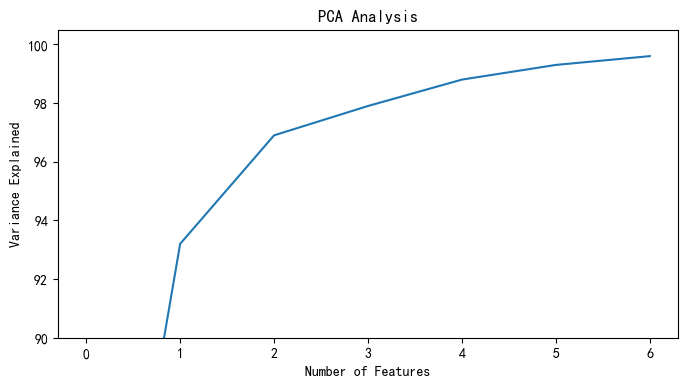


图6-2-2 PCA方法对铅钡玻璃组数据降维

由图6-2-1得：排名最前的五个主成分已经可以解释98.8%的数据，所以以下选择PCA降成5维

降维后两组数据情况如图4-2-4、图4-2-6所示。

做一个对照组，三维可视化两组降维后聚类前的数据，更加直观地对比分析两组数据聚类后划分子集的情况。

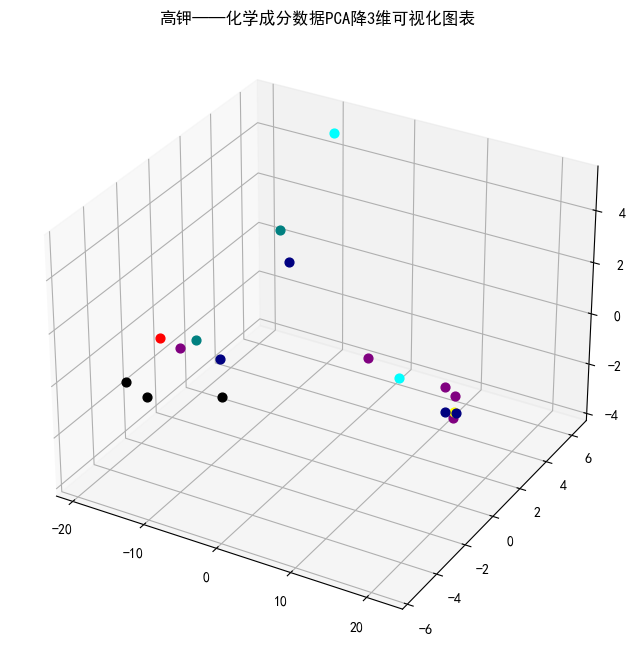


图6-2-2 对高钾玻璃数据进行PCA降维后可视化

分别采用K-means模型和DBSCAN模型对数据集进行聚类模型训练，再使用轮廓系数评价方法、Calinski-Harabasz评价方法和inertia指标分别对这两个模型的聚类结果进行评价选择两种模型中更优的模型，使得评价聚类性能准则达到最优，从而使得划分的每个子集内部紧凑，子集与子集之间相互独立。

1. 对高钾玻璃组使用K-means模型进行聚类，将这个簇内的所有数据样本的均值作为该簇的代表点，通过迭代过程把数据集划分为不同的类别。用轮廓系数评价方法、Calinski-Harabasz评价方法和inertia指标确定最优的簇的数目。评价结果如下：

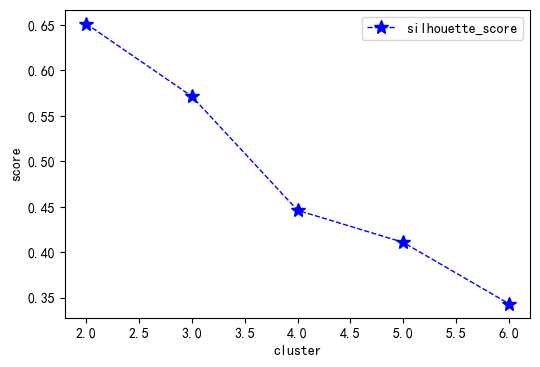


图6-2-3 轮廓系数对高钾玻璃组簇数评价

轮廓系数为： [0.6514641695284811, 0.5715250635027108, 0.44608029137320365, 0.41077750551279696, 0.343227902971504]

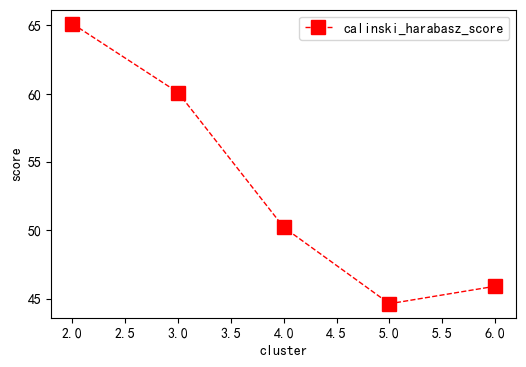


图6-2-4 Calinski-Harabasz评价方法对高钾玻璃组簇数评价

calinski\_harabasz系数为： [65.12413765971121, 60.08885909855051, 50.21754909162088, 44.627597262782444, 45.90725945840063]

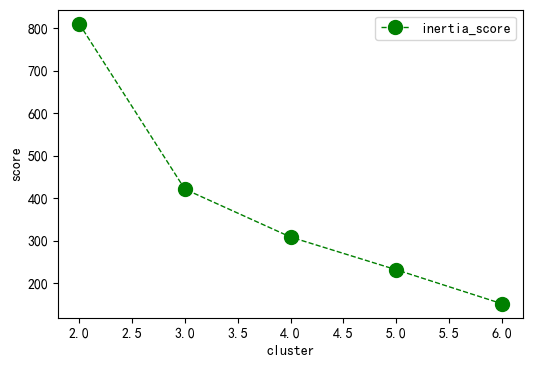


图6-2-5 inertia指标对高钾玻璃组簇数评价

inertia指标为： [810.4570201189017, 420.9231651057482, 308.21029268757565, 231.19363589914923, 151.18326969680425]

综合以上图示数据分析选择要将高钾玻璃分成三个细类更加合适。

以下进行分别采用K-means模型和DBSCAN模型对数据集进行聚类模型训练。

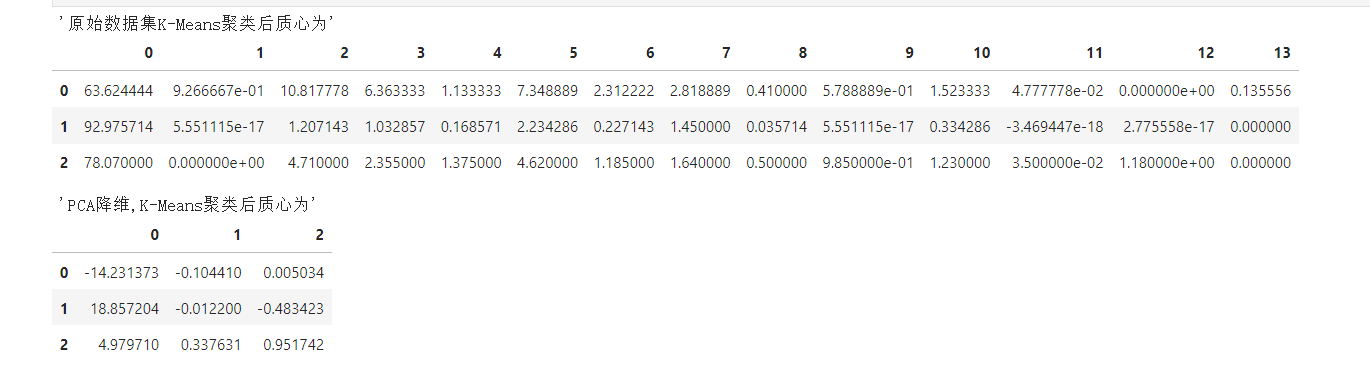


图6-2-5 原始数据集、PCA降维数据集的K-Means聚类后质心数评价

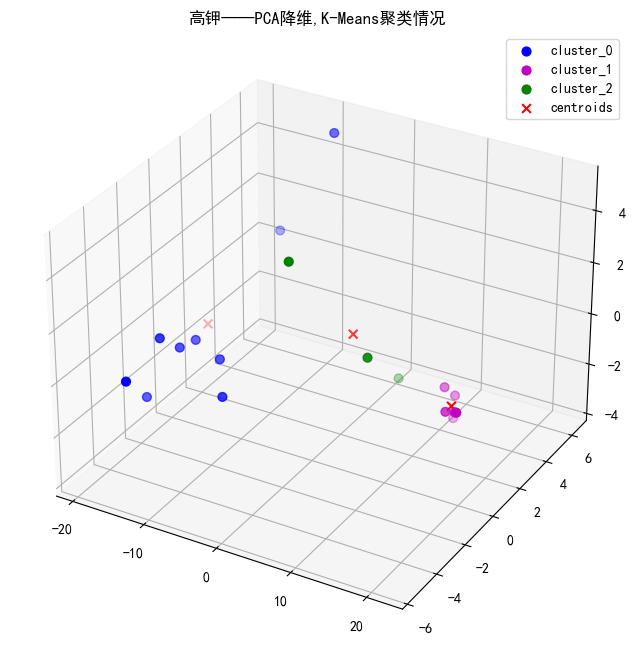


图6-2-6 高钾玻璃组K-Means聚类可视化

#### 建立高钾玻璃组通过DBSCAN聚类模型，先将簇的数据定义为密度相近的点的最大集合，将簇看作数据空间中被低密度区域分隔开的稠密对象区域，把具有密度足够的与区域划分为簇，并可以在有噪声的数据空间中发现任意形状的簇。评价聚类模型选择最优聚类簇数目。

silhouette方法、calinski\_harabasz方法进行对该模型划分簇数目评价，评价结果为：参数半径为7.1，簇类数目为2的的组合效果最优，聚类之后的簇数目有2个，噪声数据有3个，silhouette方法得分为0.5751285416476584；calinski\_harabasz方法得分为68.02868048756369。

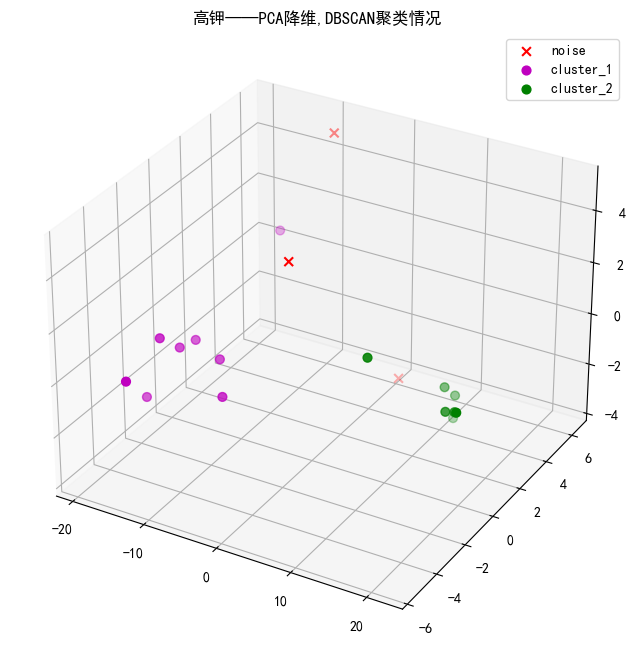


图6-2-7 高钾玻璃组DBSCAN聚类可视化

综合上面两种的可视化图与各种方法的评价指标，明显选用DBSCAN模型优于K-means模型，所以高钾玻璃组选用的划分方法是DBSCAN模型，能够将高钾玻璃再划分为两个更小的细类。

1. 与高钾玻璃组的方法一致，先对铅钡玻璃组使用K-means模型和DBSCAN模型分别进行聚类。用轮廓系数评价方法、Calinski-Harabasz评价方法和inertia指标确定最优的簇的数目。评价结果如下：

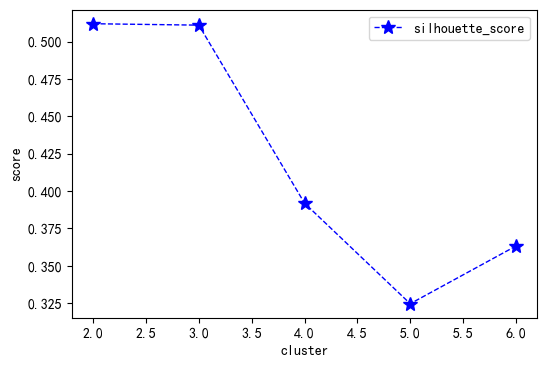


图6-2-8 轮廓系数对铅钡玻璃组簇数评价

轮廓系数为： [0.5119292021250728, 0.510976667859331, 0.391865729751237, 0.32457552783973037, 0.3631543034085193]

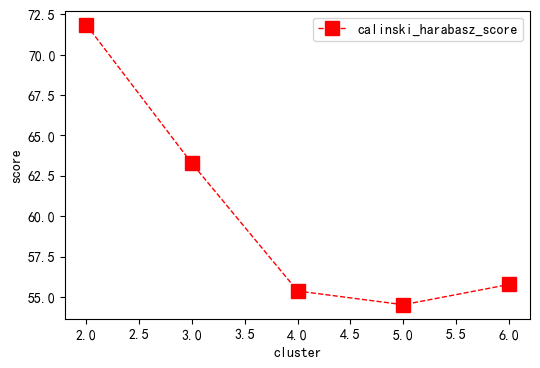


图6-2-9 Calinski-Harabasz评价方法对铅钡玻璃组簇数评价

calinski\_harabasz系数为： [71.85730416970063, 63.263797077520735, 55.36497669863502, 54.51636989737436, 55.77067395487239]

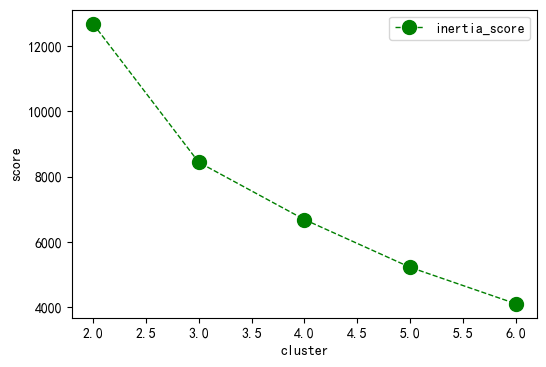


图6-2-10 inertia指标对铅钡玻璃组簇数评价

inertia指标为： [12685.953910056349, 8438.939449926063, 6683.7455458877375, 5220.010005452497, 4103.589314716419]

综合三张评价系数图表，再K-means方法中对铅钡玻璃组聚类划分选择簇的个数为3

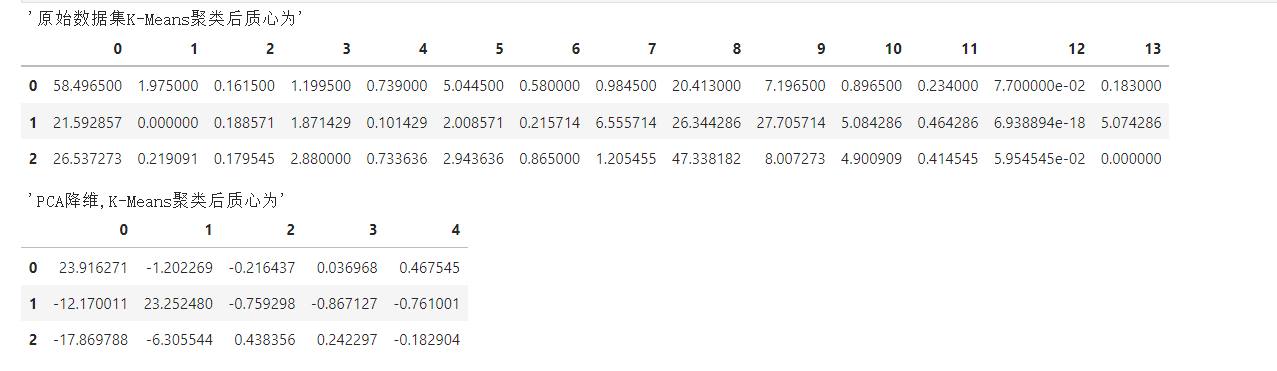


图6-2-11 原始数据集、PCA降维数据集的K-Means聚类后质心数评价

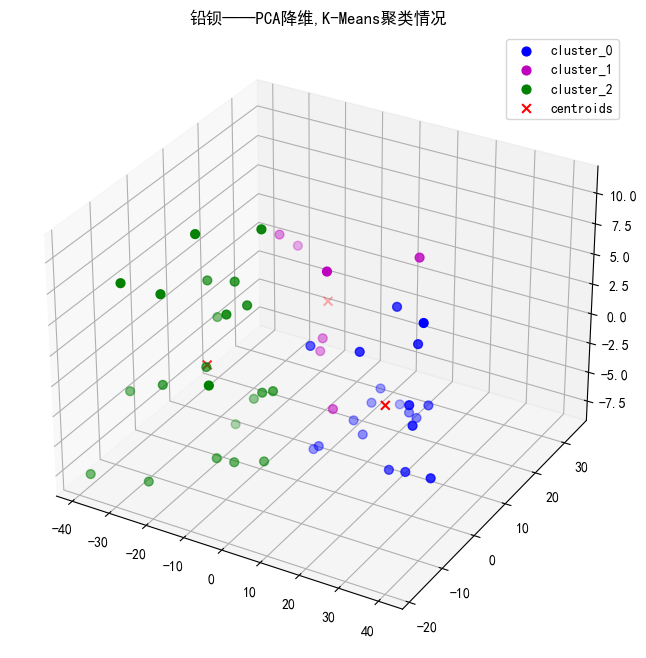


图6-2-12 铅钡玻璃组K-Means聚类可视化

#### 建立铅钡DBSCAN聚类模型，先将簇的数据定义为密度相近的点的最大集合，将簇看作数据空间中被低密度区域分隔开的稠密对象区域，把具有密度足够的与区域划分为簇，并可以在有噪声的数据空间中发现任意形状的簇。评价聚类模型选择最优聚类簇数目。

silhouette方法、calinski\_harabasz方法进行对该模型划分簇数目评价，评价结果为：参数半径为12，簇类数目为3的的组合效果最优，聚类之后的簇数目有2个，噪声数据有7个，silhouette方法得分为0.5215008141530673；calinski\_harabasz方法得分为65.99727007427376。

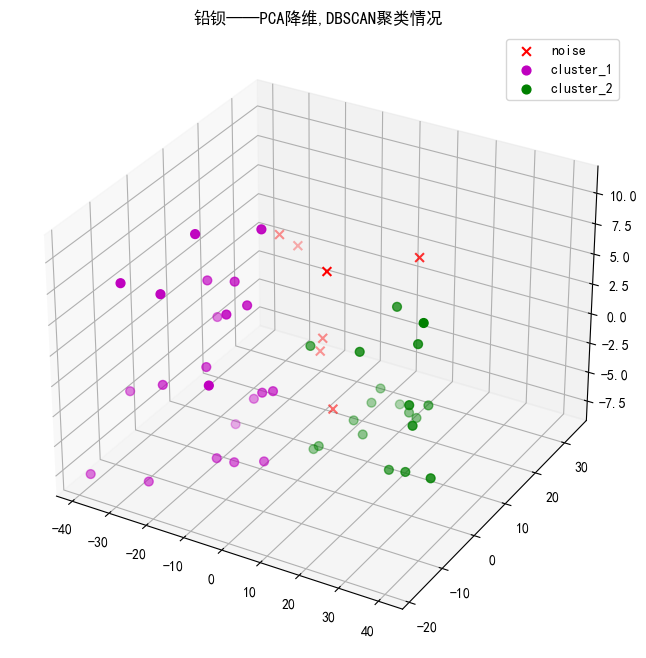


图6-2-13 铅钡玻璃组DBSCAN聚类可视化

综合上面两种的可视化图与各种方法的评价指标，明显选用K-means模型优于DBSCAN模型，所以铅钡玻璃组选用的划分方法是K-means模型，能够将铅钡玻璃再划分为3个更小的细类。

**七、问题三的求解**

**7.1 问题分析**

依据化学成分分析玻璃类别的分类规律，说明内部化学元素成分对玻璃的类别鉴别有一定的规律可循。对附件表单 3 中未知类别玻璃文物的化学成分进行分析，鉴别其所属类型。

本问是典型的二分类预测问题，常用的分类与预测算法有回归分析、决策树、人工神经网络、支持向量机等，由于维度有14维，关系较为复杂，所以本问采用深度学习中的人工神经网络，建立全连接神经网络（Fully Connected Neural Network）预测模型，并以预测结果与实际结果的均方差（Mean Square Error, MSE）作为评估指标。

**7.2 数据预处理**

在进行预测前，将进行以下数据预处理

附件表单2使用python的drop函数删除除了14个化学成分含量列和玻璃类别列之外的所有列，使用sum()函数算出每件文物相应的主要成分所占比例（空白处表示未检测到该成分，为0值）累加值，由题目限制条件，将成分比例累加和介于 85%~105%之间的数据视为有效数据（67条有效），其余的筛去（2条无效），“玻璃类型”列的标签转换成数值（“0”代表高钾，“1”代表铅钡）。使用train\_test\_split分离器函数，将矩阵划分为训练集和测试集，训练集（54条）占总数据集的80%，测试集（13条）占总数据集的20%。

附件表单3使用python的drop函数删除第一列“文物编号”和第二列“表面风化”，fillna(0)函数将空白值处填充0值，此表单数据处理好后作为训练之后模型的真实预测数据。

**7.3 模型的建立和求解**

构建全神经连接网络进行类别预测，主要步骤如下：

1. 搭建网络结构

Tensorflow是深度学习中比较受欢迎的的库之一，使用其中Keras库的Sequential模型与model类，进行神经网络模型的搭建

1. 配置模型参数并编译

在数据的输入训练之前需要设定模型的参数，配置损失函数、优化器、激活函数、训练周期等。

损失函数Loss，是用来评价模型的预测值F(x)与实际值Y的差异程度的一项指标，损失函数越小，模型的预测误差越小，预测的精确值和效果越好。为了方便计算，一般使用均方差（Mean Square Error, MSE），此时对于n个样本的损失值为：

优化器Optimizer，用于控制梯度裁剪,使用SGD随机梯度下降法，支持动量参数，支持学习衰减率。

1. 对训练数据进行拟合训练

输入“训练集”数据，训练轮数 epochs=10

1. 对训练好的模型进行评估

将“测试集”的数据输入到上一步训练好的模型中进行测试，输出最终的评价指标（MSE），并绘制模型的损失函数折线图。

部分核心代码如下：

文本

描述已自动生成

表格

描述已自动生成

图形用户界面, 文本, 应用程序

描述已自动生成

图表, 折线图

描述已自动生成

图7-1-1 全连接神经网络损失函数折线图

上图7.1是全连接神经网络模型在训练和测试过程中各自的损失变化曲线，分析可得，训练集中前2次epoch损失值极速下降，之后下降趋势明显减缓，在第7次前后到达最低点并基本保持平稳；测试集中的损失值在第4次epoch后逐渐平稳。

由此看出，此模型的最终MSE小于0.01，模型的预测精确性和效果较好。

（5）根据模型预测玻璃类型

将附件表单3数据处理好后作为以上训练完成模型的输入，输出预测值，将预测值标签再转换成对应的玻璃类型（“0”代表高钾，“1”代表铅钡），合并到原数据集中输出。

预测类型分别为['高钾', '铅钡', '铅钡', '铅钡', '铅钡', '高钾', '高钾', '铅钡']

电脑萤幕截图

描述已自动生成

图7-1-2 全连接神经网络预测结果图

八、问题四的求解

**8.1** 问题分析

通常，人们对玻璃文物的鉴定方法不仅仅是通过该文物表面风化前后的化学成分含量，还要通过同一类别的玻璃文物的不同化学成分之间的比较，才能等到更精准的鉴定结果。不同类别的玻璃文物样品，其化学成分可能具有一定的关联或者差异。因此本文对于不同类别的玻璃文物的化学成分之间的关系进行相关分析模型，评估他们之间的关联性或差异性。

**8.2 模型的建立与求解**

由于该问题建立是典型的相关性模型，我们可以选择建立皮尔逊积距相关系数矩阵来表示不同化学成分之间的相关系数。

相关系数定义为两个化学成分之间写方差和标准差的商。相关系数的取值区间是[-1,1]。如果相关系数大于0，那么这两个化学成分正相关，其中一个化学成分含量随着另一个化学成分含量的增加而增加，相关系数的值越大，相关性越强。如果相关系数为0，则说明这两个化学成分之间是相互独立的，它们之间不存在相关性。如果相关系数小于0，则这两个化学成分是负相关，即其中一个化学成分含量随另一个化学成分含量的减少而增加，这意味着两个化学成分相斥。

将各个化学成分按照表单3的顺序进行编号，利用Python语言进行编程，并可得出皮尔逊积距相关系数矩阵为：

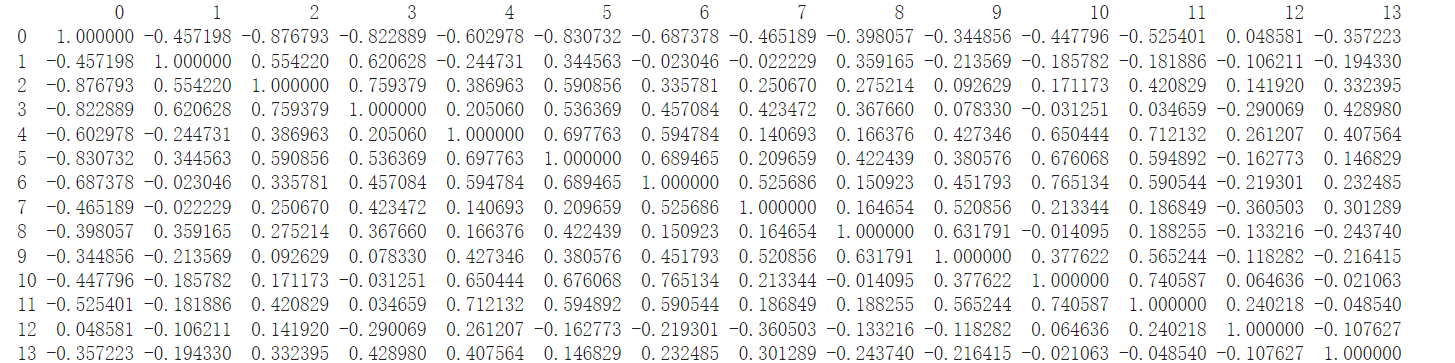


图8-2-1 高钾皮尔逊积距相关系数矩阵

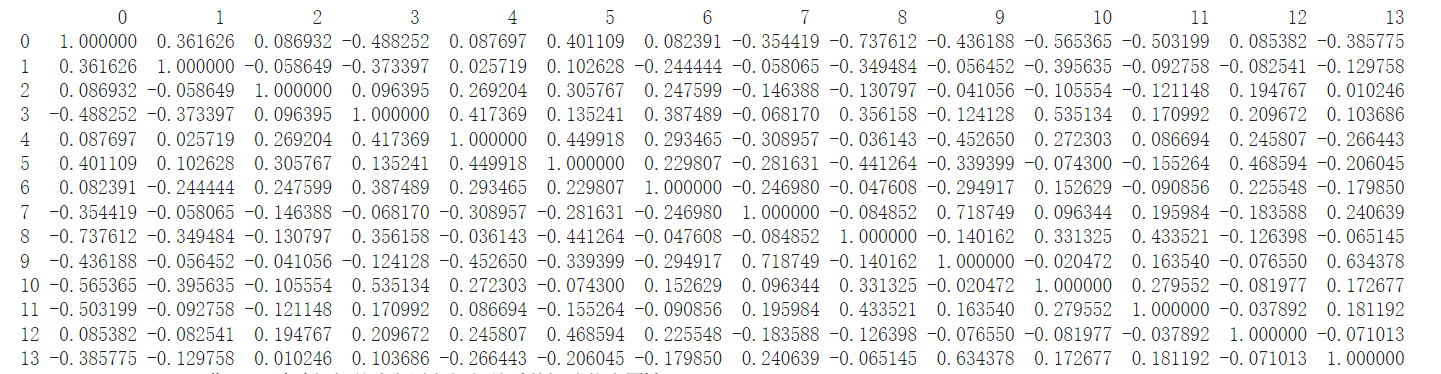


图8-2-2 铅钡皮尔逊积距相关系数矩阵

其中，我们可以根据不同类别的皮尔逊积距相关系数矩阵中各个元素值可以初步判断哪些化学成分是相关的，哪些化学成分是相异的，哪些化学成分是相斥的。在高钾和铅钡中，大多数化学成分是正相关的关系，少部分是相斥关系，不存在相异关系。当然，也可以使用热力图更直观地表示各化学成分的关系。利用jupyter工具进行可视化为：

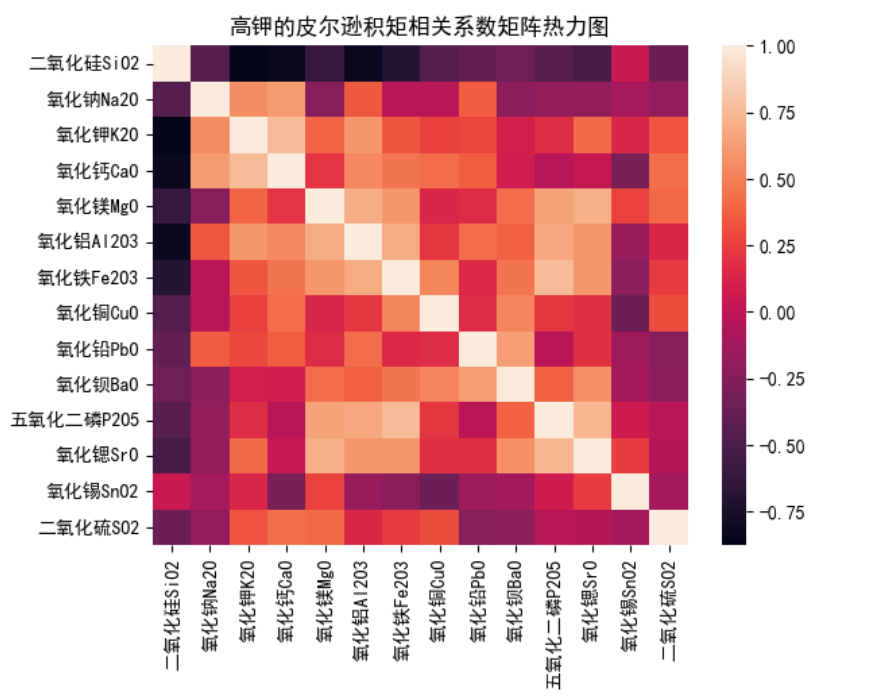


图8-2-3 高钾皮尔逊积距相关系数矩阵热力图

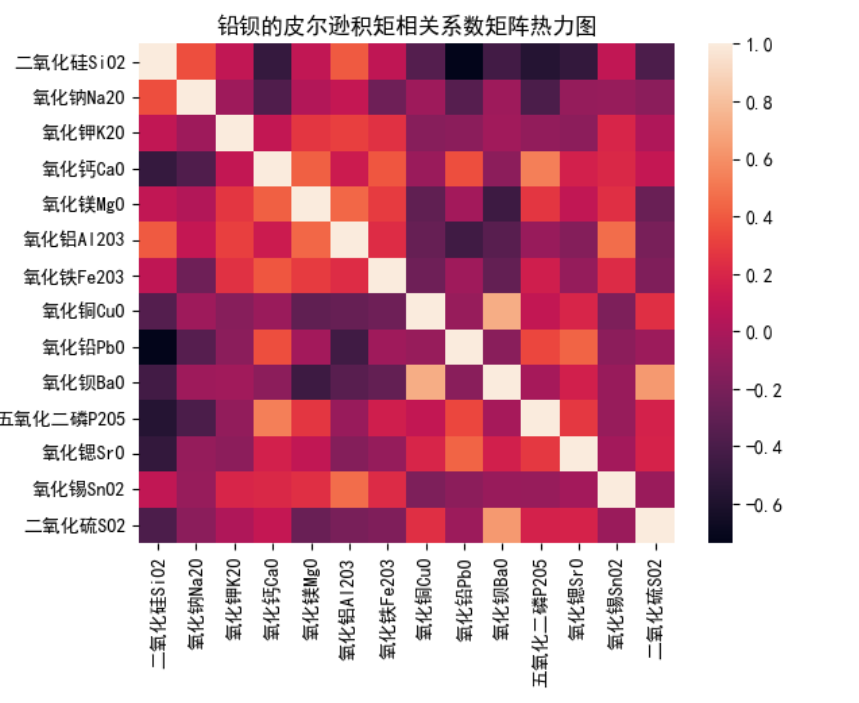


图8-2-3 铅钡皮尔逊积距相关系数矩阵热力图

为了使得我们的模型更加的精准，我们将利用SPSS工具建立计量资料相关分析，分析两个化学成分之间有无线性相关关系，关系是否密切。（其中ɑ=0.01）

表8-2-1 spss分析高钾的化学成分相关性

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **高钾的化学成分相关分析** | | | | | | | | | | | | | | | |
|  | | 二氧化硅SiO2 | 氧化钠Na2O | 氧化钾K2O | 氧化钙CaO | 氧化镁MgO | 氧化铝Al2O3 |  |  |  |  |  |  |  |  |
| 二氧化硅SiO2 | Pearson Correlation | -- |  |  |  |  |  |  |  |  |  |  |  |  |  |
| N | 18 |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 氧化钠Na2O | Pearson Correlation | -.457 | -- |  |  |  |  |  |  |  |  |  |  |  |  |
| Sig. (2-tailed) | .056 |  |  |  |  |  |  |  |  |  |  |  |  |  |
| N | 18 | 18 |  |  |  |  |  |  |  |  |  |  |  |  |
| 氧化钾K2O | Pearson Correlation | -.877\*\* | .554\* | -- |  |  |  |  |  |  |  |  |  |  |  |
| Sig. (2-tailed) | .000 | .017 |  |  |  |  |  |  |  |  |  |  |  |  |
| N | 18 | 18 | 18 |  |  |  |  |  |  |  |  |  |  |  |
| 氧化钙CaO | Pearson Correlation | -.823\*\* | .621\*\* | .759\*\* | -- |  |  |  |  |  |  |  |  |  |  |
| Sig. (2-tailed) | .000 | .006 | .000 |  |  |  |  |  |  |  |  |  |  |  |
| N | 18 | 18 | 18 | 18 |  |  |  |  |  |  |  |  |  |  |
| 氧化镁MgO | Pearson Correlation | -.603\*\* | -.245 | .387 | .205 | -- |  |  |  |  |  |  |  |  |  |
| Sig. (2-tailed) | .008 | .328 | .113 | .414 |  |  |  |  |  |  |  |  |  |  |
| N | 18 | 18 | 18 | 18 | 18 |  |  |  |  |  |  |  |  |  |
| 氧化铝Al2O3 | Pearson Correlation | -.831\*\* | .345 | .591\*\* | .536\* | .698\*\* | -- |  |  |  |  |  |  |  |  |
| Sig. (2-tailed) | .000 | .161 | .010 | .022 | .001 |  |  |  |  |  |  |  |  |  |
| N | 18 | 18 | 18 | 18 | 18 | 18 |  |  |  |  |  |  |  |  |
| 氧化铁Fe2O3 | Pearson Correlation | -.687\*\* | -.023 | .336 | .457 | .595\*\* | .689\*\* |  |  |  |  |  |  |  |  |
| Sig. (2-tailed) | .002 | .928 | .173 | .057 | .009 | .002 |  |  |  |  |  |  |  |  |
| N | 18 | 18 | 18 | 18 | 18 | 18 |  |  |  |  |  |  |  |  |
| 氧化铜CuO | Pearson Correlation | -.465 | -.022 | .251 | .423 | .141 | .210 |  |  |  |  |  |  |  |  |
| Sig. (2-tailed) | .052 | .930 | .316 | .080 | .578 | .404 |  |  |  |  |  |  |  |  |
| N | 18 | 18 | 18 | 18 | 18 | 18 |  |  |  |  |  |  |  |  |
| 氧化铅PbO | Pearson Correlation | -.398 | .359 | .275 | .368 | .166 | .422 |  |  |  |  |  |  |  |  |
| Sig. (2-tailed) | .102 | .143 | .269 | .133 | .509 | .081 |  |  |  |  |  |  |  |  |
| N | 18 | 18 | 18 | 18 | 18 | 18 |  |  |  |  |  |  |  |  |
| 氧化钡BaO | Pearson Correlation | -.345 | -.214 | .093 | .078 | .427 | .381 |  |  |  |  |  |  |  |  |
| Sig. (2-tailed) | .161 | .395 | .715 | .757 | .077 | .119 |  |  |  |  |  |  |  |  |
| N | 18 | 18 | 18 | 18 | 18 | 18 |  |  |  |  |  |  |  |  |
| 五氧化二磷P2O5 | Pearson Correlation | -.448 | -.186 | .171 | -.031 | .650\*\* | .676\*\* |  |  |  |  |  |  |  |  |
| Sig. (2-tailed) | .062 | .460 | .497 | .902 | .003 | .002 |  |  |  |  |  |  |  |  |
| N | 18 | 18 | 18 | 18 | 18 | 18 |  |  |  |  |  |  |  |  |
| 氧化锶SrO | Pearson Correlation | -.525\* | -.182 | .421 | .035 | .712\*\* | .595\*\* |  |  |  |  |  |  |  |  |
| Sig. (2-tailed) | .025 | .470 | .082 | .891 | .001 | .009 |  |  |  |  |  |  |  |  |
| N | 18 | 18 | 18 | 18 | 18 | 18 |  |  |  |  |  |  |  |  |
| 氧化锡SnO2 | Pearson Correlation | .049 | -.106 | .142 | -.290 | .261 | -.163 |  |  |  |  |  |  |  |  |
| Sig. (2-tailed) | .848 | .675 | .574 | .243 | .295 | .519 |  |  |  |  |  |  |  |  |
| N | 18 | 18 | 18 | 18 | 18 | 18 |  |  |  |  |  |  |  |  |
| 二氧化硫SO2 | Pearson Correlation | -.357 | -.194 | .332 | .429 | .408 | .147 |  |  |  |  |  |  |  |  |
| Sig. (2-tailed) | .146 | .440 | .178 | .076 | .093 | .561 |  |  |  |  |  |  |  |  |
| N | 18 | 18 | 18 | 18 | 18 | 18 |  |  |  |  |  |  |  |  |

表8-2-2 spss分析铅钡的化学成分相关性

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **铅钡的化学成分相关分析** | | | | | | | | | | | | | | | |
|  | | 二氧化硅SiO2 | 氧化钠Na2O | 氧化钾K2O | 氧化钙CaO | 氧化镁MgO | 氧化铝Al2O3 |  |  |  |  |  |  |  |  |
| 二氧化硅SiO2 | Pearson Correlation | -- |  |  |  |  |  |  |  |  |  |  |  |  |  |
| N | 49 |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 氧化钠Na2O | Pearson Correlation | .362\* | -- |  |  |  |  |  |  |  |  |  |  |  |  |
| Sig. (2-tailed) | .011 |  |  |  |  |  |  |  |  |  |  |  |  |  |
| N | 49 | 49 |  |  |  |  |  |  |  |  |  |  |  |  |
| 氧化钾K2O | Pearson Correlation | .087 | -.059 | -- |  |  |  |  |  |  |  |  |  |  |  |
| Sig. (2-tailed) | .553 | .689 |  |  |  |  |  |  |  |  |  |  |  |  |
| N | 49 | 49 | 49 |  |  |  |  |  |  |  |  |  |  |  |
| 氧化钙CaO | Pearson Correlation | -.488\*\* | -.373\*\* | .096 | -- |  |  |  |  |  |  |  |  |  |  |
| Sig. (2-tailed) | .000 | .008 | .510 |  |  |  |  |  |  |  |  |  |  |  |
| N | 49 | 49 | 49 | 49 |  |  |  |  |  |  |  |  |  |  |
| 氧化镁MgO | Pearson Correlation | .088 | .026 | .269 | .417\*\* | -- |  |  |  |  |  |  |  |  |  |
| Sig. (2-tailed) | .549 | .861 | .061 | .003 |  |  |  |  |  |  |  |  |  |  |
| N | 49 | 49 | 49 | 49 | 49 |  |  |  |  |  |  |  |  |  |
| 氧化铝Al2O3 | Pearson Correlation | .401\*\* | .103 | .306\* | .135 | .450\*\* | -- |  |  |  |  |  |  |  |  |
| Sig. (2-tailed) | .004 | .483 | .033 | .354 | .001 |  |  |  |  |  |  |  |  |  |
| N | 49 | 49 | 49 | 49 | 49 | 49 |  |  |  |  |  |  |  |  |
| 氧化铁Fe2O3 | Pearson Correlation | .082 | -.244 | .248 | .387\*\* | .293\* | .230 |  |  |  |  |  |  |  |  |
| Sig. (2-tailed) | .574 | .091 | .086 | .006 | .041 | .112 |  |  |  |  |  |  |  |  |
| N | 49 | 49 | 49 | 49 | 49 | 49 |  |  |  |  |  |  |  |  |
| 氧化铜CuO | Pearson Correlation | -.354\* | -.058 | -.146 | -.068 | -.309\* | -.282\* |  |  |  |  |  |  |  |  |
| Sig. (2-tailed) | .012 | .692 | .316 | .642 | .031 | .050 |  |  |  |  |  |  |  |  |
| N | 49 | 49 | 49 | 49 | 49 | 49 |  |  |  |  |  |  |  |  |
| 氧化铅PbO | Pearson Correlation | -.738\*\* | -.349\* | -.131 | .356\* | -.036 | -.441\*\* |  |  |  |  |  |  |  |  |
| Sig. (2-tailed) | .000 | .014 | .370 | .012 | .805 | .002 |  |  |  |  |  |  |  |  |
| N | 49 | 49 | 49 | 49 | 49 | 49 |  |  |  |  |  |  |  |  |
| 氧化钡BaO | Pearson Correlation | -.436\*\* | -.056 | -.041 | -.124 | -.453\*\* | -.339\* |  |  |  |  |  |  |  |  |
| Sig. (2-tailed) | .002 | .700 | .779 | .395 | .001 | .017 |  |  |  |  |  |  |  |  |
| N | 49 | 49 | 49 | 49 | 49 | 49 |  |  |  |  |  |  |  |  |
| 五氧化二磷P2O5 | Pearson Correlation | -.565\*\* | -.396\*\* | -.106 | .535\*\* | .272 | -.074 |  |  |  |  |  |  |  |  |
| Sig. (2-tailed) | .000 | .005 | .470 | .000 | .058 | .612 |  |  |  |  |  |  |  |  |
| N | 49 | 49 | 49 | 49 | 49 | 49 |  |  |  |  |  |  |  |  |
| 氧化锶SrO | Pearson Correlation | -.503\*\* | -.093 | -.121 | .171 | .087 | -.155 |  |  |  |  |  |  |  |  |
| Sig. (2-tailed) | .000 | .526 | .407 | .240 | .554 | .287 |  |  |  |  |  |  |  |  |
| N | 49 | 49 | 49 | 49 | 49 | 49 |  |  |  |  |  |  |  |  |
| 氧化锡SnO2 | Pearson Correlation | .085 | -.083 | .195 | .210 | .246 | .469\*\* |  |  |  |  |  |  |  |  |
| Sig. (2-tailed) | .560 | .573 | .180 | .148 | .089 | .001 |  |  |  |  |  |  |  |  |
| N | 49 | 49 | 49 | 49 | 49 | 49 |  |  |  |  |  |  |  |  |
| 二氧化硫SO2 | Pearson Correlation | -.386\*\* | -.130 | .010 | .104 | -.266 | -.206 |  |  |  |  |  |  |  |  |
| Sig. (2-tailed) | .006 | .374 | .944 | .478 | .064 | .155 |  |  |  |  |  |  |  |  |
| N | 49 | 49 | 49 | 49 | 49 | 49 |  |  |  |  |  |  |  |  |

其中，sig<=0.01时，两个化学成分之间的相关关系非常显著；sig>0.01且sig<0.05时，两个化学成分之间相关关系显著。看的出，改进后的模型优化了相关系数的数据，更符合我们预想的各个化学成分关系，不同的化学成分会存在关联关系，也会存在相异和相斥关系。这些都是取决于它们之间的相关系数的值的大小。

由此，在上述结果中，由SPSS相关分析皮尔逊系数矩阵得出以下结论，二氧化硅SiO2与氧化钙CaO、氧化铝Al2O3、氧化铅PbO、氧化钡BaO、五氧化二磷P2O5、氧化铯SrO、二氧化硫SO2关系非常显著，与氧化纳Na2O、氧化铜CuO关系一般显著，与其余化学成分关系不显著等结论，详细矩阵见附录：“SPSS数据和分析输出”文件夹中的“第4问相关分析【铅钡】.spv”与“第4问相关分析【高钾.spv】”

1. **模型评价**

**9.1模型的优点**

1. 第一题中分析玻璃文物的表面风化与其玻璃类型、纹饰和颜色的关系用的Apriori关联规则算法，此算法可以减少再数据库上进行检查的集合的数目；分析文物样品表面有无风化化学成分含量的统计规律，用的是独立样本T检验的方法，t 检验后获得的结果有助于得出结论，如果它们实际上是正确的，它们可以应用于整个总体，同时其具有很好的鲁棒性、易懂；预测已风化文物样品风化前的化学成分含量用的方法是均值比较，该方法简单移动、速度快。
2. 第二题中运用的模型是K-means和DBCSAN，K-means算法对于处理大数据集有着很好的可伸缩性和高效性，当簇是密集的并且簇与簇之间区别明显时效果很好。DBCSAN算法是基于密度的聚类算法可以发现任意形状的簇，这对于有噪声的数据很有利，可以过滤掉“噪声点”和“离群点”。

**9.2模型的缺点**

1. 用Apriori关联规则算法，每次增加频繁项集的大小，会重新扫描整个数据集，所以当数据集很大时，会显著降低频繁项集的发现速度。T检验会受环境影响，独立t检验可以确定样本组的区别，但不能控制环境的影响，环境变化可能会影响 t 检验的输出。均值对极端值及其敏感，有时不能反映数据整体情况。
2. K-means算法，只有在初始时通过随机产生的簇中心是簇中真实存在的点，当簇内内含有异常点，会导致均值严重偏离；并且K-means对初始值很敏感。DBCSAN算法当数据点的密度不均匀各簇的簇内距离相差很大时聚类的质量很差。

**9.3模型的优化**

1. 在均值比较中应该对数据值进行排序，并且计算均值前去掉高端低端的2%，应避免在两端截去太多的数据，否则导致有价值的信息被截掉。
2. 对K-means算法的优化：应该在每次迭代后的簇中心要在簇的样本点中选取，选取的标准就是当样本点成为新的簇中心后能提高簇类的质量。