针对古代玻璃化学成分分析与鉴别的研究

摘要

玻璃制品在被埋藏后，极易受到环境的影响而风化，风化过程会改变其内部各种化学成分的含量，从而影响对其种类的判断。本文围绕玻璃文物分类的问题，基于附件中给出的数据并借助 SPSS 等数据分析软件，建立了玻璃类型与风化情况及文物中各类化学成分含量之间的分类统计模型。

针对问题一：为了描述玻璃文物的表面风化情况与其类型、颜色和纹饰之间的关系，我们采用了**卡方独立性检验**与 **Yates 校正卡方检验**，根据计算结果得出玻璃文物表面是否风化与其玻璃类型有着**强相关性**，而与颜色和纹饰的相关性则**较弱**。而对于第二小问，我们首先根据玻璃类型将数据分成两类，接着使用 **Pearson 相关性分析**，挑选出了一些对是否风化影响较为显著的化学成分，在**高钾玻璃**中，**氧化钾**的含量对于是否风化有着较为重要的影响，而在**铅钡玻璃**中则是**氧化铅**的含量。接着以这些成分为自变量，是否风化为因变量，建立了**二元 Logistic 回归模型**，检验显著性均在 0.9 以上且对于原数据集的分类正确百分比超过了 95%，模型的拟合度较好。最后对于风化前含量的预测，我们在前两小问的基础上，建立了**基于综合评价法的文物相似度评价模型**，用与风化文物最为相似的未风化文物的化学成分含量，来推断其风化前的数据。

针对问题二：为了探究高钾玻璃和铅钡玻璃的分类规律，本文首先利用 **Pearson 相关性分析**，去除了对玻璃类型影响较弱的一些化学成分，接着以剩下的化学成分为自变量，玻璃类型为因变量，使用**支持向量机**生成分类超平面。通过对超平面的系数进行分析可以发现，**二氧化硅及铅钡含量较高**的玻璃一般可以分入**铅钡玻璃类**，而**二氧化硅含量较低，钾钙含量较高**的则划入**高钾玻璃类**。在第二小问中，我们选用了 **K-均值聚类**和**层次聚类**算法，在两种主要玻璃类型中，分别以**钾钙含量和铅钡含量**作为特征数据进行聚类，并根据结果进行亚类划分，最后进行了**敏感性分析**。

针对问题三：在问题二得出了玻璃类型划分标准的基础上，将未知类别玻璃文物的化学成分代入到已求得的分类标准中进行划分，之后又进行了亚类划分。为了探究分类结果的敏感性，我们对分类模型的参数进行了多次调整，得出的分类结果与初始的分类结果均无较大变化，这说明了预测结果对分类方法及其中参数的**敏感性较弱**。

针对问题四：为了探究不同类别的玻璃文物样品化学成分之间的关联关系，我们采用 **Pearson 相关性分析**刻画同类玻璃化学成分之间的关联性大小，筛选出每个类别主要关联的化学成分。通过建立关于主要关联的化学成分的**多元线性回归模型**，量化分析了其不同化学成分彼此的关联程度。通过比较不同类别的回归方程自变量的种类，分析不同玻璃内部化学成分互相关联的差异性，推测可能引起差异的原因。

关键字： Pearson 相关性分析　支持向量机　二元 Logistic 回归　聚类　综合评价模型

一、 问题重述

1.1 问题背景

玻璃是中西方丝绸之路中早期贸易往来的宝贵物证。早期的玻璃从西方传入我国，而在吸收了西方的技术之后，我国本土也开始了玻璃的制作。在炼制玻璃时，加入的助熔剂不同，得到的玻璃类型也不同，如铅钡玻璃、钾玻璃等。当古代玻璃被埋藏时，极易受到环境的影响而风化。

1.2 问题要求

现在给出一批古代玻璃品的相关数据，其中包括每个玻璃的类型、纹饰、颜色以及有无风化等情况。同时，根据技术检测的结果，也给出了每个文物各类成分的占比。我们团队需要根据上述数据，解决下面这些问题：

（1）分析这些文物的风化情况与其玻璃类型、纹饰、颜色是否存在关系，并定量分析不同类型的玻璃的风化情况与化学成分含量之间的规律，最后预测其风化前的化学成分含量。

（2）根据化学成分含量，分析高钾玻璃与铅钡玻璃的分类规律，并选择合适的化学成分，对每类进行二次亚类划分，分析结果的合理性和敏感性。

（3）利用第二问得到的规律，鉴别附件表单 3 中文物的所属类型并分析结果的敏感性。

（4）分析不同类别文物化学成分之间的关联并比较其差异性。

二、 问题分析

2.1 问题一的分析

问题一要求根据附件表单 1 中的数据，分析玻璃文物表面是否风化与其纹饰、颜色和玻璃类型之间的关系。由于是否风化、纹饰、颜色以及玻璃类型均为定类变量，因此可以利用 Pearson 卡方检验来判断各变量之间是否存在相关性。考虑到其中某些情况不适用卡方检验，我们又使用了 Yates 校正卡方检验，得到了是否风化与三种因素之间的相关性强弱。接着，我们从附件表单 2 中的数据可以看出，化学成分含量均为定量变量，而是否风化则是定类变量，因此我们构建了二元 Logistic 回归模型，得到了各种化学成分含量与是否风化之间的定量关系。最后，对于未风化前数据的预测，由于数据量较小，因此我们构建了评价模型，构建文物相似性指标，并用相似未风化文物的数据来进行推断预测。

2.2 问题二的分析

问题二要求分析文物化学含量与其玻璃种类之间的关联，为了避免某些与玻璃类型无关的化学成分对后续分析造成影响，首先进行 Pearson 相关性分析，保留与玻璃类型有较强相关性的化学成分作为自变量，而后使用支持向量机算法，得到玻璃类型与化学成分含量之间的划分规律。而对于亚类划分问题，则可采用 K-均值聚类和层次聚类算法进行二次分类，比较分类标准对分类方法的敏感性。

2.3 问题三的分析

问题三要求根据附件表单 3 中玻璃文物的各类化学成分含量，进行玻璃类型的划分，由于在问题二中已经求得玻璃类型与化学成分含量之间的划分规律，因此可以将已知的数据代入，得出结果。最后进行敏感性分析。

2.4 问题四的分析

三、 模型假设

为了构建更加精确的数学模型，本文根据实际情况作出下列合理的假设：

- 文物的玻璃类型、纹饰以及颜色相互独立；
- 文物在遭受风化后，其化学成分含量会与风化前有较大的区别；
- 在对文物表面的化学成分含量进行分析时，分析的结果不会出现数据上的差错；
- 文物不同部位颜色、纹饰基本与文物主要颜色、纹饰一致。

四、 符号说明

表 **1**　符号说明

符号	定义
x_i	二元 Logistic 回归的自变量,$i = 1, 2, \cdots$
y	二元 Logistic 回归的因变量
χ^2	卡方值
w	支持向量机法向量
$P_{x,y}$	Pearson 相关性系数

五、 模型建立与求解

5.1 数据预处理

本题首先进行数据的预处理。根据题意，附件表单 2 中缺失的数据表示未检测到该成分，因此不存在因数据缺失导致的异常值。其次，由于检测手段等原因，本题中将成分比例累加和介于 85% 至 105% 之间的数据视为有效数据。因此，对表单 2 中的数据进行累和处理后，可以发现编号为 15 与 17 的文物检测数据不符合要求，因此在接下来的数据分析过程中不再考虑这两组数据。

5.2 问题一的模型建立与求解

根据题意，我们将问题分为三个小问。对于第一小问，我们使用卡方独立性检验以及 Yates 校正卡方检验，来描述各个变量之间的差异性。而在第二小问中，为了分析文物是否风化与其各种类化学成分之间的关联，先使用 Pearson 相关性分析，找出了文物风化时影响较为显著的一些化学成分，而后采用二元 Logistic 回归模型，建立了各种类化学成分与是否风化之间的定量关系。在第三问中，由于数据量较少，因此我们利用评价与分级模型来进行风化前数据的预测。

5.2.1 问题一第 (1) 问模型建立与求解

以下是本题的流程图：

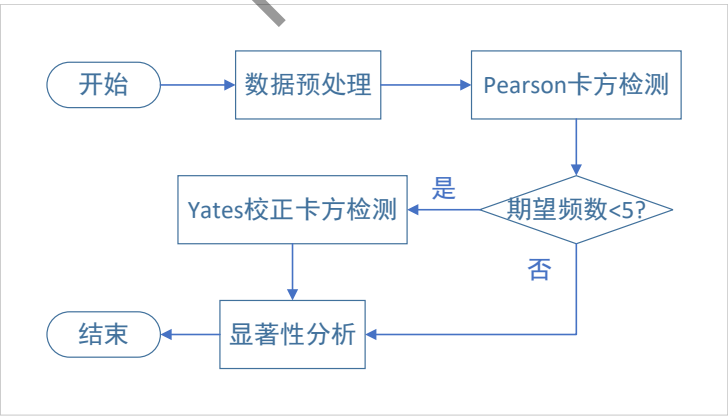

图 **1**　卡方独立性检验流程图

卡方独立性检验：由表单 1 中的数据可以看出，表面是否风化、纹饰、类型以及颜色均为定类变量，因此为了分析定类变量之间的联系，我们采用了卡方独立性检验来描述它们之间的差异性。

本文首先提出假设：H_1: 表面风化和纹饰数据不存在显著性差异。

表 **2**　卡方独立性检验结果

题目	名称	表面风化		合计	χ^2	校正 χ^2	P
		无风化	风化				
纹饰	A	11	9	20	5.747	5.747	0.056*
	B	0	6	6			
	C	13	15	28			
合计		24	30	54			
类型	铅钡	12	24	36	5.4	4.134	0.020**
	高钾	12	6	18			
合计		24	30	54			
颜色	浅绿	2	1	3	6.287	6.287	0.507
	浅蓝	8	12	20			
	深绿	3	4	7			
	深蓝	2	0	2			
	紫	2	2	4			
	绿	1	0	1			
	蓝绿	6	9	15			
	黑	0	2	2			
合计		24	30	54			
注：***、**、* 分别代表 1%、5%、10% 的显著性水平							

接着计算卡方值，自由度和显著性水平。H_1 的卡方值的计算公式如 (1) 所示。

$$\chi^2 = \sum_{i=1}^{6} \frac{(O_i - E_i)^2}{E_i} \tag{1}$$

其中 O_i 为附件表单 1 中纹饰种类的观测值，E_i 为纹饰种类在与表面是否风化独立假设下的期望值。卡方检验的原理是：通过计算得到卡方值后，再根据自由度得到相应的 P 值，与显著性水平对比后，判断是否拒绝原假设。同理，可以分别提出下列假设：

H_2：表面风化和玻璃类型数据不存在显著性差异。

H_3：表面风化和颜色数据不存在显著性差异。

通过计算 P 值来判断是否接受或拒绝原假设，结果详见表 (2)。在计算类型因素的卡方值时，出现了期望频数小于 5 的情况，为了避免过度拒绝原假设，我们使用了 **Yates 校正卡方检测**对其重新进行计算。

根据表 (2) 所示的结果，当显著性水平 α=0.05 时：

基于表面风化和纹饰，显著性 P 值为 0.056，水平上不呈现显著性，接受假设 H_1，

因此表面风化和纹饰数据不存在显著性差异。

基于表面风化和玻璃类型，显著性 P 值为 0.020，水平上呈现显著性，拒绝假设 H_2，因此表面风化和玻璃类型数据存在显著性差异。

基于表面风化和颜色，显著性 P 值为 0.507，水平上不呈现显著性，接受假设 H_3，因此表面风化和颜色数据不存在显著性差异。

因此，玻璃文物表面是否风化与其玻璃类型有**强相关性**，与纹饰和颜色的相关性则**较弱**。

5.2.2 问题一第 (2) 问模型建立与求解

考虑到是否风化为定类变量，而文物中各化学成分含量为定量变量，因此可以考虑使用二元 Logistic 回归模型。

Part 1: 数据预处理

在建立二元 Logistic 回归模型前，由于化学成分个数较多，一些与是否风化无较大关联的成分会影响后续的拟合，因此先进行数据预处理，也即分析各种化学成分与是否风化之间的相关性，采用 Pearson 相关系数，计算公式如下：

$$P_{x,y} = \frac{cov(x,y)}{\sigma_x \sigma_y} = \frac{E[(x - x_i)(y - y_i)]}{\sigma_x \sigma_y} \tag{2}$$

其中 x 分别代表 14 中化学成分，y 代表是否风化 (风化为 1，未风化为 0)。

计算结果见下表 (由于结果过多，此处仅列举相关性较高的几种化学成分，其余见附录)。

表 **3**　高钾玻璃中化学成分与是否风化的相关性

化学成分	二氧化硅	氧化钾	氧化钙	氧化镁	氧化铝
相关性	0.871	-0.803	-0.654	-0.601	-0.739

表 **4**　铅钡玻璃中化学成分与是否风化的相关性

化学成分	二氧化硅	氧化钠	氧化钙	氧化铅	五氧化二磷
相关性	-0.804	-0.408	0.424	0.716	0.545

从上方的相关性表格中可以看出，在高钾玻璃中，除去玻璃的主成分二氧化硅之外，对是否风化影响最显著的是氧化钾的含量，而在铅钡玻璃中则是氧化铅。

Part 2：建立二元 Logistic 回归模型

二元 Logistic 回归分析的模型为：

$$\ln(\frac{y}{1-y}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m \tag{3}$$

其中 $\beta_0, \beta_1, \cdots, \beta_m$ 为偏回归系数，与 $x_1, x_2, \cdots, x_m$ 无相关性。

Part 3: 求解回归系数并进行分析

依据附件表单 2 中的数据，依照表 (3) 和 (4) 中的顺序，记高钾玻璃中二氧化硅、氧化钾、氧化钙、氧化镁、氧化铝的含量分别为 $x_{11}, x_{12}, x_{13}, x_{14}$ 和 x_{15}，铅钡玻璃中二氧化硅、氧化钠、氧化钙、氧化铅和五氧化二磷的含量分别为 $x_{21}, x_{22}, x_{23}, x_{24}$ 和 x_{25}。y 为表示是否风化的二分类变量，其中 1 表示风化，0 表示未风化。那么 Logistic 回归模型可以表示为：

$$\begin{aligned} \ln(\frac{y_1}{1-y_1}) =& \beta_{10} + \beta_{11} x_{11} + \beta_{12} x_{12} + \cdots + \beta_{15} x_{15} \\ \ln(\frac{y_2}{1-y_2}) =& \beta_{20} + \beta_{21} x_{21} + \beta_{22} x_{22} + \cdots + \beta_{25} x_{25} \end{aligned} \tag{4}$$

通过使用 SPSS 进行求解，得到了回归系数分析表 (5) 和 (6)。

表 **5**　高钾玻璃回归系数分析表

变量	B	标准误差	显著性
x_{11}	2.982	2213.854	0.999
x_{12}	-1.020	2228.127	1.000
x_{13}	8.519	7232.835	0.999
x_{14}	20.704	14783.333	0.999
x_{15}	-10.733	5498.378	0.998
常量	-246.021	215793.670	0.999

表 **6**　铅钡玻璃回归系数分析表

变量	B	标准误差	显著性
x_{21}	-0.126	0.079	0.910
x_{22}	0.107	0.518	0.837
x_{23}	-0.375	0.467	0.421
x_{24}	0.131	0.075	0.781
x_{25}	0.392	0.256	0.826
常量	0.110	4.446	0.980

因此，我们得到的两个二元 Logistic 回归模型可以表示为

$$\begin{aligned}\ln(\frac{y_1}{1-y_1}) &= -246.021 + 2.982x_{11} - 1.020x_{12} + 8.519x_{13} + 20.704x_{14} - 10.733x_{15} \\ \ln(\frac{y_2}{1-y_2}) &= 0.110 - 0.126x_{21} + 0.107x_{22} - 0.375x_{23} + 0.131x_{24} + 0.392x_{25}\end{aligned} \tag{5}$$

Part 4: 误差分析

利用 SPSS，可以得到回归方程的各项显著性指标，如表 (7) 和 (8) 所示。

表 **7**　高钾玻璃回归方程显著性指标

内戈尔科 R 方	霍斯默-莱梅肖显著性	总体正确百分比
0.999	0.998	100

表 **8**　铅钡玻璃回归方程显著性指标

内戈尔科 R 方	霍斯默-莱梅肖显著性	总体正确百分比
0.843	0.923	98.0

从表中的结果可以得出，通过二元 Logistic 回归拟合的模型 R^2 均在 0.8 以上，显著性在 0.9 以上，总体正确百分比超过 95%，可以认为模型的拟合度均较好。

5.2.3 问题一第 (3) 问模型建立与求解

Part 1：分级综合评价指标体系的构建

基于前两小问所得结果，首先筛选出所需的能够反映相似度的特征变量，为后续评价模型的指标体系的构建做准备。

Step 1：对于高钾和铅钡玻璃而言，由于从制作工艺、化学结构等角度，两种玻璃均存在较大差异，因此它们的风化过程几乎不同，所以我们不能用一个未风化的铅钡玻璃的化学成分去预测高钾玻璃风化前的化学成分。因此，在问题解决初始，我们先将玻璃划分为两大类：高钾和铅钡，下文的模型与算法均是基于两种不同类型的玻璃。

Step 2：根据前两小问所进行的文物化学成分含量与是否风化的相关性分析，综合了 Pearson 相关系数以及 Logistic 回归的结果，对于高钾和铅钡玻璃，我们需要选择与风化相关性较低的成分，并且由于某些相关性较低的化学成分含量极低（如氧化锡等），因此我们不选择这些成分。于是分别筛选出如下化学成分作为主要指标，见表 (9) 和 (10).

表 **9**　高钾玻璃指标筛选

化学成分	氧化锡	氧化铜	氧化钠	二氧化硫	氧化钡	氧化铅	五氧化二磷
相关性	-0.171	-0.290	-0.310	-0.314	-0.345	-0.388	-0.425
化学成分	氧化锶	氧化铁	氧化镁	氧化钙	氧化铝	氧化钾	二氧化硅
相关性	-0.461	-0.516	-0.601	-0.654	-0.739	-0.803	0.871

表 **10**　铅钡玻璃指标筛选

化学成分	氧化镁	氧化锡	氧化铁	氧化钾	氧化钡	氧化铜	二氧化硫
相关性	0.008	0.052	-0.081	-0.156	0.170	0.172	0.194
化学成分	氧化铝	氧化锶	氧化钠	氧化钙	五氧化二磷	氧化铅	二氧化硅
相关性	-0.249	0.287	-0.408	0.424	0.545	0.716	-0.804

同时根据第一问所研究的三个指标，由于已经使用了类型指标，我们对颜色和纹路两个特征进行选择，从而近似反映文物的用途、工艺等外在特点。

Step 3：由于化学成分的相关系数在我们的评价体系中是逆向化指标，因此我们首先要对其进行正向化处理。

$$a_i' = \frac{1}{a_i},\ i = 1, 2, 3, 4$$

Step 4：由于颜色和纹路是两个定类指标，因此我们需要对它们做定量处理。为此，我们建立如下映射关系：

（1）纹路：共有三种纹饰 A,B,C，我们定义 $y_{ij} = 1$，若 i 与 j 纹饰相同，反之则 $y_{ij} = 0$。其中 i 为待检测的风化样品，j 为未风化样品，且 y_{ij} 为正向化指标。

（2）颜色：根据颜色深浅以及相似程度，我们对七种颜色做出排序：黑、紫、深蓝、浅蓝、蓝绿、深绿、浅绿。接着定义距离 $D_1(i, j)$ 为两种文物颜色在上述排序中的距离，为了将 $D_1(i, j)$ 化为正向化指标，可以定义 $Z_{ij} = 7 - D_1(i, j)$。

2. 指标体系的初步建立基于上述指标的遴选分析以及可以用来反映我们评价体系中的权重的相关性系数，以一个高钾玻璃为例，一级指标是它的用途 R_1 与类型 R_2, 而 R_1 可以进一步分为纹饰与颜色两个二级指标，R_2 分为氧化铁、氧化铜、五氧化二磷、氧化镁含量这四个二级指标。因此我们文物相似性评价指标体系构架为：

Part 2：基于综合评价的文物风化过程相似性分级模型的建立

1. 基于相关系数确定评价指标权重

类型的二级指标间的权重：$w_i = a_i' / \sum a_i', i = 1, 2, 3, 4$

用途的二级指标间的权重：w_1', w_2'

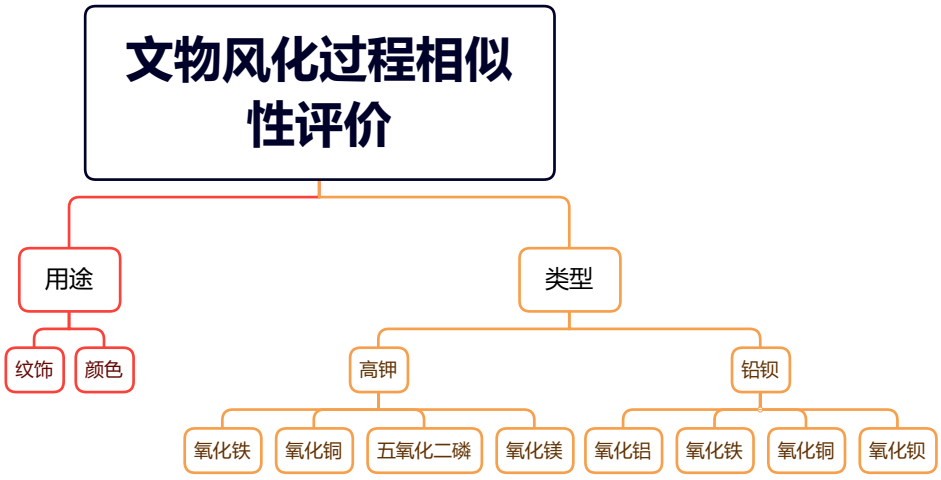

图 **2**　文物风化过程相似性评价指标体系

两个一级指标间的权重，经查阅文献可知类型更加重要，因此记用途权重为 W_1，类型权重为 W_2, 且满足：$\frac{W_2}{W_1} = C(C > 1)$，本题中取 C=2。

2. 综合评价模型的建立

文物相似性综合评价模型是通过一定的数学模型以及假设经验，将多个特征的评价值，按照一定的定量评价标准，线性加权得到一个综合评价值，通过认为选择的分级方式，实现对文物相似性的综合评价模型。同时设 d_{ij} 为类型的 2 级指标所得的评价值，那么可以定义为待检测样品与未风化样品间该成分含量之差的绝对值，其中 i 为高钾或铅钡类型下的风化样本数，$j = 1, 2, 3, 4$。记 $D_1(i, j)$ 为颜色的评价值，$D_2(i, j)$ 为纹饰的评价值。那么二级指标的综合评价值可以表示为：

$$g(x_i) = w_1 d_{i1} + w_2 d_{i2} + w_3 d_{i3} + w_4 d_{i4},$$
$$h(x_i) = w_1^{'} D_1 + w_2^{'} D_2,$$

其中 x_i 为第 i 个待检测样品，$i = 1, 2, \cdots, N$。

对于一级指标类型与用途，我们规定 $g(x_i)$ 为类型，$h(x_i)$ 为用途，而最终的综合评价值 Y，可以由 $Y = f(g(x_i), h(x_i)) = W_1 g(x_i) + W_2 h(x_i)$ 来表示.

由文物相似性分级综合评价指标所得的评价值以相应的权重系数来加权，将其求和，作为最终的综合评价值，这样得到的评价值大小符合上下文模型假设，也具有一定的合理性。而综合评价值 Y 的大小与待预测风化样品和非风化样品之间的相似度高低呈正相关关系，这样就可以根据 Y 值，进行分级阶梯模型的建立。

Part 3：两种类型的未风化样品的分级阶梯模型的建立

文物相似性阶梯模型标准设定如表 (11) 所示：

Part 4: 模型求解

利用 EXCEL，我们对数据进行了处理与计算，得到了每个待预测风化样品与未风

表 **11**　分级阶梯模型标准设定

评价值 Y	[500,∞）	[100,500)	[50,100)	[25,50)	[10,25)	[0,10)
等级 T	6	5	4	3	2	1

化样品相关性的最终综合评价值 Y，以 7 号样品为例：

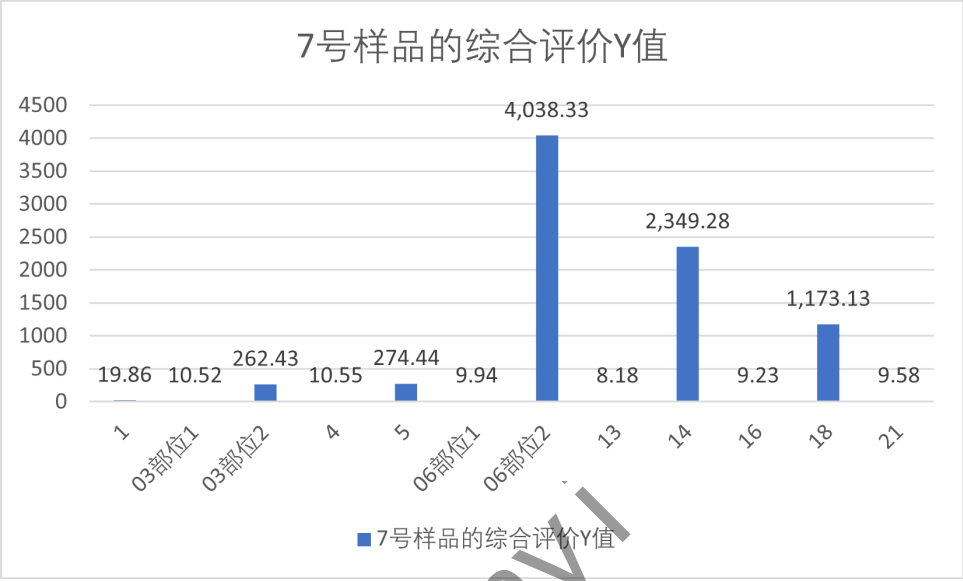

图 **3**　**7** 号样品的综合评价 **Y** 值

综合评价值 Y	19.86209	10.51969	262.4271	10.54858	274.4431	9.942397
等级 T	2	2	5	2	5	1
综合评价值 Y	4038.326	8.180939	2349.284	9.229809	1173.132	9.581381
等级 T	6	1	6	1	6	1

选择等级为 5 和 6 的样本并按照 Y 值对 14 种化学成分进行加权，即为最终结果。其余风化样本同理可进行计算。由于数据量较多，故仅列出 7 号样品的预测值，见表 (12)，其余见附录。

5.3 问题二的模型建立与求解

5.3.1 数据预处理

为了避免某些与玻璃类型无关的化学成分对后续分析造成影响，首先进行 Pearson 相关性分析，与问题 1 类似，可以得到下面的相关系数表 (同样，由于数据过多，在此仅列举相关性较大的几种成分)。

表 **12**　**7** 号样品风化前化学成分预测值

二氧化硅	氧化钠	氧化钾	氧化钙	氧化镁	氧化铝	氧化铁
65.35431	1.133981	9.910183	4.028612	1.115644	8.59288	2.15812
氧化铜	氧化铅	氧化钡	五氧化二磷	氧化锶	氧化锡	二氧化硫
2.878202	0.123779	0.688214	2.501209	0.060653	0	0.012201

表 **13**　化学成分与玻璃类型的相关性

化学成分	二氧化硅	氧化钾	氧化钙	氧化铅	氧化钡
相关性	-0.513	-0.890	-0.688	0.845	0.651

由于风化会对文物的各种化学成分含量都造成较大的影响，因此，为了玻璃类型分类的准确性，本小题会将风化文物与未风化文物分开处理。

5.3.2 支持向量机模型建立

Part 1：合并变量

由文献 [1] 及题意可知，氧化钾含量较高的文物中氧化钙含量通常也较高，而氧化铅与氧化钡也有相同的规律，这一点在 Pearson 相关系数表中也有体现，经计算，氧化钾与氧化钙的相关系数达到 0.767，而氧化铅和氧化钡则是 0.582，均具有**较强的正相关性**，因此可以将四个变量合成为**钾钙综合含量**与**铅钡综合含量**，这里采用加权的方式，权重则由各变量与玻璃类型的相关性绝对值所决定。

假设合成的钾钙综合含量为 p_1，铅钡综合含量为 p_2，氧化钾，氧化钙，氧化铅和氧化钡的含量分别为 z_1, z_2, z_3, z_4，对应的权重为 w_1, w_2, w_3, w_4，则有如下计算公式：

$$
\begin{aligned}
p_1 =& w_1 z_1 + w_2 z_2 \\
p_2 =& w_3 z_3 + w_4 z_4 \\
w_i =& \frac{P_i}{P_1 + P_2} \quad i = 1, 2 \\
w_i =& \frac{P_i}{P_3 + P_4} \quad i = 3, 4
\end{aligned}
\tag{6}
$$

其中，$P_i (i = 1, 2, 3, 4)$ 分别为四种化学成分与玻璃类型之间的 Pearson 相关性绝对值。

由于所给数据集数据规模较小的特点，我们考虑使用支持向量机 (SVM) 算法对两种玻璃类型进行二分类。支持向量机算法在解决非线性、小样本以及高维模式识别中具有很多优势。此外它也是一个凸二次优化问题，所得到的极值解同时也是最优解。根据前文特征选择的结果，我们对表单 2 中的未风化的 37 个样本，以及 3 个特征（即 3 维），

利用支持向量机算法对高钾与钠钡两种玻璃类型进行有监督的学习，同时将样本进行交叉验证，提高学习的准确率，防止过拟合。

Part 2：线性支持向量机模型介绍

1. 线性分类模型

给定一个两类分类器数据集 $D=\{(x^{(n)},y^{(n)})\}_{n=1}^{N}$，其中 $y_n\in\{+1,-1\}$，如果两类样本是线性可分的，即存在一个超平面

$$w^{\mathrm{T}}x+b=0$$

将两类样本分开，那么对于每个样本都有 $y^{(n)}(w^{\mathrm{T}}x^{(n)}+b)>0$。

2. 间隔

间隔用来描述分割超平面与样本之间的距离。利用空间中点到平面的距离公式，我们可以得到，数据集中每个样本点 $x^{(n)}$ 到分割超平面之间的距离为

$$\gamma^{(n)}=\frac{||w^{\mathrm{T}}x^{(n)}+b||}{||w||}=\frac{y^{(n)}(w^{\mathrm{T}}x^{(n)}+b)}{||w||},$$

而间隔就被定义为整个数据集中所有样本点到分割超平面之间的最短距离。所求的间隔越大，其分割超平面对两个数据集的划分越稳定，也能够避免噪声等因素的影响。而支持向量机的目标就是希望找到这样一个超平面，距离每个样本最远，也即最大化

$$M=\frac{2}{||w||}.$$

3. 模型设计

基于上述的问题，在添加限制条件后，可以得到如下的拉格朗日优化问题：

$$\min_{w,b}\ \frac{1}{2}||w||^2$$

$$\text{s.t. } y_i(w^{\mathrm{T}}x_i+b)\geqslant 1,\ i=1,\cdots,n.$$

根据拉格朗日定理，得到该优化问题的对偶形式：

$$\min_{\alpha}\ \psi(\alpha)=\min_{\alpha}\ \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}y_iy_j(x_ix_j)\alpha_i\alpha_j-\sum_{i=1}^{N}\alpha_i$$

$$\text{s.t. } \alpha_i\geqslant 0,\ \forall i,$$

$$\sum_{i=1}^{N}y_i\alpha_i=0.$$

而当训练集中的样本在特征空间中不是线性可分时，就无法找到最优解。为此，我们引入松弛变量 ξ，同时加入惩罚系数 C，得到：

$$\min_{\mathbf{w},b,\xi},\ \frac{1}{2}||\mathbf{w}||^2+C\sum_{i=1}^{N}\xi_i$$

$$\text{s.t. } y_i(\mathbf{w}x_i+b)\geqslant 1-\xi_i,\forall\, i.$$

同时引入核函数 $K(x_i, x_j)$，得到下式：

$$\min_{\alpha} \ \psi(\alpha) = \min_{\alpha} \ \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} y_i y_j K(x_i, x_j)\alpha_i\alpha_j - \sum_{i=1}^{N}\alpha_i$$

$$\text{s.t. } 0 \leqslant \alpha_i \leqslant C, \forall\, i,$$

$$\sum_{i=1}^{N} y_i\alpha_i = 0.$$

而 SMO 算法可用来求解该式所描述的二次规划问题，将整个二次规划问题分解成多个子问题求解。

Part 3: 模型求解

在本题中，设 $w = (w_1, w_2, w_3)$，二氧化硅的含量为 p_0，则超平面可以表示为

$$w_1 p_0 + w_2 p_1 + w_3 p_2 + b = 0 \tag{7}$$

将附件 2 中的未风化文物数据集代入求解，得到超平面的系数如表 (14) 所示。

表 **14**　未风化文物超平面系数

系数	w_1	w_2	w_3	b
值	-0.6323	-1.307	0.538	43.505

同理，可以得到风化文物数据集对应的超平面系数 (15)。

表 **15**　风化文物超平面系数

系数	w_1	w_2	w_3	b
值	-0.047	0.0003	0.0144	3.323

故两个分类超平面可以分别表示为

$$-0.6323x_1 - 1.307x_2 + 0.538x_3 + 43.505 = 0$$

$$-0.047x_1 + 0.0003x_2 + 0.0144x_3 + 3.323 = 0$$

Part 4：模型准确性检验及结果解释

由几何知识可知，分类超平面的法向量为 $w = (w_1, w_2, w_3)$，因此从得到的超平面方程 (5 . 3 . 2) 可知，**二氧化硅及钾钙综合含量较高的**通常为**高钾玻璃**，而**二氧化硅含量较低，铅钡综合含量较高的**则为**铅钡玻璃**，通过查阅文献 [3]，发现模型得到的分类规律与现实情况基本相同，可以认为该模型得出的分类平面可信度较高。

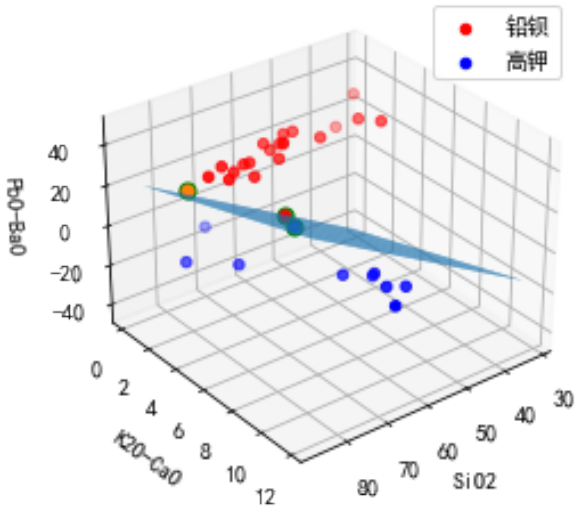

图 **4**　未风化超平面

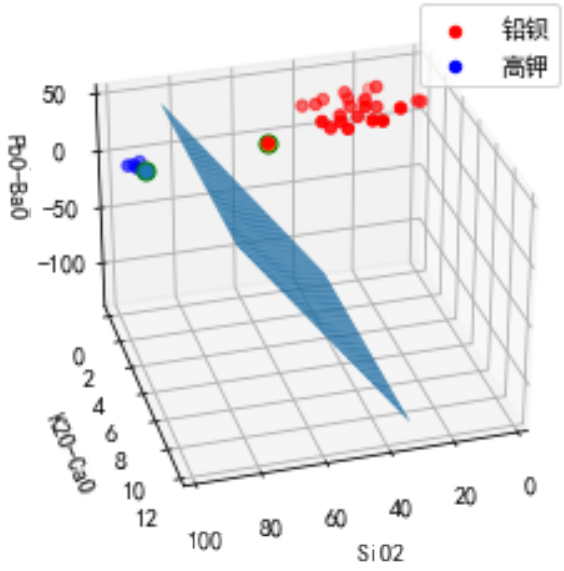

图 **5**　风化超平面

5.3.3 亚类划分

聚类算法是一种区别于分类的无监督学习算法，即对无标签的样本点根据数据及数据间的信息关系，对数据对象进行分组。聚类的最终目的是使得组内的对象是相似的，而不同组中的对象之间则是有区别的。[2]

Part 1：K-均值聚类算法

K-均值聚类算法是一种经典的聚类算法，原理是利用距离描述数据集内各个数据点的“相似度”，将数据集分成 K 个类簇。K-均值聚类本质上是最小化类簇内部各点的距离和，当 K 值增大，距离和会减小。但当到达最优的 k 值时，距离和的降幅就会变小。基于这样的经验性准则，我们可以用图 (6) 和 (7) 中的方法来进行 K 值的选取。因此，我们选取 K=2。

在本题中，我们采取常用的欧几里得距离，分别对高钾和铅钡两类玻璃进行亚类划分，迭代聚类 10 次，选取分类效果最好的情形，即亚类数据之间距离最大的聚类结果作为最终结果，如图 (8) 和 (9) 所示。

通过 K-均值聚类算法，我们将高钾玻璃进一步分为高钾高钙与低钾低钙两种亚类，其中高钾高钙玻璃氧化钾平均含量在 10% 以上，氧化钙平均含量在 7% 以上；铅钡玻璃则进一步分为高铅与低铅两种亚类，其中高铅玻璃氧化铅平均含量在 30% 以上，而两种铅钡亚类玻璃的氧化钡含量都在 10% 左右，这个分类结果与文献 [3] 中给出的亚类划分方法十分相近，证明了分类标准的合理性。

Part 2: 层次聚类法

层次聚类是由下而上的一种聚类方法。首先仅以一个样本为一类构造 n 个类簇，标记该类的平台高度为 0。接着合并距离最近的两个类为新类簇，标记该类的平台高度为

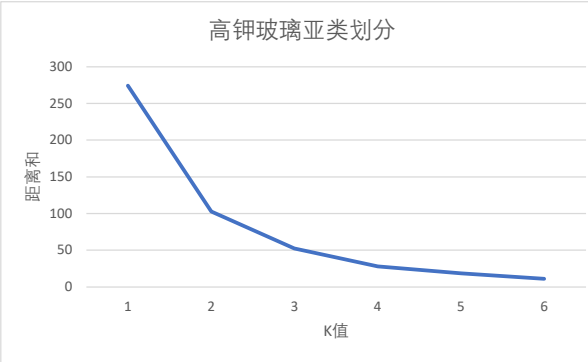

图 6　高钾亚类不同 K 值划分结果

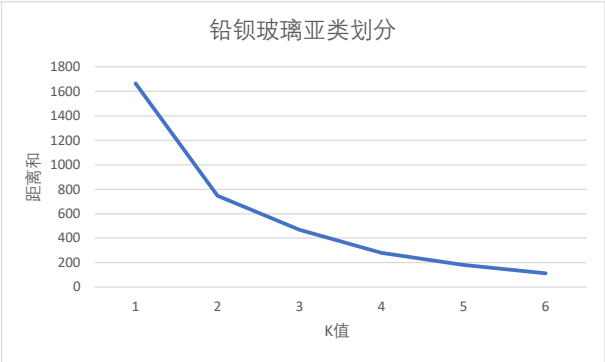

图 7　铅钡亚类不同 K 值划分结果

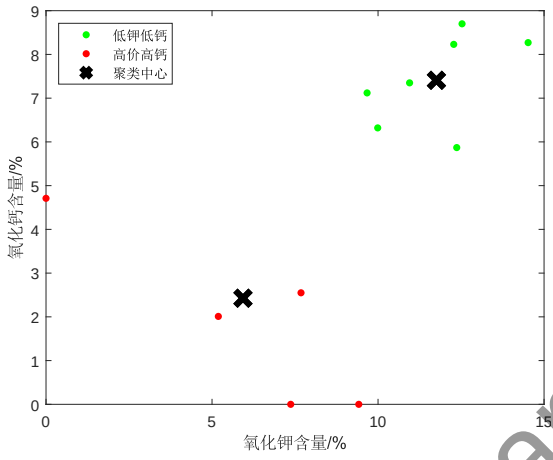

图 8　高钾亚类划分结果

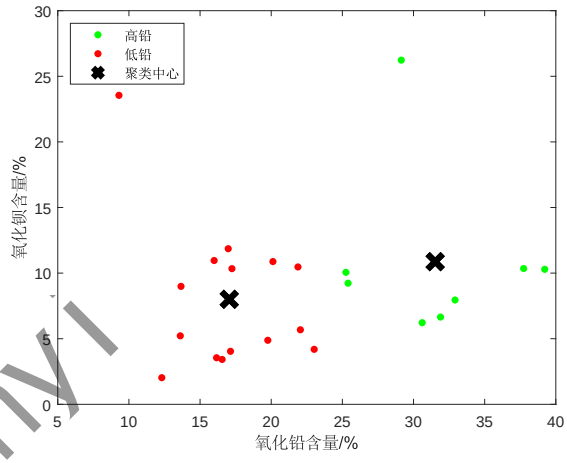

图 9　铅钡亚类划分结果

1。仿此最终得到一个最大的类，其下可分为若干类簇，实现样本的层次聚类。

本题中我们选取 p=2 时的闵氏距离，利用层次聚类方法分别对高钾和铅钡两类进行亚类划分，以此对我们制定的分类标准进行敏感性分析。

高钾玻璃亚类划分的层次聚类中出现了噪声 21，源数据中未检出其氧化钾含量，根据其氧化钙含量可推测其属于低钾低钙类。剔除噪声 21 后，其余的分类结果详见表 (16)，除了 06 部分 2 被错分到高钾高钙之外，其余数据均符合我们制定的分类标准，两组聚类结果之间仅有 8% 的差异。同样地，我们对铅钡玻璃的亚类划分也进行了敏感性分析，具体结果可见附录，最终得到两种聚类方法所得类别之间的差异在 10% 左右，分类结果之间较小的差异表明分类标准对分类方法的敏感性较低。

5.4 问题三的模型建立与求解

5.4.1 鉴别未知文物类型

在问题二已经得到了高钾玻璃和铅钡玻璃划分规律的基础上，我们将各个文物的化学成分含量代入到分类超平面中，并根据问题二中的亚类划分标准进行进一步的划分，

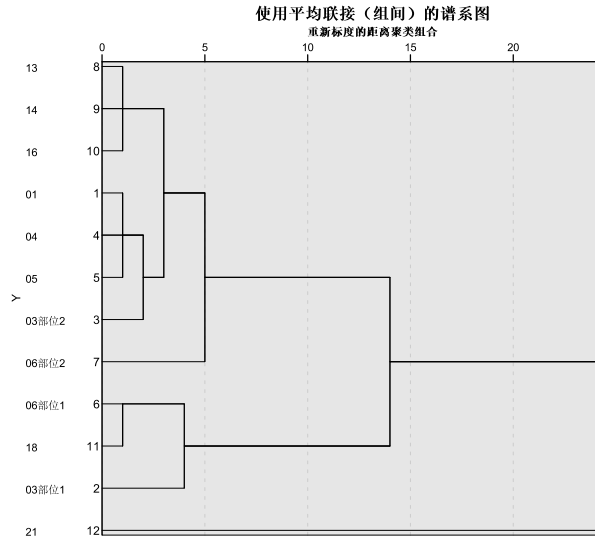

图 **10**　层次聚类高钾亚类划分结果

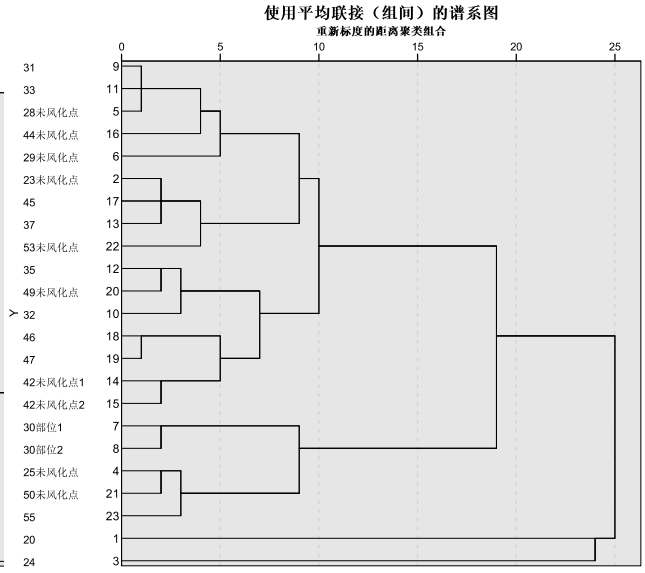

图 **11**　层次聚类铅钡亚类划分结果

表 **16**　高钾亚类划分敏感度分析表

文物采样点	21	03 部位 1	06 部位 1	06 部位 2	18	4
氧化钾 (K2O)	0	5.19	7.37	7.68	9.42	9.67
氧化钙 (CaO)	4.71	2.01	0	5.41	0	7.12
K-means 聚类类别	2	2	2	2	2	1
层次聚类类别	噪声	2	2	1	2	1
文物采样点	1	5	14	03 部位 2	13	16
氧化钾 (K2O)	9.99	10.95	12.28	12.37	12.53	14.52
氧化钙 (CaO)	6.32	7.35	8.23	5.87	8.7	8.27
K-means 聚类类别	1	1	1	1	1	1
层次聚类类别	1	1	1	1	1	1

最终得到预测结果，如表 (17) 所示。

5.4.2 敏感度分析

由支持向量机的特点可知，法向量大小对于最终的分类结果有着较大的影响，因此，我们对其中法向量大小进行敏感度分析，使其值增加 1% 或是 5%，得到相应的分类情况，如表 (18) 所示。

法向量的大小改变基本没有引起最终分类结果的变化，说明最终分类结果对于模型参数的敏感度较低，不会因为参数的微小改变而改变，较为稳定。

表 **17**　问题三的预测结果

文物编号	A1	A2	A3	A4	A5	A6	A7	A8
SiO2	78.45	37.75	31.95	35.47	64.29	93.17	90.83	51.12
氧化钾	0	0	1.36	0.79	0.37	1.35	0.98	0.23
氧化钙	6.08	7.63	7.19	2.89	1.64	0.64	1.12	0.89
氧化铅	0	34.3	39.58	24.28	12.23	0	0	21.24
氧化钡	0	0	4.69	8.31	2.16	0	0	11.34
钾钙综合	2.6752	3.3572	3.9252	1.714	0.9288	1.0376	1.0416	0.5204
铅钡综合	0	19.208	24.2284	17.2532	7.7992	0	0	16.884
预测类别	铅钡玻璃	高钾玻璃	高钾玻璃	高钾玻璃	高钾玻璃	铅钡玻璃	铅钡玻璃	高钾玻璃
亚类划分	低钾低钙	高铅	低钾低钙	低铅	低铅	低钾低钙	低钾低钙	低铅

表 **18**　问题三敏感度分析

编号	原结果	$\Delta w_1 = 1\%$	$\Delta w_1 = 5\%$	$\Delta w_2 = 1\%$	$\Delta w_2 = 5\%$	$\Delta w_3 = 1\%$	$\Delta w_3 = 5\%$
A1	铅钡	铅钡	铅钡	铅钡	铅钡	铅钡	铅钡
A2	铅钡	铅钡	铅钡	铅钡	铅钡	铅钡	铅钡
A3	高钾	高钾	高钾	高钾	高钾	高钾	高钾
A4	高钾	高钾	高钾	高钾	高钾	高钾	高钾
A5	铅钡	铅钡	铅钡	铅钡	铅钡	铅钡	铅钡
A6	高钾	高钾	高钾	高钾	高钾	高钾	高钾
A7	高钾	高钾	高钾	高钾	高钾	高钾	高钾
A8	高钾	高钾	高钾	高钾	高钾	高钾	高钾

5.5 问题四的模型建立与求解

5.5.1 多元线性回归模型的建立

多元线性回归分析的模型为：

$$\begin{cases} y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m + \varepsilon,\ i = 1, 2, \cdots, n \\ \varepsilon \sim N(0, \sigma^2), \end{cases} \tag{8}$$

其中 $\beta_0, \beta_1, \cdots, \beta_m, \varepsilon^2$ 为偏回归系数，与 $x_1, x_2, \cdots, x_m$ 无相关性，ε 为随机误差项。

Pearson 相关性分析表明高钾玻璃中，二氧化硅与氧化铝，氧化铁、氧化锶、五氧化二磷、氧化钾、氧化钙以及氧化镁彼此之间的相关性显著，因此将氧化铝，氧化铁、氧化锶、五氧化二磷、氧化钾、氧化钙以及氧化镁依次记为 x_1 至 x_7，y 为二氧化硅。

假设因变量与自变量之间存在线性关系，那么总体线性回归模型可以表示为

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_7 x_7 + \varepsilon \tag{9}$$

其中 ε 为随机误差项 $\varepsilon \sim N(0, \sigma^2)$。

通过 SPSS 软件求解多元线性规划模型，得到拟合优度 $R^2 = 0.959$，表明线性回归效果较好，验证了 Pearson 相关性的结论，即这些化学成分含量之间存在显著关联性。具体回归系数见下表。

变量	B	标准误差	显著性
常量	95.486	3.838	0.000
x_1	-1.294	0.819	0.189
x_2	-3.615	3.010	0.296
x_3	15.599	51.676	0.778
x_4	2.073	3.562	0.592
x_5	-1.415	0.647	0.094
x_6	-0.113	1.756	0.952
x_7	-1.579	1.933	0.460

因此，高钾玻璃中有关二氧化硅的回归方程可以描述为

$$y = 95.486 - 1.294x_1 - 3.615x_2 + 15.599x_3 + 2.073x_4 - 1.415x_5 - 0.113x_6 - 1.579x_7 + \varepsilon$$

在铅钡玻璃的主要化学成分中，同样可以得到二氧化硅 (y)、氧化铜 (x_1)、氧化铅 (x_2)、氧化钡 (x_3)、氧化锡 (x_4) 和氧化钙 (x_5) 之间的线性回归方程:

$$y = 88.379 + 0.418x_1 - 0.609x_2 - 1.767x_3 + 18.684x_4 - 4.409x_5 + \varepsilon$$

比较两个方程的自变量，其关联主要成分除了氧化钙之外，全部互不相同，这可能是由于两种玻璃在烧制时添加了共同的稳定剂石灰石，经煅烧后残留成为氧化钙。此外，两种玻璃制作配方、主要成分不同，各种化学组分互相的理化作用有很大差异。比如在铅钡玻璃中，氧化钙与氧化锡含量的关联性比高钾玻璃中更显著，而在高钾玻璃中有着较强关联性的五氧化二磷、氧化铁含量等却未在铅钡玻璃中表现出明显关联性。

六、 敏感度分析

针对支持向量机算法的敏感度分析，我们从以下几个方面做出分析与评价：

1. 模型的泛化能力

泛化能力指算法对新数据的预测能力。当一个模型在训练数据上得到了非常好的效果，即过拟合时，对于新数据的预测能力会变差。由于题目所给的训练集数量较少，不适合划分出一部分数据作为测试集，因此我们利用 Python 中的 sdv 库，通过 Copula 算法，新生成了一些和原样本数据具有一定相似性的含有标签的样本，并对这些新产生的样本使用支持向量机模型，得到预测的二分类结果，再与原标签进行对比。从图 (12) 中可以看出，预测的准确率较高，这说明该模型泛化能力较强。

2. 模型的鲁棒性

鲁棒性指训练好的模型对于输入扰动或对抗样本的性能。我们对数据加入些许小扰动，观察模型的分类效果。利用 Python 语言，根据样本数据分别生成一些正态分布的随机噪声，以二氧化硅为例，从图 (13) 中可以看出，模型对于噪声数据的分类情况与原数据基本相同，由此可见模型的鲁棒性较好。

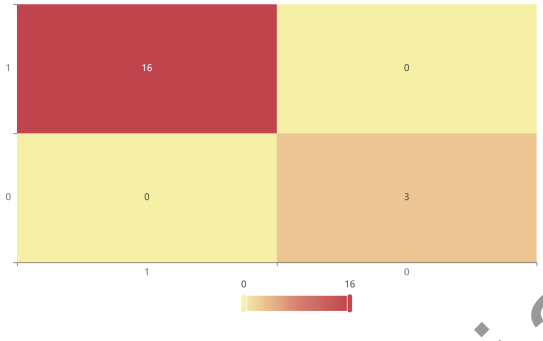

图 **12**　预测准确率

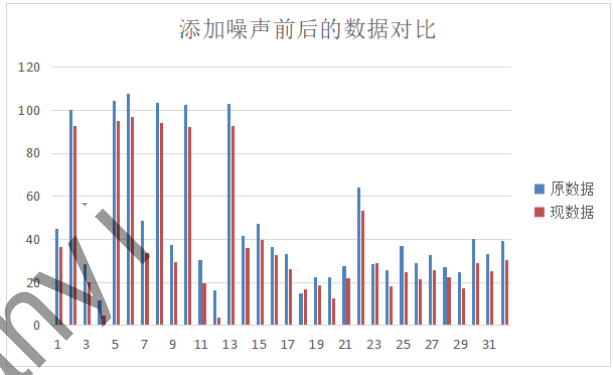

图 **13**　添加噪声前后的数据对比

3. 模型的参数输入

通过改变模型的输入参数，观察分类情况。这里我们改变模型中的惩罚参数 C，C 越大，对误分类的惩罚越大，模型会倾向于将训练集全部正确分类，这会导致训练集测试时准确率很高，但泛化能力较弱。C 值减小，对误分类的惩罚减小，模型会允许对训练集的分类出现一些错误，加强泛化能力。

以风化类型的训练为例，改变 C 值分别为 0.5,1,1.5,50，得到的超平面系数几乎没有发生变化，说明模型能够很好的收敛到最优解。

七、 模型的评价

7.1 模型的优点

- 在使用 Logistic 回归及支持向量机前，先通过 Pearson 相关性分析，去除了对因变量影响较小的因素，避免其对后续的拟合及分类造成干扰。
- 使用支持向量机进行分类，并进行了敏感性分析，得到的模型可信度较高，且与实际情况高度符合。

表 **19**　改变惩罚参数时得到的超平面系数

C	w_1	w_2	w_3	b
0.5	-0.6321	-1.305	0.537	43.501
1	-0.6323	-1.307	0.538	43.505
1.5	-0.6325	-1.306	0.538	43.507
50	-0.6328	-1.305	0.538	43.507

- 在探究文物玻璃类型划分规律时，考虑到了风化会对文物表面的化学成分含量造成较大的影响，因此对风化文物和未风化文物采取了分别处理的方式。

7.2 模型的缺点

- 对数量较少的个案采用了二元 Logistic 回归和支持向量机，可能会导致过拟合。

参考文献

[1] 王枫云, 陈亚楠. 古代丝绸之路 (中国段) 沿线城镇兴衰的内在机理及其启示 [J]. 西南民族大学学报 (人文社科版),2018,39(9):206-213.

[2] PANG N T, MICHANEL S, VIPIN K. Introduction to data mining[M]. 范明, 范宏建,, 译. 2 版. 北京: 人民邮电出版社, 2011.

[3] 伏修锋, 干福熹. 基于多元统计分析方法对一批中国南方和西南地区的古玻璃成分的研究 [J]. 文物保护与考古科学,2006(04):6-13.DOI:10.16334/j.cnki.cn31-1652/k.2006.04.002.

[4] Platt J. Sequential minimal optimization: A fast algorithm for training support vector machines[J]. 1998.

附录 A　高钾玻璃中化学成分与是否风化的相关性总表

化学成分	二氧化硅	氧化钠	氧化钾	氧化钙	氧化镁	氧化铝	氧化铁
相关性	.871**	-0.310	-.803**	-.654**	-.601**	-.739**	-.516*
化学成分	氧化铜	氧化铅	氧化钡	五氧化二磷	氧化锶	氧化锡	二氧化硫
相关性	-0.290	-0.388	-0.345	-0.425	-0.461	-0.171	-0.314

附录 B　铅钡玻璃中化学成分与是否风化的相关性总表

化学成分	二氧化硅	氧化钠	氧化钾	氧化钙	氧化镁	氧化铝	氧化铁
相关性	-.804**	-.408**	-0.156	.424**	0.008	-0.249	-0.081
化学成分	氧化铜	氧化铅	氧化钡	五氧化二磷	氧化锶	氧化锡	二氧化硫
相关性	0.172	.716**	0.170	.545**	.287*	0.052	0.194

附录 C　化学成分与类型的相关性总表

化学成分	二氧化硅	氧化钠	氧化钾	氧化钙	氧化镁	氧化铝	氧化铁
相关性	-.513**	0.227	-.890**	-.688**	-.340*	-0.330	-.398*
化学成分	氧化铜	氧化铅	氧化钡	五氧化二磷	氧化锶	氧化锡	二氧化硫
相关性	-0.258	.845**	.651**	-0.100	.483**	-0.178	0.044

附录 D　K-均值聚类–matlab 源程序

```
clc
clear
data=xlsread('julei.csv');
x=[data(:,2),data(:,3)];
[idx,center]=kmeans(x,2,'replicate',1000,'display','final');
plot(x(idx==1,1),x(idx==1,2),'g.','MarkerSize',14)
xlabel('氧化铅含量/%');
ylabel('氧化钡含量/%');
hold on
plot(x(idx==2,1),x(idx==2,2),'r.','MarkerSize',14)
hold on
```

```
plot(x(idx==3,3),x(idx==3,2),'b.','MarkerSize',14)
for i=1:2
plot(center(:,1),center(:,2),'kx','MarkerSize',14,'LineWidth',4)
end
hold off
legend('高铅','低铅','聚类中心','Location','NW')
```

附录 E　鲁棒性 Python 源程序

```
import pandas as pd
from sklearn import svm
from sklearn.model_selection import train_test_split as ts
import matplotlib.pyplot as plt
import numpy as np
from mpl_toolkits.mplot3d import Axes3D
import matplotlib as mpl

data_robust=pd.read_excel("F:/SVM2.xlsx")
data_robust

X = data_robust.loc[:,['SiO2','K2O-CaO','PbO-BaO']]
y = data_robust.loc[:,['类型']]
# X_train, X_test, y_train, y_test = ts(
#     X, y, test_size=2, shuffle=True
# )
# X_test

X1_mean=data_robust.loc[:,['SiO2']].mean()
print(X1_mean)
X1_sigma=data_robust.loc[:,['SiO2']].std()
print(X1_sigma)

X2_mean=data_robust.loc[:,['K2O-CaO']].mean()
print(X2_mean)
X2_sigma=data_robust.loc[:,['K2O-CaO']].std()
print(X2_sigma)

X3_mean=data_robust.loc[:,['PbO-BaO']].mean()
print(X3_mean)
X3_sigma=data_robust.loc[:,['PbO-BaO']].std()
print(X3_sigma)

# x_Train=X_train.loc[:,'SiO2']
# x_Train
```

```
X.loc[:,0]=data_robust.loc[:,'SiO2']
X.loc[:,0]

noise_factor=0.05
X.loc[:,'SiO2'] += noise_factor * np.random.normal(loc=X1_mean, scale=X1_sigma,
    size=X.loc[:,'SiO2'].shape)
# X_train.loc[:,'SiO2']-data_robust.loc[:,'SiO2']
X.loc[:,'SiO2']

data_robust.loc[:,['SiO2']]

data=pd.read_excel("F:/SVM2.xlsx")
data

X1 = data.loc[:,['SiO2','K2O-CaO','PbO-BaO']]
y1 = data.loc[:,['类型']]
X_train, X_test, y_train, y_test = ts(
X1, y1, test_size=2, shuffle=True
)
X_test

y_train=y_train.values.ravel()
y_test=y_test.values.ravel()
clf_linear = svm.SVC(kernel='linear',C=1.5)
clf_linear.fit(X_train,y_train)
score_linear = clf_linear.score(X_test,y_test)
print("The score of linear is : %f"%score_linear)

X.drop(0, axis = 1)

result_robust=clf_linear.predict(X.drop(0, axis = 1))

X_array = np.array(X_train, dtype=float)
y_array = np.array(y_train, dtype=int)
pos = X_array[np.where(y_array==1)]
neg = X_array[np.where(y_array==0)]

from sklearn.metrics import accuracy_score,precision_score,\
recall_score,f1_score,cohen_kappa_score
print("accuracy_score is")
accuracy_score(result_robust,y)
```


附录 F　支持向量机预测 Python 源程序

```
import pandas as pd
```

```
from sklearn import svm
from sklearn.model_selection import train_test_split as ts
import matplotlib.pyplot as plt
import numpy as np
from mpl_toolkits.mplot3d import Axes3D
import matplotlib as mpl

data=pd.read_excel("F:/SVM2.xlsx")
data

X = data.loc[:,['SiO2','K2O-CaO','PbO-BaO']]
y = data.loc[:,['类型']]
X_train, X_test, y_train, y_test = ts(
X, y, test_size=2, shuffle=True
)
X_test



y_test

y_train=y_train.values.ravel()
y_test=y_test.values.ravel()
clf_linear = svm.SVC(kernel='linear',C=50)
clf_linear.fit(X_train,y_train)
score_linear = clf_linear.score(X_test,y_test)
print("The score of linear is : %f"%score_linear)

w = clf_linear.coef_
b = clf_linear.intercept_
print()
print(w)
print(b)

data.loc[:,['PbO-BaO']].min()

X_array = np.array(X_train, dtype=float)
y_array = np.array(y_train, dtype=int)
pos = X_array[np.where(y_array==1)]
neg = X_array[np.where(y_array==0)]


#绘图
#设置默认字体
mpl.rcParams['font.sans-serif'] = [u'SimHei']
mpl.rcParams['axes.unicode_minus'] = False
fig = plt.figure()
ax = fig.add_subplot(111, projection='3d')
```

```
#画散点图
ax.scatter(pos[:,0], pos[:,1], pos[:,2], c='r', label='pos')
ax.scatter(neg[:,0], neg[:,1], neg[:,2], c='b', label='neg')
#绘制超平面,alpha为设置平面透明度
x = np.arange(0.0,42.2612,1)
y = np.arange(0.0,2.816,0.05)
x, y = np.meshgrid(x, y)
z = (w[0,0]*x + w[0,1]*y + b) / (-w[0,2])
surf = ax.plot_surface(x, y, z,alpha=1)
# ax.plot_surface(X,Y,Z1,alpha=0.6)
# # ax.plot_surface(X,Y,Z2,alpha=0.6,)
# ax.plot_surface(X,Y,Z3,alpha=0.6)
#设置轴标签
ax.set_xlabel("特征1",fontsize=10)
ax.set_ylabel("特征2",fontsize=10)
ax.set_zlabel("特征3",fontsize=10)
#设置图例
ax.legend(loc='best')
n_Support_vector = clf_linear.n_support_ # 支持向量个数
sv_idx = clf_linear.support_ # 支持向量索引
X = np.array(X_train,dtype=float)
for i in range(len(sv_idx)):
ax.scatter(X[sv_idx[i],0], X[sv_idx[i],1], X[sv_idx[i],2],s=50,marker='o', edgecolors='g')
#保存图片
plt.savefig("Result.png")
plt.show()


from mpl_toolkits import mplot3d

# 定义三维显示的函数
def plot_3D(elev=30,azim=30,X=X,y=y):
#绘图
#设置默认字体
mpl.rcParams['font.sans-serif'] = [u'SimHei']
mpl.rcParams['axes.unicode_minus'] = False
fig = plt.figure()
ax = fig.add_subplot(111, projection='3d')
#画散点图
ax.scatter3D(pos[:,0], pos[:,1], pos[:,2], c='r', label='铅钡')
ax.scatter3D(neg[:,0], neg[:,1], neg[:,2], c='b', label='高钾')

#绘制超平面,alpha为设置平面透明度
x = np.arange(31.94,85.05,1)
y = np.arange(0.14,11.77,0.5)
x, y = np.meshgrid(x, y)
z = (w[0,0]*x + w[0,1]*y + b) / (-w[0,2])
```

```
surf = ax.plot_surface(x, y, z,alpha=1)

ax.view_init(elev=elev,azim=azim)

#设置轴标签
ax.set_xlabel("SiO2",fontsize=10)
ax.set_ylabel("K2O-CaO",fontsize=10)
ax.set_zlabel("PbO-BaO",fontsize=10)
#设置图例
ax.legend(loc='best')
n_Support_vector = clf_linear.n_support_ # 支持向量个数
sv_idx = clf_linear.support_ # 支持向量索引
X = np.array(X_train,dtype=float)
for i in range(len(sv_idx)):
ax.scatter(X[sv_idx[i],0], X[sv_idx[i],1], X[sv_idx[i],2],s=50,marker='o', edgecolors='g')

# 交互式显示三维图像：可以动态修正evel和amin
from ipywidgets import interact,fixed
interact(plot_3D,elev=[0,30,60,90], azip=(-180,180),X=fixed(X),y=fixed(y))

plt.show()

pred=pd.read_excel("F:\风化预测.xlsx")
pred

pred_data=pred.loc[:,['SiO2','K2O-CaO','PbO-BaO']]
pred_data

result=clf_linear.predict(pred_data)

result

from mlxtend.evaluate import bias_variance_decomp

X_train=np.array(X_train)
X_test=np.array(X_test)
y_train=np.array(y_train)
y_test=np.array(y_test)

mse,bias, var = bias_variance_decomp(clf_linear, X_train, y_train, X_test,
    y_test,loss='mse', num_rounds=200, random_seed=1)

mse

bias

var
```

附录 G　铅钡亚类划分敏感度分析表

文物采样点	氧化铅 (PbO)	氧化钡 (BaO)	K-means 聚类类别	层次聚类类别
30 部位 1	39.22	10.29	2	2
30 部位 2	37.74	10.35	2	2
55	32.92	7.95	2	2
25 未风化点	31.90	6.65	2	2
50 未风化点	30.61	6.22	2	2
24	29.14	26.23	2	2
47	25.40	9.23	2	1
46	25.25	10.06	2	1
49 未风化点	23.02	4.19	1	1
35	22.05	5.68	1	1
42 未风化点 1	21.88	10.47	1	1
42 未风化点 2	20.12	10.88	1	1
32	19.76	4.88	1	1
37	17.24	10.34	1	1
28 未风化点	17.14	4.04	1	1
23 未风化点	16.98	11.86	1	1
31	16.55	3.42	1	1
33	16.16	3.55	1	1
45	15.99	10.96	1	1
53 未风化点	13.66	8.99	1	1
44 未风化点	13.61	5.22	1	1
29 未风化点	12.31	2.03	1	1
20	9.30	23.55	1	2

附录 H　风化鲁棒性预测结果

附录 I　支撑材料列表

预测结果	PbO-BaO	K2O-CaO	SiO2	原标签值	是否预测正确
1	26.5608	1.6176	44.80563	1	1
0	0	0.4708	99.98897	0	1
1	29.802	0.6512	28.40864	1	1
1	31.6448	1.4036	11.81484	1	1
0	0	0.6032	104.2273	0	1
0	0	0.6076	107.5115	0	1
1	20.6468	1.662	48.64324	1	1
0	0	0.8824	103.5769	0	1
1	26.3332	1.2892	37.48312	1	1
0	0	1.1448	102.3492	0	1
1	30.7268	0.6336	30.22366	1	1
1	32.3532	1.5484	16.29641	1	1
0	0	0.4136	103.1255	0	1
1	30.468	0.4832	41.51974	1	1
1	28.0668	0.2412	47.28142	1	1
1	31.9212	0.2992	36.50958	1	1
1	37.3536	0.4884	32.96153	1	1
1	42.2612	0.8228	14.9115	1	1
1	29.0016	2.4288	22.39537	1	1
1	36.7236	2.3056	22.38254	1	1
1	26.4944	2.816	27.71435	1	1
0	12.014	1.42	64.22867	1	0
1	21.8248	2.0152	28.62666	1	1
1	30.888	1.4036	25.53166	1	1
1	26.468	1.5752	36.77061	1	1
1	28.7504	2.2572	29.11179	1	1
1	30.3568	0.9988	32.86135	1	1
1	34.1552	1.5828	27.2063	1	1
1	32.7376	0	24.56826	1	1
1	29.898	0.5324	40.40229	1	1
1	32.868	0.5764	33.28103	1	1
1	25.4064	1.726	39.05108	1	1

gaojia.csv	svm 可视化-未风化.ipynb	高钾回归.mat
julei.csv	svm 可视化-未风化.py	鲁棒性.ipynb
svm 分析-风化.ipynb	高钾 K-means 聚类.m	鲁棒性.py
svm 分析-风化.py	高钾回归.m	铅钡 K-means 聚类.m