# 古代玻璃珠成分分析模型

## 摘要

玻璃珠是我国古代玻璃器物中一种常见的类型，具有悠久的历史。本文针对一批古代玻璃珠的数据进行处理和分析，采用灰色关联分析、随机森林和偏最小二乘回归等方法，建立玻璃珠成分分析相关的数学模型，并对题目中提出的未知数据预测等问题作出了有效解答。

问题一中，我们采取了利用Python绘制柱状图的方式，简单直观地反映了玻璃珠的类型与出产年代、纹饰的关系。

问题二中，我们通过计算分析风化前后化学成分含量变化，并给出预测风化前化学成分含量的方式，利用灰色关联分析法分析各种化学成分与风化的相关性，以阐释推测的准确性。

问题三中，我们采用决策树的方法来求解玻璃珠类型的分类规律，但决策树的分类准确性较低，我们采取了多个决策树即随机森林的方式提高预测的准确率。

问题四中，我们先使用随机森林预测出原料产地，之后对不同的产地进行偏最小二乘回归，拟合出PbO、BaO含量与各微量元素含量的关系。

**关键词**：关联性分析、类型鉴别、灰色关联分析、决策树、线性回归

# 目录

[一、问题重述 3](#_Toc137036790)

[二、问题一分析与解答 4](#_Toc137036791)

[2.1 问题一分析 4](#_Toc137036792)

[2.2 问题一求解 4](#_Toc137036793)

[三、问题二分析与解答 5](#_Toc137036794)

[3.1 问题二分析 5](#_Toc137036795)

[3.2 计算分析风化前后化学成分含量变化 5](#_Toc137036796)

[3.3 灰色关联分析法阐释各种化学成分与风化的相关性 6](#_Toc137036797)

[四、问题三分析与解答 8](#_Toc137036798)

[4.1 问题三分析 8](#_Toc137036799)

[4.2 决策树求解玻璃珠类型的分类规律 8](#_Toc137036800)

[4.3 随机森林求解玻璃珠类型的分类规律 9](#_Toc137036801)

[4.4 未知类型样本的预测 10](#_Toc137036802)

[五、问题四分析与解答 11](#_Toc137036803)

[5.1 问腿四分析 11](#_Toc137036804)

[5.2 随机森林预测出产地区 11](#_Toc137036805)

[5.3 偏最小二乘回归拟合PbO、BaO含量与各微量元素含量的关系 12](#_Toc137036806)

[六、模型评价 13](#_Toc137036807)

[6.1 问题一模型评价 13](#_Toc137036808)

[6.2 问题二模型评价 14](#_Toc137036809)

[6.3 问题三模型评价 14](#_Toc137036810)

[6.4 问题四模型评价 14](#_Toc137036811)

[附录 14](#_Toc137036812)

[问题一绘图Python代码 14](#_Toc137036813)

## 一、问题重述

玻璃珠是我国古代玻璃器物中最常见的一类器型，从西周开始就已存在了，至战国时期则大为流行。战国秦汉时期，玻璃珠饰作为一类重要的随葬品，经常出现在墓葬中。由于各地制作玻璃的技术不同，玻璃的化学成分能反映其产地的特征，所以通过对玻璃样本进行化学成分分析，以此来了解它们的产地、来源等历史信息，往往能够获得古代文化交流和贸易等方面的信息。

玻璃的主要原料是石英砂（主要化学成分是SiO2），烧制中加入不同的材料以帮助降低熔点，材料不同其化学成分不同。按照化学成分分类，我国出土的先秦至两汉时期的玻璃大致可分为三种类型：

钠钙玻璃，烧制中加入石灰和草木灰，即以氧化钠NaO作为助熔剂、氧化钙CaO作为稳定剂的硅酸盐玻璃，通常被认为是从西方引入的玻璃品种；

铅钡玻璃，烧制中加入硝石和方铅矿等，以氧化铅PbO、氧化钡BeO作为助熔剂的硅酸盐玻璃，通常被认为是我国自己发明的玻璃品种；例如楚文化的玻璃以铅钡玻璃为主。

钾玻璃，以氧化钾K2O作为助熔剂的硅酸盐玻璃，主要流行于我国岭南以及东南亚和印度等区域。

现有某地出土的一批玻璃珠样本，主要来源于戎人墓和秦人墓，其中秦人墓分为平民墓和贵族墓，样本的基本信息见附件1，主要成分所占比例数据见附件2，（F、G说明）部分样本的主要化学成分及微量元素含量比例数据见附件3。请依据这些数据进行分析，回答以下问题：

**问题1：**依据附件1的数据，分析玻璃珠的类型与出产年代、纹饰的关系。

**问题2：**玻璃珠出土后可能会发生表面风化，依据附件1和附件2，分析表面风化后样品化学成分含量的变化。根据一个表面有风化的样本分析数据，是否能推测出其风化前化学成分含量可能的比例？阐述推测的准确性。

**问题3:** 依据附件1和附件2，找出玻璃珠类型的分类规律。对未知类型的新样本进行分析，判断该样本所属类型，并对分类效果进行评价。

**问题4**：依据附件3数据，分析PbO 、BaO 这两种主要化学成分与哪些微量成分含量相关，对其含量产生什么样的影响？

（备注：对文物样品需要分析其原料所属产地，不同产地的原料其微量元素含量不同，因此需要研究原料（主要化学成分）与哪些微量元素含量密切相关。用另一种分析仪器测得部分样本更多的元素，其成分含量包括主要成分和微量成分，主要成分主要体现原料来源，微量成分为该样本中所有成分的含量数据，这些微量成分是由不同的主要成分带入的。）

## 二、问题一分析与解答

### 2.1 问题一分析

问题一要求依据附件1的数据，分析玻璃珠的类型与出产年代、纹饰的关系。由于本体中的变量均为定类变量，我们将采用Python绘制柱状图，对玻璃珠的类型与出产年代、纹饰的关系进行直观的展示。

### 2.2 问题一求解

从附件1的数据中可以看出，人造颜料玻璃珠没有特定纹饰，出产年代为战国或战国末。对于其他类型的玻璃珠，我们使用Python绘制了两幅柱状图，第一幅图以出产年代为横轴，以纹饰为纵轴，可以看出依据出产年代和纹饰可以将玻璃珠分为五类。第二幅图以这五类玻璃珠为横轴，以玻璃珠的成分类型为纵轴，可以看出战国末单色玻璃珠全部为铅钡玻璃，春秋早期费昂斯纹饰的玻璃珠全部为高钾玻璃，其他三类玻璃珠有可能是高钾、铅钡和高钾铅钡玻璃中的任意一种，通过柱状图可以直观看出各种类型的比例。

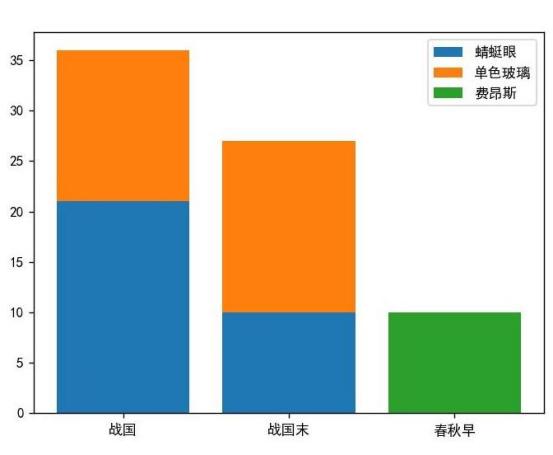


图2-1 纹饰与出产年代的柱状图

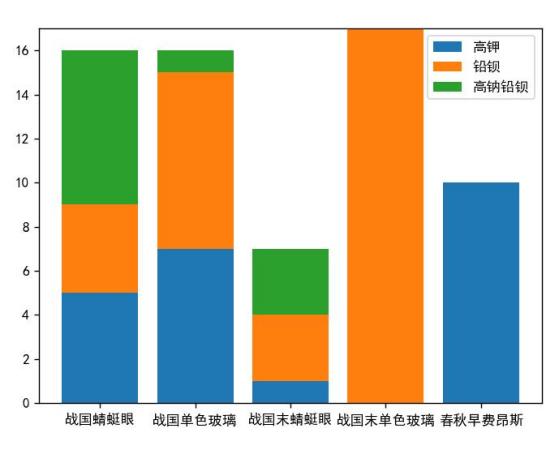


图2-2 类型与纹饰、出产年代的柱状图

## 三、问题二分析与解答

### 3.1 问题二分析

问题二分为两小问：第一问要求依据附件1和附件2，分析表面风化后样品化学成分含量的变化；第二问要求根据一个表面有风化的样本分析数据，是否能推测出其风化前化学成分含量可能的比例并阐述推测的准确性。我们考虑先采用求平均值、列表的方式直观展示表面风化前后样品化学成分含量的变化，再采用灰色关联分析法阐释各种化学成分与风化的相关性，即比例推测的准确性。

### 3.2 计算分析风化前后化学成分含量变化

四种类型的玻璃珠在表面风化时产生的变化是不同的，对于四种类型的玻璃珠，我们分别对每类中风化和未风化的样本，计算各种化学成分的平均值作为标准值，可以看出表面风化前后各种化学成分的变化（注：数据均为该成分所占比例数值）。



图3-1 表面风化前后各种化学成分的变化

给出某种类型表面风化的样本，已知某种化学成分含量为，从上表中查出该成分已风化标准值为，未风化标准值为，则预测样本风化前该成分含量。

### 3.3 灰色关联分析法阐释各种化学成分与风化的相关性

不一定每种化学成分的变化都与风化密切相关，因此这种推测对不同化学成分的准确性不同，我们采用灰色关联分析来阐释各种化学成分与风化的相关性。

第一步需要确定分析数列。首先确定反映系统行为特征的参考数列和影响系统行为的比较数列。通常称第一种序列为参考数列。后一种数列称比较数列。

参考序列(母序列)公式为：



比较序列(子序列)公式为：



第二步，对变量进行无量纲化处理。由于系统中的各种因素的数据量纲可能不同，会造成比较结果较差甚至无法正确得到结果的情况。因此在进行灰色关联分析的适合，首先要进行数据的无量纲化处理。本文采用的是均值化处理的方法。

均值化处理公式：



其中k对应时间段, i对应比较数列中的第i行。

第三步，计算关联系数。公式如下：



其中ρ∈(0, ∞), ρ称为分辨系数。而ρ的大小决定分辨能力，分辨能力随ρ的减小而增大，反之亦然。一般ρ的取值是(0，1)。当ρ≤0.5463时，分辨能力最好。本文取0.1。

第四步，计算关联度。关联系数时母序列和子序列在各个时刻的关联程度值，所以这个值不只有一个。但是由于值过多会导致信息分散，不便于进行整体的比较，因此对各个时刻的关联系数集中为一个平均值是十分必要的。作为母序列和子序列的关联程度的该值的公式如下：



第五步，关联度排序。在构建模型的最后需要对关联度按大小排序，当在算出与两个母子序列的关联系数后，计算各类关联系数的平均值，即为关联度。

采用SPSSPRO软件，以风化情况为母序列，各化学成分含量为特征序列，得到各化学成分含量与是否风化的灰色关联度排名如下，关联度越高，预测结果越有说服力。

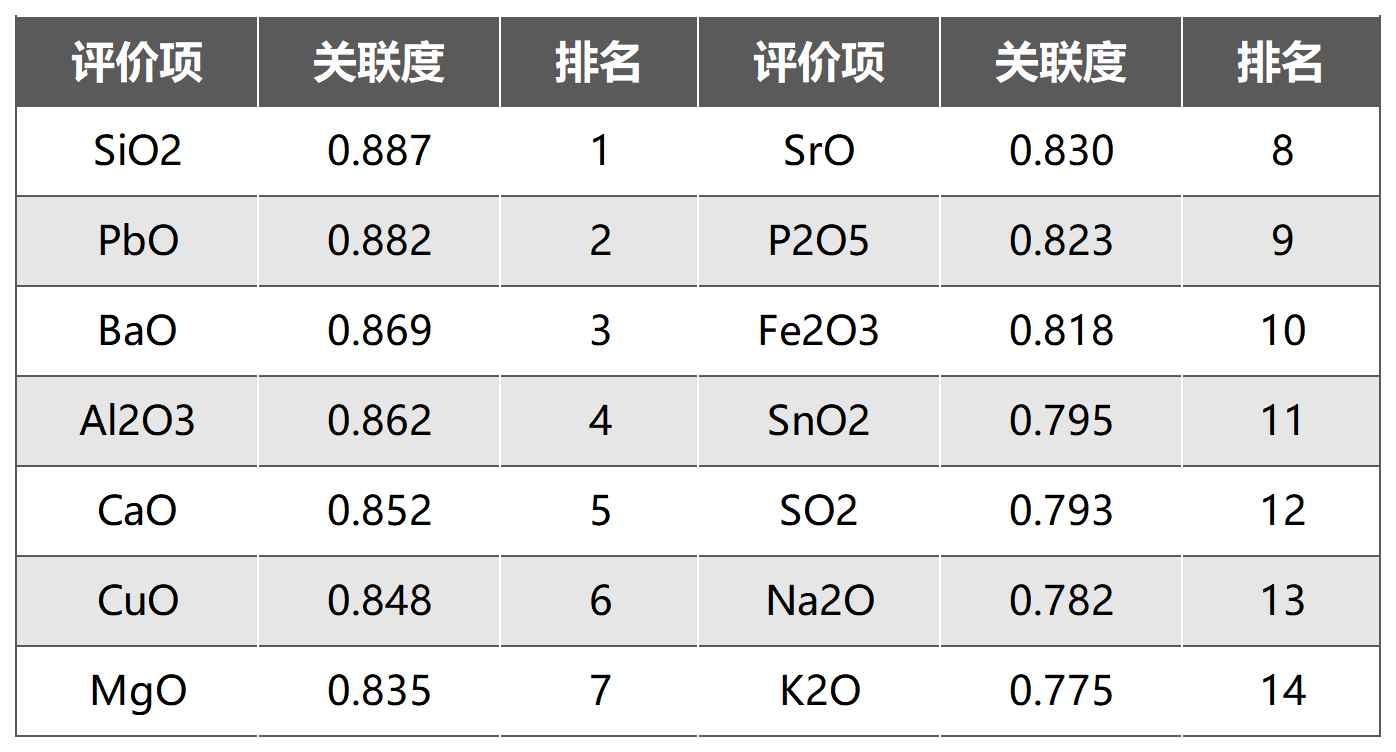


图3-2 各化学成分含量与是否风化的灰色关联度排名

## 四、问题三分析与解答

### 4.1 问题三分析

问题三分为两小问，第一问要求依据附件1和附件2，找出玻璃珠类型的分类规律；第二问要求对未知类型的新样本进行分析，判断该样本所属类型，并对分类效果进行评价。我们考虑采用决策树找出分类规律，并将未知样本采用决策树进行分析。决策树是一种机器学习方法，可用来做分类或回归，其基本思想是根据一些属性和阈值，将数据集划分成不同的子集，直到每个子集都属于同一类别或满足某个停止条件。决策树的每个内部节点表示一个属性上的判断，每个分支代表一个判断结果，每个叶节点代表一个分类结果或一个回归值。

决策树的生成算法有很多，比如ID3，C4.5，CART等。这些算法的主要区别在于如何选择划分属性和阈值，以及如何避免过拟合。一般来说，划分属性和阈值的选择是基于信息增益，信息增益率或基尼指数等准则，目的是使得划分后的子集尽可能纯净，即属于同一类别或具有相似的回归值。决策树的优点有易于理解和解释，需要很少的数据预处理、可以处理多种类型的数据等，但也有一些缺点，如容易过拟合、对噪声数据敏感、可能不稳定等。

### 4.2 决策树求解玻璃珠类型的分类规律

采用SPSSPRO软件，选择类型为分类变量，各化学成分含量及表面是否风化为自变量，得到的决策树结构如图4-1。

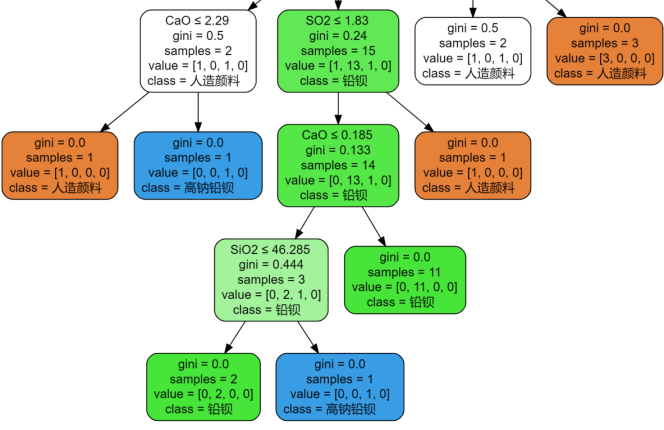
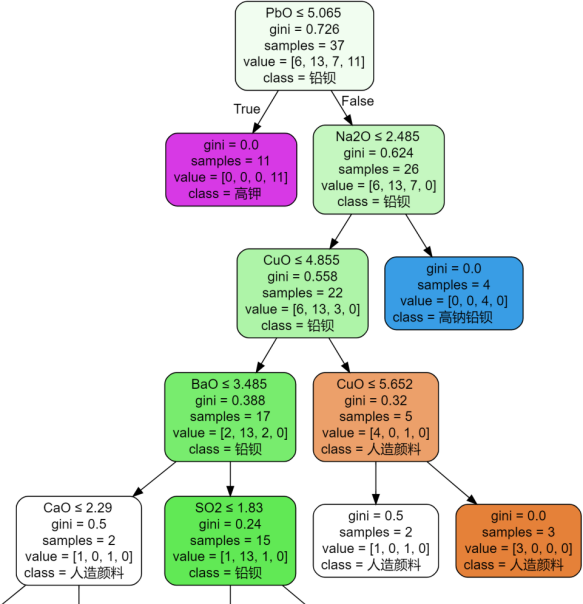


图4-1 玻璃珠类型分类决策树



图4-2 决策树准确率（0.7的训练数据）

在该决策树中，我们选取70%的数据作为训练数据，30%的数据作为测试数据，由图4-2可见，测试数据的预测结果的正确率仅有81%，有接近五分之一的误判可能性。

### 4.3 随机森林求解玻璃珠类型的分类规律

对此，我们考虑采取随机森林模型进行优化。随机森林模型就是多个决策树组成的森林，即由于对于同一组数据可能存在很多种决策树分类方式，我们可以生成多个不同的决策树组成一个随机森林，将要预测的数据分别输入这些决策树。由于随机森林中决策树的数量众多，不同决策树的不准确性得到弥补，我们就能得到相对准确的预测结果。

我们仍选取70%的数据作为训练数据，30%的数据作为测试数据，生成100个决策树组成的随机森林，得到的测试集准确率如图4-3。



图4-3 随机森林准确率（0.7的训练数据）

可见其预测准确率已可达93%。而当我们选取80%的数据作为训练数据，其预测准确率约可达100%，如图4-4所示。



图4-4 随机森林准确率（0.8的训练数据）

因此随机森林模型对于此问题的求解是较为准确的。我们将全部的数据作为训练数据生成一个随机森林，这样能够进一步提高模型的准确率。

### 4.4 未知类型样本的预测

依靠上面得到的随机森林，我们可以对第二个问题进行求解。将未知数据放入模型中进行预测，得到根据各化学成分含量及表面是否风化对未知类型样本的预测结果如下（注：1代表未风化，2代表已风化）。

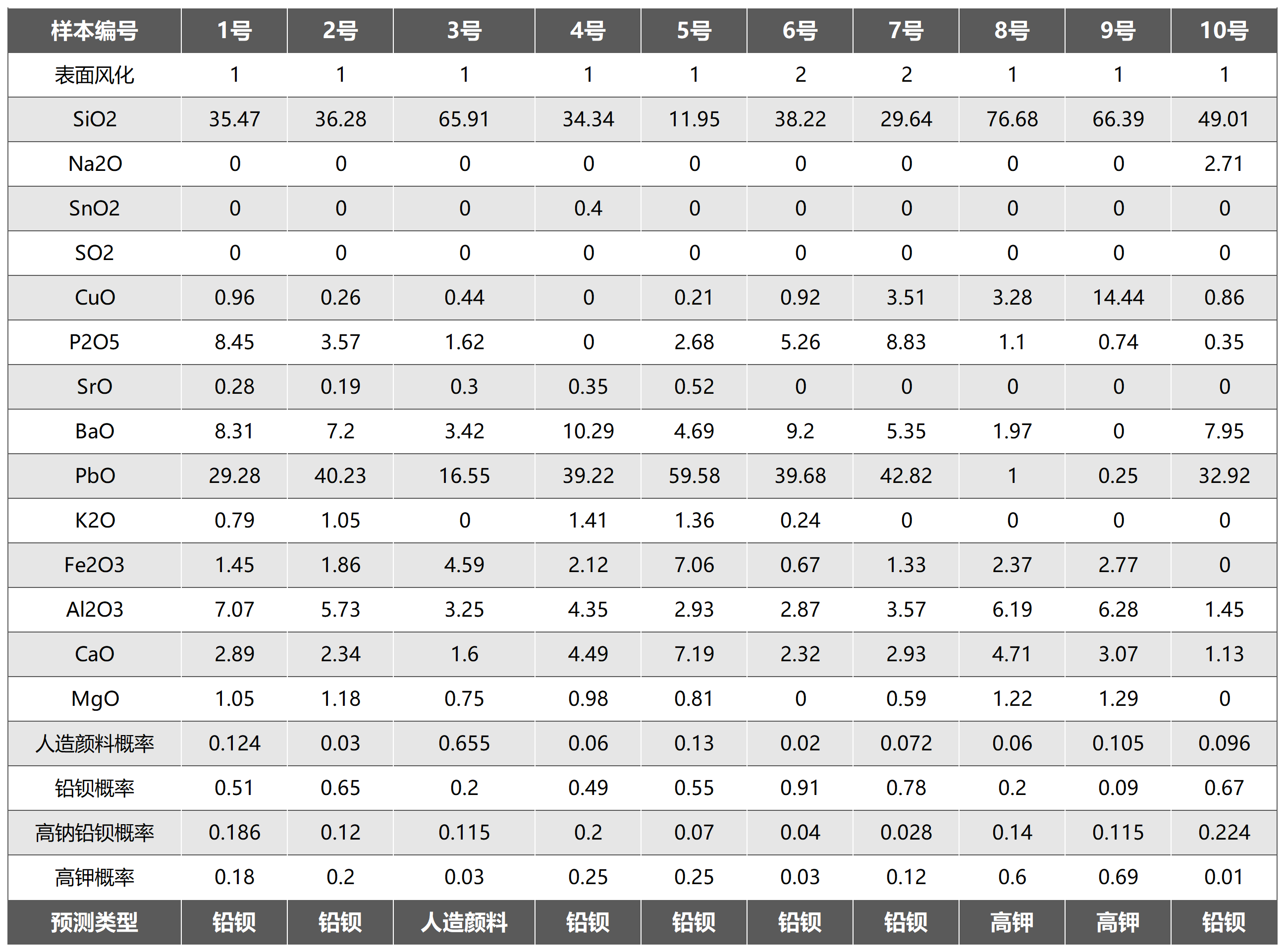


图4-5 未知类型样本的预测

## 五、问题四分析与解答

### 5.1 问腿四分析

问题四要求依据附件3数据，分析PbO 、BaO 这两种主要化学成分与哪些微量成分含量相关，对其含量产生什么样的影响。根据问题中的假设，不同出产地区主要化学成分与微量元素的关系不同，因此在分析附件3给出的数据前，我们必须找到其中数据的出产地区。而题目假设主要成分主要体现原料来源，因此我们应首先找到主要化学成分与出产地区的关系，并根据附件3中的主要化学成分推定其出产地区。仿照问题三中的方法，我们可以采取随机森林的方法来找到其关系，并对其出产地区做出预测。在预测出它们的出产地区后，我们可以采用偏最小二乘回归（PLSR）分别拟合不同产地的玻璃珠PbO、BaO含量与各微量元素含量的关系。

### 5.2 随机森林预测出产地区

运用SPSSPRO软件，以出产地区为分类变量，各主要化学成分含量为自变量，得到的随机森林中决策树如图5-1，并根据得到的随机森林推定样本的出产地区。

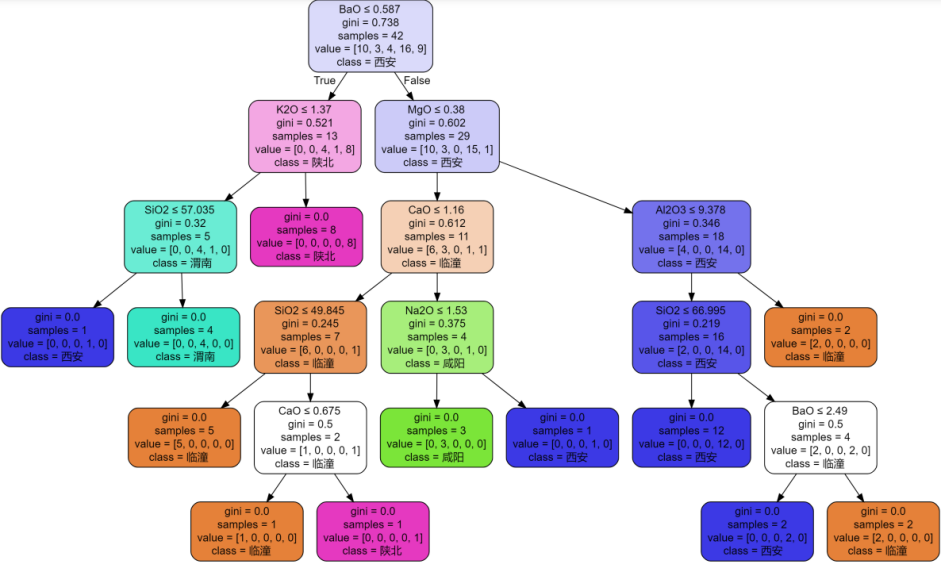


图5-1 出产地区与主要化学成分的随机森林中的决策树



图5-2 随机森林的预测准确性

### 5.3 偏最小二乘回归拟合PbO、BaO含量与各微量元素含量的关系

得到样本的出产地区后，我们使用偏最小二乘回归（PLSR）拟合PbO、BaO含量与各微量元素含量的关系。偏最小二乘回归可以用来分析多个自变量和多个因变量之间的关系，集主成分分析、典型相关分析和多元线性回归分析3种分析方法的优点于一身。它与主成分分析法都试图提取出反映数据变异的最大信息，但主成分分析法只考虑一个自变量矩阵，而偏最小二乘法还有一个“响应”矩阵，因此具有预测功能。

偏最小二乘回归的基本思想是，通过投影的方式，将自变量矩阵X和因变量矩阵Y分别映射到一个新的空间，得到一组新的变量，称为潜在变量，或者叫做因子。其优点是可以处理自变量数量多于观测数量的情况，以及自变量之间存在多重共线性的情况。

采用SPSSPRO软件，以PbO 、BaO含量为因变量，各微量元素含量和出土地区为自变量，得到的拟合结果如下（注：出土地区按临潼、陕北、西安、咸阳、渭南依次量化为1、2、3、4、5）。

主要成分含量=常数+系数对应因变量值。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **PbO** | **BaO** |  | **PbO** | **BaO** |
| 常数 | 74.458 | 3.116 | Eu | 3.571 | 3.19 |
| 出土地区 | -4.086 | 0.444 | Sm | -3.27 | -8.84 |
| Ti | -0.004 | -0.002 | Nd | 2.78 | -3.633 |
| Li | 1.256 | 0.533 | As | 0.02 | 0.004 |
| Se | 1.749 | 1.276 | La | -3.171 | 0.19 |
| Y | -4.399 | 5.122 | Ce | 1.594 | 1.403 |
| Sr | -0.012 | 0.001 | Cs | -1.982 | -1.609 |
| Rb | -2.172 | -1.25 | Sn | -0.018 | -0.006 |
| Cu | 0 | 0 | Sb | -0.013 | 0.001 |
| Dy | 5.004 | 2.454 | Ge | 2.17 | 4.402 |
| Bi | 0 | -0.001 | In | -0.725 | -0.475 |
| U | -14.381 | -0.467 | Cd | 1.487 | -0.236 |
| Ho | -1.094 | -1.173 | Ag | 0.01 | -0.002 |
| Th | -11.762 | -7.821 | Mo | 7.876 | 5.627 |
| Au | 1.237 | -1.106 | Nb | -14.573 | 3.192 |
| Tl | 0.478 | -0.418 | Ga | 1.405 | 2.935 |
| Hf | -2.66 | -1.112 | Be | -9.879 | -3.446 |
| Ta | 0.358 | 0.023 | Zr | 1.887 | -0.322 |
| W | -0.697 | 0.785 | Zn | -0.011 | -0.007 |
| Yb | -2.989 | -4.126 | B | 0.291 | -0.186 |
| Lu | -0.176 | 0.331 | Ni | 0.733 | -0.213 |
| Er | 1.151 | -1.273 | Co | -0.017 | -0.026 |
| Tm | -0.373 | -0.304 | V | 4.792 | 3.137 |
| Pr | -2.253 | 0.598 | Cr | -0.203 | -0.361 |
| Tb | -0.423 | -0.92 | Sc | -6.141 | -2.717 |
| Gd | -6.426 | -7.993 |  |  |  |

图5-3 PbO ,BaO含量与微量元素含量的线性关系

## 六、模型评价

### 6.1 问题一模型评价

问题一所采取的画柱状图的方式可以直观地展示玻璃珠的类型与出产年代、纹饰的关系，但缺乏恰当的数学模型对其进行定量分析，没有得出相应的数学规律，可以考虑采取更加准确严谨的理论模型对其进行分析。

### 6.2 问题二模型评价

问题二采用平均值作为标准值准确性有待考证，最好采用更多元的统计值得出标准值；灰色关联分析得到的各成分关联度相差不大，应采用斯皮尔曼系数进行验证。

### 6.3 问题三模型评价

问题三采取的随机森林模型在80%的训练数据，20%的测试数据时能够达到接近100%的准确率，说明该模型已经能够较准确地预测玻璃珠的类型。

### 6.4 问题四模型评价

问题四中得到的随机森林准确度不够高，测试集的数据准确率在60%至90%上下浮动，这可能意味着随机森林出现了过拟合或者随机森林模型并不适用于本问题的求解。 偏最小二乘法是一个线性回归模型，但我们并不能保证PbO、BaO含量与各微量元素含量的关系一定是线性的，可能存在求解其关系的更优模型。

## 附录

### 问题一绘图Python代码

import pandas as pd

import matplotlib.pyplot as plt

import numpy as np

plt.rcParams["font.sans-serif"]=["SimHei"]

df = pd.read\_excel('data.xls')

wens=np.zeros((3,3),dtype=int)

labels=["战国","战国末","春秋早"]

for i in range(0,len(df["实验编号"])):

if df["纹饰"][i]=="蜻蜓眼":

j=0

elif df["纹饰"][i]=="单色玻璃":

j=1

elif df["纹饰"][i]=="费昂斯":

j=2

if df["出产年代"][i]=="战国":

wens[j][0]+=1

elif df["出产年代"][i]=="战国末":

wens[j][1]+=1

elif df["出产年代"][i]=="春秋早":

wens[j][2]+=1

plt.bar(labels,wens[0],label="蜻蜓眼")

plt.bar(labels,wens[1],bottom=wens[0],label="单色玻璃")

plt.bar(labels,wens[2],bottom=wens[1],label="费昂斯")

plt.legend()

plt.show()

types=np.zeros((3,5),dtype=int)

labelws=["战国蜻蜓眼","战国单色玻璃","战国末蜻蜓眼","战国末单色玻璃","春秋早费昂斯"]

for i in range(0,len(df["实验编号"])):

if df["纹饰"][i]=="蜻蜓眼" and df["出产年代"][i]=="战国":

j=0

elif df["纹饰"][i]=="单色玻璃" and df["出产年代"][i]=="战国":

j=1

elif df["纹饰"][i]=="蜻蜓眼" and df["出产年代"][i]=="战国末":

j=2

elif df["纹饰"][i]=="单色玻璃" and df["出产年代"][i]=="战国末":

j=3

elif df["纹饰"][i]=="费昂斯" and df["出产年代"][i]=="春秋早":

j=4

if df["类型"][i]=="高钾":

types[0][j]+=1

elif df["类型"][i]=="铅钡":

types[1][j]+=1

elif df["类型"][i]=="高钠铅钡":

types[2][j]+=1

plt.bar(labelws,types[0],label="高钾")

plt.bar(labelws,types[1],bottom=types[0],label="铅钡")

plt.bar(labelws,types[2],bottom=(types[0]+types[1]),label="高钠铅钡")

plt.legend()

plt.show()