

基于最小二乘回归模型的红外光谱溶液浓度预测

摘要

最小二乘回归模型通过最小化误差的平方和寻找数据的最佳函数匹配，可以简便地求得未知的数据，并使得这些求得的数据与实际数据之间误差的平方和为最小，因此本文选用最小二乘回归模型作为矫正模型。我们将 $\frac{2}{3}$ 的数据用于训练集合，剩下 $\frac{1}{3}$ 的数据用于检测集合。我们发现吸光度-波长图像下方的数据预测效果良好，上方的数据预测效果。

最后，我们对本模型进行评价，提出了可以用偏最小二乘、机器学习算法进行预测的展望。

关键词：红外光谱；Savitzky-Golay平滑处理；混合溶液浓度检测；预测模型；普通最小二乘回归模型

1 问题重述

目前有两种溶液U(VI)和U(VI)按照一定的浓度与硝酸溶液混合。经过红外光线扫描，得到该混合溶液在一定波长下的频谱。试根据现有频谱数据建立预测两种溶液的浓度的数学模型，并用该模型预测出待检混合溶液样本中的两种溶液的浓度。

2 问题分析

2.1 光谱数据预处理

2.1.1 数据的平滑处理

2.1.2 聚类分析

混合溶液的光谱图呈现明显的上下分离趋势，因此我们使用K-means聚类算法将频谱数据分成2类，分别建立模型，寻找参数，提高预测效果。

2.2 预测模型建立与检验

普通最小二乘法通过最小化误差的平方和寻找数据的最佳函数匹配，可以简便地求得未知的数据，并使得这些求得的数据与实际数据之间误差的平方和为最小，较为常用。在本题中，我们选择普通最小二乘法模型作为预测模型。我们仅利用 $\frac{2}{3}$ 的数据集来建立模型，剩下 $\frac{1}{3}$ 的数据集用来对模型进行误差检验，识别模型预测效果。

2.3 预测待检混合溶液浓度

模型建立与误差检验完成后，使用该模型预测出待检混合溶液样本中的两种溶液的浓度。研究思路如1所示。

3 模型假设

1. 扫描溶液的红外光线是平行单色光。
2. 本文涉及的溶液均为均匀的、非散射的吸光物质溶液。
3. 每次测量的光程不变。

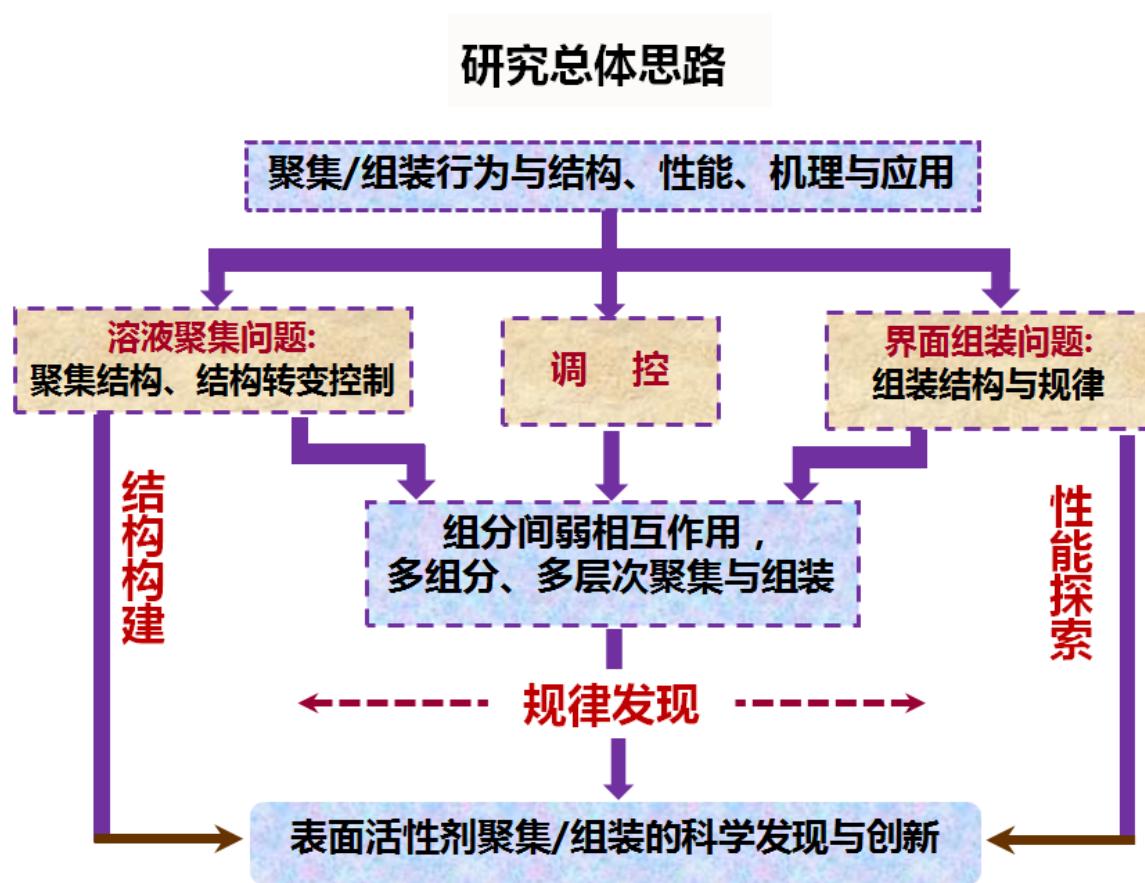


图 1: 研究思路图

4 符号说明

表 1: 这是一个表格

符号	说明	单位
\int	积分符号	
W_0	区分高峰和低峰的一个临界值	
M_t	简单移动平均项	

5 模型的建立与求解

5.1 红外光谱检测方法

含有各类官能团和化学键的化学物质对各类光的吸收、散射、发射的特征不同，因此可以用来了解其组成特点和数量[2]。由于技术和应用的不同，通常把红外光谱区划分为3个区：近红外区，中红外区和远红外区。近红外光谱区位于 $780 \sim 2500\text{nm}$ ，是人们认识最早的非可见光光谱，其中位于 $780 \sim 1100\text{nm}$ 的光谱区域称为短波近红外区，位于 $1100 \sim 2526\text{nm}$ 的光谱区域称为长波近红外区，如2所示。近红外光谱技术具有测量过程简单、分析速度快、成本低、效率高、样品不需要预处理、可同时测定样品的多个组分、测量重现性好等特点。

分子的非谐振性振动使得分子振动从基态向高能级跃迁，近红外光谱由此产生，主要记录含氢基团 (C-H、O-H、N-H、S-H、P-H等) 振动的倍频和合频吸收。近红外光通过待测样品时，不同样品具有不同的基团组成及含量，对应的近红外光的吸收程度也不同。光吸收的程度通常用吸光度 A (Absorbance) 来表示，指光线通过某一溶液或物质前的入射光强度与该光线通过溶液或物质后的透射光强度比值的对数。朗伯-比尔定律 (Beer-Lambert Law) 可以定量描述样本吸光度特性，是近红外光谱分析技术的基础，其表达式为：

$$A = \log \left(\frac{I_0}{I_T} \right) = \log \left(\frac{1}{T} \right) = Kbc, \quad (1)$$

其中：

A: 吸光度；

I_0 : 入射光强度；

T: 为透射比 (透过光强度比入射光强度)；

c: 被测样品组分浓度；

b: 液层厚度；

K: 比例常数，与入射光的波长，待测物质的性质和溶液的温度等因素有关。



图 2: 近红外光谱区在电磁波中的位置

当一束平行单色光垂直通过某一均匀、非散射的吸光物质溶液时，其吸光度 A 与溶液液层厚度 b 和样品浓度 c 的乘积成正比。

5.2 数据预处理

5.2.1 数据除噪

本题提供了两种溶液用红外光线扫描后在一定波长下的频谱数据和待测溶液样本的光谱数据。对光谱的三次测量进行平均后得到的结果如图 3 所示。通过观察图像，我们发现数据在峰值位置存在明显的振荡，故需要对数据进行平滑处理。

平滑处理可以使数据变得更加平滑，便于进一步的分析和理解，有效减少数据中的噪声和异常值，提高数据的可靠性和准确性。我们分别采用Movmedian平滑处理、五点三次平滑处理、Savitzky-Golay (SG)平滑处理，然后对此三种处理的结果进行比较，选择处理效果最好的方法。

{Insert Figure 3 here. }

5.2.2 分类分析

由于聚类分析后的结果与图像不符，我们以光谱数据900nm处的浓度作为分类标准，浓度大于1的为一类，浓度小于1的为一类，分类后的效果如图9所示。

6 模型分析

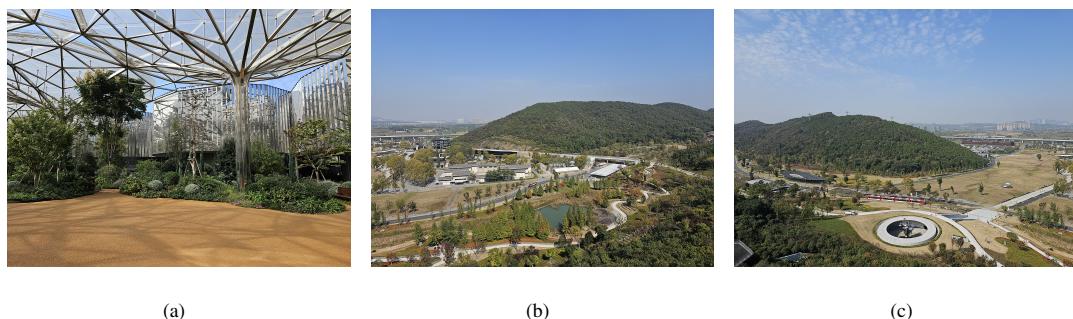


图 3: 并列放置的3张江苏园博园图@20231029

7 模型总结

7.1 模型的优点

7.2 模型的缺点

7.3 模型的推广与改进

这篇文章结束了!

参考文献

- [1] 熊婵. 基于多维多模式超光谱系统的复杂混合溶液成分分析[D]. 天津大学, 2012.
- [2] 谈爱玲. 水中石油类污染物光纤光谱检测方法的研究[D]. 燕山大学, 2012.
- [3] 孙明顺. 毕赤酵母发酵过程中甲醇浓度的红外光谱分析方法研究[D]. 山东大学, 2016.

A 附录

A.1 问题一

```
1 q=2;w=2;e=2;
2 X(:,1)=X(:,1).^q;
3 X(:,2)=X(:,2).^w;
4 X(:,3)=X(:,3).^e;
5 Y=nongdu_up(1:56,:);
6 [B]=OLS(Y,X)
7 x=unknownabs';
8 for j=1:7
9 X1(:,j)=x(:,upi(j,2));
10 end
11 Y0=X1*B;
```

A.2 问题二

A.3 问题三

A.4 问题四

A.5 参数可靠性检验

1. 插入表格

表1 描述性统计

a	c1	c2	c3	c4	c5
a	c1	c2	c3	c4	c5
a	c1	c2	c3	c4	c5
a	c1	c2	c3	c4	c5
a	c1	c2	c3	c4	c5
a	c1	c2	c3	c4	c5
a	c1	c2	c3	c4	c5
a	c1	c2	c3	c4	c5

2. 公式对齐

$$\begin{cases} y = 1, & -1 < x < 4, \\ y = 2, & x > 6, \\ y = 3, & -5 < x, \end{cases}$$