2022年C题讲解与真题复现

主讲人：颖老师

01

赛题解析

一、赛题解析

1. 问题背景

【真题阐述】 你拥有一批古代玻璃文物的化学成分数据，文物类型已知为**“高钾玻璃”**或**“铅钡玻璃”**，另有部分为未知类型。数据来源为表面不同部位的检测点，部分文物表面已风化。数据具有成分型特征，总和应在85%~105%之间为有效。现基于这些数据完成以下四个问题：

【问题一】：首先要求以**风化与否为结果变量**，分析数据集中**定类变量之间的关系**。接下来要求分析不同类型玻璃文物风化前后化学成分的变化规律，并给出通过风化后成分数据**预测风化前成分数据的方法**。

【问题二】：首先要求分析已经给出的**高钾、铅钡两类**玻璃化学成分，给出其**分类规律**；在此基础上，对每类玻璃进行**亚类划分**，并针对分类依据、结果**做模型评估**。

【问题三】：要求**利用问题二**中建立的亚类分类模型，对未知类别的玻璃文物的化学成分进行分析，其中包括处于不同风化状态的玻璃文物。

【问题四】：要求针对不同类别的玻璃文物样品，分析其**化学成分之间的关联关系**，并比较不同类别之间的化学成分**关联关系的差异性**。

一、赛题解析

2. 问题分析

【总体分析】：本题是一个关于古代玻璃文物的化学成分数据的数据挖掘类问题。**聚焦于3个问题**：

- 【其一】考察风化对不同类型玻璃文物化学成分含量的影响，并能通过风化后成分**预测**风化前成分；
- 【其二】根据不同文物的化学成分对文物样品进行合理**分类**，给出分类标准，考察同类文物化学成分相似性、异类文物化学成分差异性；
- 【其三】对风化后文物进行分类，确保分类模型足够**稳健**，能包容一部分成分预测带来的误差。

注意

1. 从数据特性来看，成分数据具有**高维、稀疏、严重右偏**等特性，并且各成分累加和为定值100%。因此，本题数据相对特殊，需要对数据做一定的预处理。
2. 从模型选择来看，本体数据**样本量较小**，且分析目的均与实际情况息息相关，因此模型应当追求其可解释性，**不宜用过于复杂的模型**。

02

问题一的分析与求解

求解之前

数据预处理

一、 删除无效数据

删除累加和不在85%-105%范围内的数据，确保数据有效性。

二、缺失值处理分组

所有缺失值统一填充为0.04%：0.04%为仪器测量下限值，避免CLR变换产生无穷值(I
符合实际检测能力限制 。

三、CLR变换（优势：解决数据偏斜问题，保留成分相对关系）

- 中心对数比变换公式：

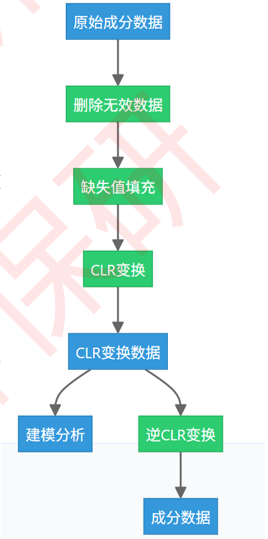
$$clr(x) = [\ln(x_1/g_m(x)), \ln(x_2/g_m(x)), ..., \ln(x\_D/g_m(x))]$$

其中 $g_m(x) = \exp(1/D * \sum \ln(x_k))$.

- **CLR变换理解(相同绝对值的成分在不同样本中相对重要性不同)：**

[30,30,40] → [-0.10,-0.10,0.19], [30,10,60] → [0.13,-0.96,0.82]

CLR是艾奇逊空间到欧氏空间的等距变换。



二、问题一的分析与求解

问题一分析

【问题一】：首先要求以**风化与否为结果变量**，分析数据集中**定类变量之间的关系**。接下来要求分析不同类型玻璃文物风化前后化学成分的变化规律，并给出通过风化后成分数据**预测风化前成分数据的方法**。求解步骤如下：

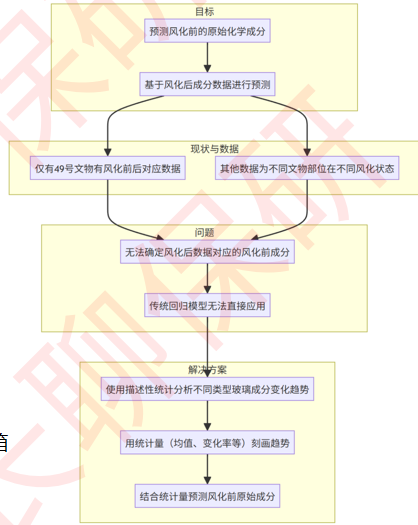
一、 变量关系检验

- 使用**列联表分析**构建风化状态与类型/纹饰/颜色的交叉表
- 应用**假设检验**方法（卡方检验/Fisher精确检验）验证关联性

确认：玻璃类型与风化显著相关（$p<0.05$），**纹饰/颜色无显著影响**

二、成分变化规律分析

- 按玻璃类型（高钾/铅钡）分组。
- 进行**描述性统计**：计算各成分风化前后均值/中位数，绘制成分变化箱线图。 发现规律：**不同类型呈现相反变化趋势（如Na_2O）**。



二、问题一的分析与求解

问题一分析

【问题一】：首先要求以**风化与否为结果变量**，分析数据集中**定类变量之间的关系**。接下来要求分析不同类型玻璃文物风化前后化学成分的变化规律，并给出通过风化后成分数据**预测风化前成分数据的方法**。求解步骤如下：

三、预测模型构建

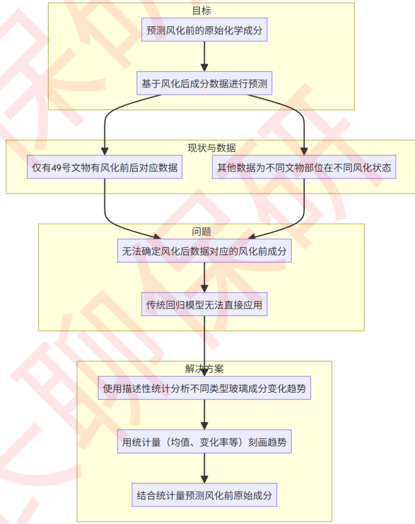
- **均值差计算**：分类型分成分独立计算
- **补偿预测公式**：解决非配对数据问题（仅49号文物有配对数据）

四、模型验证与应用

- 利用49号文物配对数据验证预测准确性
- 输出预测结果支持文物断代与工艺分析

【方法总结】

列联表分析 → 假设检验 → 描述性统计 → 均值差补偿模型 → 本题应用



03

问题二的分析与求解

三、问题二的分析与求解

问题二分析

【问题二】：核心目标在于**根据文物化学成分给出文物分类依据**，并解释分类的现实合理性，并且论证该分类模型的稳健性能确保对风化文物原始成分的预测偏差不会显著影响分类结果。

求解步骤如下：

一、 步骤1：数据准备

- 仅使用未风化点数据：排除风化数据对分类的影响；
- 关键变量选择：Na_2O, CaO, MgO, Al_2O_3, Fe_2O_3等核心化学成分；
- 所有化学成分数据经过CLR变换处理。

二、步骤2：主分类（高钾/铅钡）

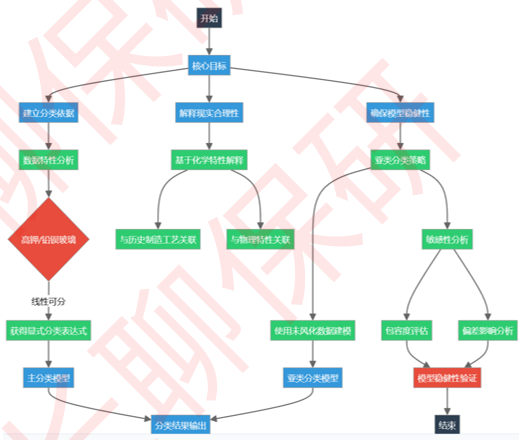
- 使用**支持向量机(SVM)**进行主分类；
- 选择**线性核函数**：因数据线性可分；
- 关键分类特征：PbO和BaO的CLR变换值。

分类边界：

$\xi_{PbO} + 3.51\xi_{BaO} = 5.97$

铅钡玻璃：$\xi_{PbO} + 3.51\xi_{BaO} > 5.97$

高钾玻璃：$\xi_{PbO} + 3.51\xi_{BaO} < 5.97$



三、问题二的分析与求解

问题二分析

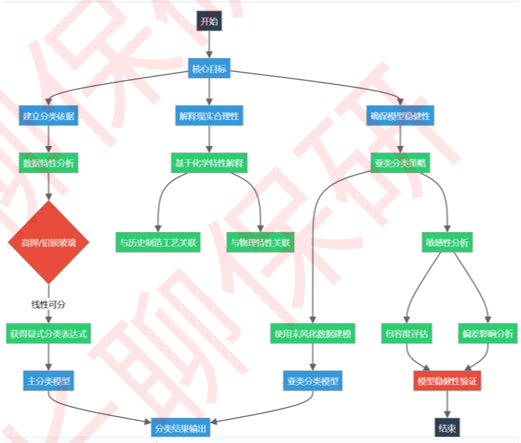
【问题二】：核心目标在于**根据文物化学成分给出文物分类依据**，并解释分类的现实合理性，并且论证该分类模型的稳健性能确保对风化文物原始成分的预测偏差不会显著影响分类结果。

求解步骤如下：

三、 步骤3：亚类划分探索

使用**层次聚类法**进行探索性分析（**距离度量：欧氏距离；聚类方法：Ward最小方差法**）

- 高钾玻璃聚类结果：
  - 亚类1：高钠高钙
  - 亚类2：低钠高钙
  - 亚类3：低钙高镁
- 铅钡玻璃聚类结果：
  - 亚类1：高钠
  - 亚类2：低钠



三、问题二的分析与求解

问题二分析

【问题二】：核心目标在于**根据文物化学成分给出文物分类依据**，并解释分类的现实合理性，并且论证该分类模型的稳健性能确保对风化文物原始成分的预测偏差不会显著影响分类结果。

求解步骤如下：

四、 步骤4：亚类划分建模

- 基于聚类结果，使用**SVM确定精确边界**
- 高钾玻璃亚类划分：
- 铅钡玻璃亚类划分：基于Na_2O含量阈值

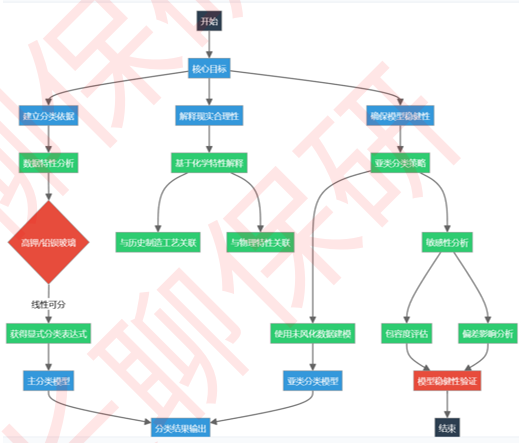
$$\xi_{MgO} + 2.10\xi_{CaO} = 1.51$$

$$\xi_{MgO} + 2.10\xi_{CaO} > 1.51 \rightarrow \text{高钙低镁}$$

$$\xi_{MgO} + 2.10\xi_{CaO} < 1.51 \rightarrow \text{低钙高镁}$$

五、步骤5：敏感性分析

- 验证分类器对风化数据"还原"后的分类效果
- 对分类边界施加扰动测试稳定性
- 结果：模型对预测偏差**具有良好包容度**
- 结论：分类**模型稳健**，可有效应用于风化文物

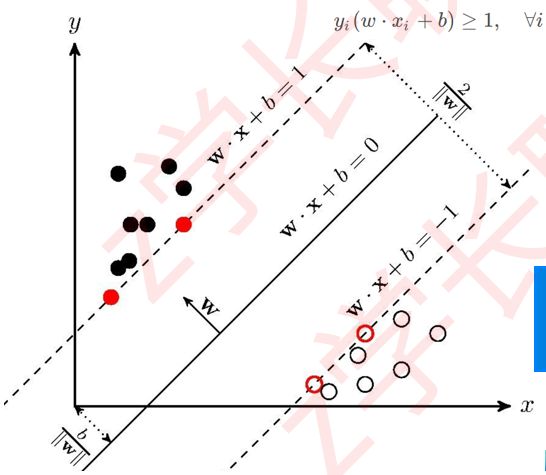


01

支持向量机分类回归与实践

一、支持向量机分类回归与实践

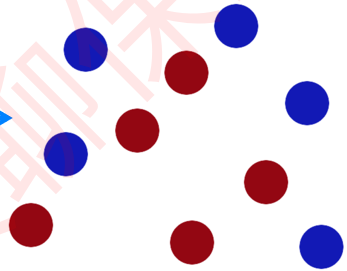
1. 理论介绍-实际案例



支持向量机是一种用于分类和回归的**监督学习**算法，其核心思想是**找到一个最优超平面，以最大化数据分类的间隔。**

为了找到这个最佳的超平面，SVM使用支持向量，即离超平面最近的一些数据点。这些支持向量决定了超平面的位置和方向。

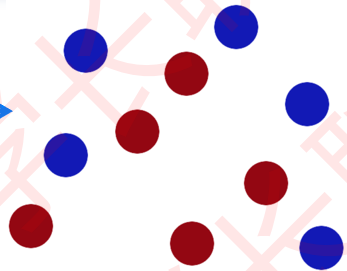
若是线性不可分的?



一、支持向量机分类回归与实践

1. 理论介绍-实际案例

若是线性不可分的?



这时，就引入**松弛变量**，允许一定的分类错误

$$y_i(w \cdot x_i + b) \geq 1, \quad \forall i$$

$$y_i(w \cdot x_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad \forall i$$

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

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

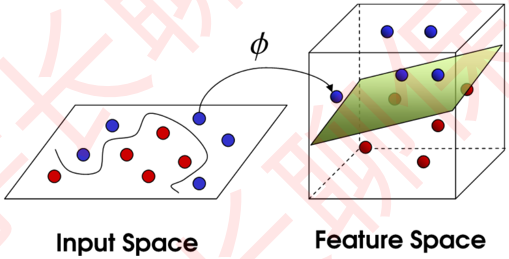
SVM 的优势在于它**可以处理高维数据集，同时具有较强的泛化能力**。它适用于线性和非线性分类问题，可以通过使用不同的核函数来处理非线性关系。

常见的核函数包括线性核、多项式核、径向基函数（RBF）核等。

一、支持向量机分类回归与实践

1. 理论介绍-实际案例

若是复杂且线性不可分的呢?



对于复杂数据，SVM 采用**核方法（引入核函数）**，将低维数据映射到高维空间，使其在高维空间线性可分。

常见核函数包括：

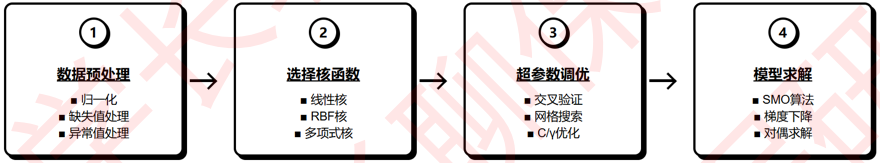
- **线性核**：$K(x_i, x_j) = x_i \cdot x_j$
- **多项式核**：$K(x_i, x_j) = (x_i \cdot x_j + c)^d$
- **高斯核（RBF 核）**：$K(x_i, x_j) = \exp(-\gamma ||x_i - x_j||^2)$

通过核函数，SVM 在高维特征空间找到最优超平面，实现非线性分类。

一、支持向量机分类回归与实践

1. 理论介绍-实际案例

训练SVM的具体步骤（评估指标与其他分类任务一致）：



推荐选择：

- **小规模数据集（<10,000样本）：** 优先使用交叉验证 + 网格搜索 + SMO + 对偶求解，这是最经典且可靠的组合。
- **中大规模数据集（10,000-100,000样本）：** 继续使用SMO作为核心求解器，结合交叉验证 + 随机搜索（代替网格搜索以减少计算量）。
- **超大规模数据集（>100,000样本）：** 考虑梯度下降（尤其是SGD）或数据采样，结合C优化和降维技术。
- **非线性问题:** 必须使用对偶求解 + 核函数，SMO是最佳实现方式。

03

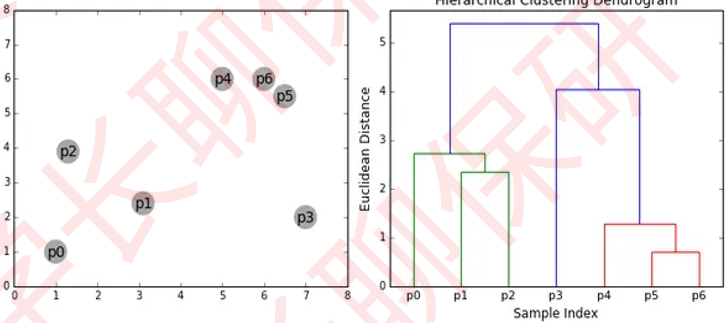
层次聚类算法与实践

三、层次聚类算法与实践

1. 理论介绍-实际案例

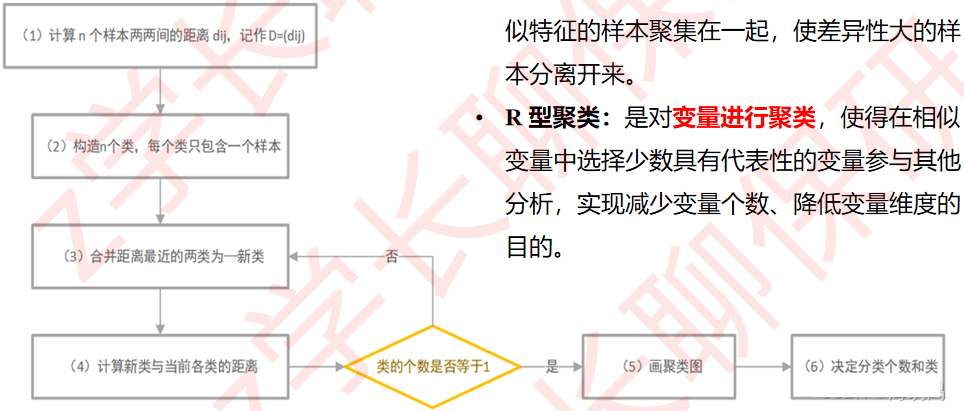
首先**将一定数量的样本或指标各自看成一类**，然后根据样本的亲疏程度，将亲疏程度**最高的两类进行合并**，然后考虑合并后的类与其他类之间的亲疏程度，再进行合并。重复这一过程，直到将所有的样本(或指标)合并为一类。

核心：合并相似项



三、层次聚类算法与实践

1. 理论介绍-实际案例



- **Q 型聚类**：是对**样本进行聚类**，它使具有相似特征的样本聚集在一起，使差异性大的样本分离开来。
- **R 型聚类**：是对**变量进行聚类**，使得在相似变量中选择少数具有代表性的变量参与其他分析，实现减少变量个数、降低变量维度的目的。

三、层次聚类算法与实践

1. 理论介绍-实际案例

距离度量的选择：

距离度量是**层次聚类的关键**，决定了数据点之间的相似性。常用的距离度量包括：

- **欧几里得距离：**适用于连续数值型数据，计算公式为：

$$d(x,y)=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}$$

- **曼哈顿距离：**适用于离散数据，计算公式为：

$$d(x,y)=\sum_{i=1}^{n}|x_i-y_i|$$

- **余弦相似度：**适用于文本数据，计算两个向量之间的夹角，公式为：

$$\text{cosine}(x,y)=\frac{x\cdot y}{\|x\|\|y\|}$$

三、层次聚类算法与实践

1. 理论介绍-实际案例

聚类合并策略：

在自底向上的层次聚类中，合并策略决定了如何计算聚类之间的距离。常见的合并策略包括：

- **单链接**：聚类之间的距离为**两个聚类中最近的两个点之间的距离**。此策略可能导致“链式效应”，使得聚类结果呈现出长条形状。
- **全链接**：聚类之间的距离为两个聚类中**最远的两个点之间的距离**。此策略倾向于形成紧凑的聚类。
- **平均链接**：聚类之间的距离为**所有点对的平均距离**，综合考虑了聚类内部的所有点。
- **Ward法**：通过**最小化聚类内的方差**来选择合并的聚类，通常能够产生更均匀的聚类结果。

04

问题三的分析与求解

四、问题三的分析与求解

问题三分析

【问题三】：核心目标在于**进一步通过未分类的数据检验分类模型的有效性。**新数据中存在风化文物，这类文物完成原始成分预测后，需要考察预测误差在何种范围内时，分类结果保持稳定。具体求解步骤如下：

一、 步骤1：原始成分预测

- 对未知类别文物进行原始成分预测
- 对于风化文物，使用问题一建立的预测模型：

预测公式：$\xi predict = \xi erode + \Delta\xi mean$

所有预测结果转换为CLR值用于分类。

二、步骤2：多重SVM分类

- 首先**使用主分类器区分高钾/铅钡玻璃**：

分类边界：$\xi PbO + 3.51\xi BaO = 5.97$

铅钡玻璃：$\xi PbO + 3.51\xi BaO > 5.97$

高钾玻璃：$\xi PbO + 3.51\xi BaO < 5.97$

四、问题三的分析与求解

问题三分析

【问题三】：核心目标在于**进一步通过未分类的数据检验分类模型的有效性。**新数据中存在风化文物，这类文物完成原始成分预测后，需要考察预测误差在何种范围内时，分类结果保持稳定。具体求解步骤如下：

二、步骤2：多重SVM分类

- **根据大类结果选择亚类分类器**：

高钾亚类边界：$\xi MgO + 2.10\xi CaO = 1.51$

铅钡亚类边界：Na_2O含量阈值

大类	亚类	包含样本
铅钡	高钠	A5
	低钠	A2, A3, A4, A8
高钾	高钙低镁	A1, A6, A7
	低钙高镁	无

三、步骤3：敏感性分析

结论：**模型对MgO和CaO扰动具有较强鲁棒性，Na_2O扰动影响较大**

四、步骤4：特殊样本分析（A5案例）

- A5样本被分类为铅钡-高钠亚类
- 特殊现象：风化后仍保持较高钠含量

四、问题三的分析与求解

问题三分析

【问题三】：核心目标在于**进一步通过未分类的数据检验分类模型的有效性。**新数据中存在风化文物，这类文物完成原始成分预测后，需要考察预测误差在何种范围内时，分类结果保持稳定。具体求解步骤如下：

三、步骤3：敏感性分析

结论：**模型对MgO和CaO扰动具有较强鲁棒性，Na_2O扰动影响较大**

四、步骤4：特殊样本分析（A5案例）

- **A5样本被分类为铅钡-高钠亚类**
- 特殊现象：风化后仍保持较高钠含量
- 解释：铅钡玻璃风化通常会使钠含量下降，A5风化后仍保持高钠，表明**其风化前属于高钠亚类**
- 结论：分类结果具有考古学合理性
- 模型有效性：**除A5外所有样本均被正确分类，A5的分类结果具有可解释性**

05

问题四的分析与求解

五、问题四的分析与求解

问题四分析

【问题四】：核心目标在于**建立合适的模型来提取数据的有效特征**，其一要**控制表面风化的对成分相对含量的影响**，其二要**兼顾部分样本量偏少的亚类**。最终形成的特征要具有可比较性。具体求解步骤如下：

五、问题四的分析与求解

问题四结果

高钾玻璃分析

前4个主成分解释92.11%变异

主成分载荷

- PC1: 氧化钙(CaO)、氧化锶(SrO)
- PC2: 氧化铁(Fe_2O_3)、氧化钡(BaO)
- PC3: 氧化铜(CuO)、二氧化硫(SO_2)

高度相关组合

- 二氧化硅(SiO_2)-氧化铝(Al_2O_3)
- 氧化钠(Na_2O)-氧化钾(K_2O)-二氧化硫(SO_2)
- 氧化铁(Fe_2O_3)-氧化铜(CuO)

化学机理

SiO_2和Al_2O_3形成Si-O-Al键
Na_2O和K_2O作为同族元素性质相似

铅钡玻璃分析

前4个主成分解释90.92%变异

主成分载荷

- PC1: Al_2O_3、SiO_2、SO_2
- PC2: Na_2O、CaO、Fe_2O_3、BaO、P_2O_5

高度相关组合

- BaO-CuO-MgO(负)-Fe_2O_3(负)
- P_2O_5-CaO-Na_2O(负)

化学机理

Ba和Cu提供氧化层保护
MgO和Fe_2O_3的阻塞效应
混合碱效应影响

类别比较

特征	高钾玻璃	铅钡玻璃
变异解释度	92.11%	90.92%
主要关联元素	碱金属(Na,K)	碱土金属(Ba,Ca)
关键负相关	较少	显著(MgO/Fe_2O_3)
硅铝关联	强(SiO_2-Al_2O_3)	强(SiO_2-Al_2O_3)
硫元素关联	与碱金属相关	与硅铝相关
抗风化机制	不明显	明显(氧化层保护)

谢谢

主讲人：颖老师