**Modeling and Simulation** 建模与仿真, 2022, 11(2), 405-410
Published Online March 2022 in Hans. http://www.hanspub.org/journal/mos
https://doi.org/10.12677/mos.2022.112037

Hans汉斯

# 基于中红外光谱数据对中药材的鉴别分析

**刘松亭[1]，许雯婷[2]**

[1]安徽职业技术学院基础教学部，安徽 合肥
[2]安徽职业技术学院环境与化工学院，安徽 合肥

收稿日期：2022年2月20日；录用日期：2022年3月11日；发布日期：2022年3月18日

## 摘 要

**文章首先选取中红外光谱数据差异较为明显的波数区域，并使用K-means聚类分析法对其进行分析，从而鉴别出药材种类；再次选取同种药材在不同产地的中红外光谱波峰处数值，对这些数据针对产地和需鉴定药材种类进行均值处理，并对相应波段的差值绝对值的方差进行估计，方差最小的则确定为目标药材的产地；最后，采用暴力枚举法对数据进行分析，并利用方差估计进一步进行判断，得出对应种类和产地。**

## 关键词

**K-Means聚类分析法，方差，暴力枚举法**

# Identification of Chinese Medicinal Materials by Using Mid-Infrared Spectroscopy Data

**Songting Liu[1], Wenting Xu[2]**

[1]Department of Basic Education, Anhui Vocational and Technical College, Hefei Anhui
[2]School of Environment and Chemical Engineering, Anhui Vocational and Technical College, Hefei Anhui

Received: Feb. 20th, 2022; accepted: Mar. 11th, 2022; published: Mar. 18th, 2022

## Abstract

**In this paper, the wavenumber regions with obvious differences in mid-infrared spectral data were firstly selected and analyzed by k-means clustering analysis, so as to identify the medicinal materials. The values at the mid-infrared spectrum peaks of the same medicinal materials in different producing areas were selected again, and these data were averaged according to the origin**

文章引用: 刘松亭, 许雯婷. 基于中红外光谱数据对中药材的鉴别分析[J]. 建模与仿真, 2022, 11(2): 405-410.
DOI: 10.12677/mos.2022.112037

**and the species of medicinal materials to be identified. The variance of the absolute difference values of the corresponding bands was estimated, and the one with the smallest variance was determined as the origin of the target medicinal materials. Finally, the violence enumeration method was used to analyze the data, and variance estimation was used to further judge the corresponding species and origin.**

## Keywords

**K-Means Clustering Analysis, Variance, Violence Enumeration Method**



## 1. 问题背景

本文以 2021 年高教社杯全国大学生数学建模竞赛 E 题为依托进行分析。中药材的种类及产地对中药材的药性有很大影响，因此，对中药材的种类及产地的判定十分重要。不同中药材表现的光谱特征差异较大，即使来自不同产地的同一药材，因其无机元素的化学成分、有机物等存在的差异性，在近红外、中红外光谱的照射下也会表现出不同的光谱特征，因此可以利用这些特征来鉴别中药材的种类及产地。本文主要目标是通过中红外光谱数据进行数据分析，从而对药材的种类和产地进行鉴别分析[1]。对所给药材的中红外光谱数据进行数据处理，筛出异常数据，选取波峰差异明显区域，利用 K-means 聚类分析法，得到药材分类结果；接下来，根据已知药材种类数据，并依据相应波段的差值绝对值方差进行估计分析，得出目标药材的种类判定结果[2] [3] [4]；最后，依据给定种类和产地的中药材的近红外光谱数据，利用暴力枚举法判定未知药材的种类及产地，并根据方差进行进一步精细判定。

## 2. 药材种类的鉴别模型

### 2.1. 异常数据处理

题中附件 1 给出了所需判断种类的药材中红外数据，由于所给数据量较大，因此对数据在 EXCEL 进行排序进行处理后发现 NO64、NO136、NO201 这三个编号数据与整体数据差异较大，因此本文首先将该异常数据清洗掉。

### 2.2. 模型建立与分析

选择波数区域(652-662)、(1144-1154)、(3242-3252)数据差异较为明显三个波段的数据进行均值处理，建立欧式距离分析处理较为合适使用 K-means 聚类分析法对三个区域的数据进行分析。

建立欧式距离的表达式如下：

$$d_{ij}=\sqrt{\left(x_i-x_j\right)^2+\left(y_i-y_j\right)^2} \tag{1}$$

$$\left(i,j=652,\cdots,662;1144,\cdots,1154;3242,\cdots,3252\right)$$

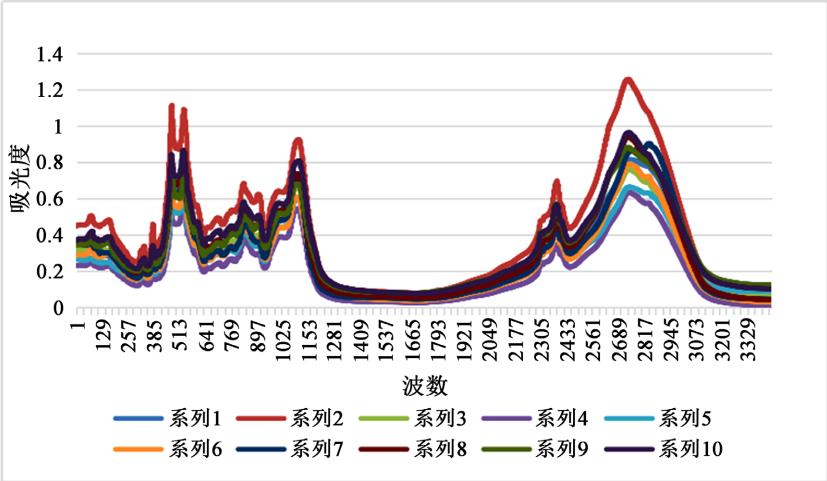
其中：$d_{ij}$ 表示第 $i$ 个样本和第 $j$ 个样本之间的距离。

按照吸光性由低到高分为三类，利用 K-means 聚类分析法，对上述三个区域的数据进行迭代处理，得到 10 组结果。因此，分析得出所给药材数据中有 10 种药材。

## 3. 同种药材的不同产地的鉴别模型

### 3.1. 数据分析

根据题中提供的数据显示，同种药材在不同产地的中红外光谱数据曲线图变化趋势相同，但变化量有所差异。根据题目中给出的数据按照产地构建中红外光谱数据曲线图，如下如图 1。




注：该图为产地 1 的药材中红外光谱数据图，其他产地同理，故不列入正文。

**Figure 1.** Mid-infrared spectrum data of crude drugs from place 1
**图 1.** 产地 1 的药材中红外光谱数据图

其次，对中红外光谱数据曲线图选取波峰处数据作为对比参照，观察图中波峰区间进行分组，选取 6 组波峰，分别为(551-650)、(1011-1110)、(1371-1470)、(1601-1700)、(2901-3000)、(3251-3350)。因此，本考虑对已知药材的中红外光谱图的数据与其不同产地药材的中红外光谱数据的方差数据进行分析，从而鉴别同种药材的不同产地。

### 3.2. 模型建立与分析

考虑对已知药材的中红外光谱图的数据与其不同产地药材的中红外光谱数据的方差数据进行分析，从而鉴别同种药材的不同产地。

分别对该区域内的数据按照产地和种类分别进行均值处理，首先按照产地分类进行均值处理(表 1)：

$$\overline{OP_{n\partial k}} = \frac{\sum_{m=1}^{100} OP_{n\partial km}}{100} \tag{2}$$

$$\left(\partial = 1,2,\cdots,6; k = 550,1000,1370,1600,2900,3250; n = 1,2,\cdots,11\right)$$

$\partial$：波峰区间分组；

$\overline{OP_{n\partial k}}$：针对第 $n$ 个产地的第 $\partial$ 组的均值处理结果；

$OP_{n\partial km}$：针对第 $n$ 个产地的第 $\partial$ 组的第 $k+m$ 个数据。

其次，按照产地分类进行均值处理：

$$\overline{NO_{\partial\omega k}} = \frac{\sum_{m=1}^{100} NO_{\partial\omega km}}{100} \tag{3}$$

$$(\omega = 3,14,38,48,58,71,79,86,89,110,134,152,227,331,618)$$

$\overline{NO_{\partial\omega k}}$：针对第 $\omega$ 种药材的第 $\partial$ 组的均值处理结果；

$NO_{\partial\omega km}$：针对第 $\omega$ 种药材的的第 $\partial$ 组的每个数据。

随后，为了估计第 $\omega$ 个未知产地药材与第 $n$ 个产地的逼近效果，对相应波段的差值绝对值的方差进行估计：

相应波段的差值的绝对值：

$$D_{\partial\omega n} = \left|\overline{NO_{\partial\omega k}} - \overline{OP_{n\partial k}}\right| \tag{4}$$

相应波段的差值的平均值：

$$\overline{D_{\omega n}} = \frac{\sum_{\partial=1}^{6} D_{\partial\omega n}}{6} \tag{5}$$

相应波段的差值的方差：

$$SD_{\omega n}^{2} = \frac{\sum_{\partial=1}^{6}\left(D_{\partial\omega n} - \overline{D_{\omega n}}\right)}{6} \tag{6}$$

**Table 1.** Data variance processing results for 11 producing areas
**表 1.** 针对 11 个产地的数据方差处理结果

| OP | 3 | 14 | 38 | … | 318 | 618 |
|---|---|---|---|---|---|---|
| 1 | 0.041229258 | 0.076267529 | 0.196008508 | … | 0.150713629 | 0.015073855 |
| 2 | 0.021294789 | 0.048420743 | 0.52132589 | … | 0.010298385 | 0.106432349 |
| 3 | 0.038354827 | 0.076170891 | 0.20349037 | … | 0.141256111 | 0.014251981 |
| … | … | … | … | … | … | … |
| 9 | 0.050031399 | 0.099868488 | 0.417746227 | … | 0.049239923 | 0.08598518 |
| 10 | 0.028460967 | 0.069955355 | 0.301700638 | … | 0.079006015 | 0.035101808 |
| 11 | 0.031578723 | 0.070726081 | 0.27886763 | … | 0.092444691 | 0.030137438 |

最终判断下列序号对应的未知产地药材的产地结果如下表 2：

**Table 2.** Origin of 15 kinds of medicinal materials
**表 2.** 15 种药材产地

| NO | 3 | 14 | 38 | 48 | 58 | 71 | 79 | 86 | 89 | 110 | 134 | 152 | 227 | 331 | 618 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| OP | 8 | 8 | 4 | 1 | 8 | 11 | 4 | 2 | 4 | 10 | 4 | 7 | 2 | 3 | 3 |

## 4. 未知药材的不同产地的鉴别模型

### 4.1. 模型建立

以给定的 A、B、C 三种种类的药材的近红外光谱数据作为参考依据，判定未知药材的种类和产地。利用暴力枚举法在 C++中建立循环遍历模型，首先筛选出同一光数处的 A、B、C 三种种类的药材数据，

选取药材种类所在列作为主要关键字，进行循环遍历，得到药材分类，其次选择药材产地所在列作为次要关键字再次进行循环遍历，最终得到目标药材的种类和产地。算法程序如下图 2：

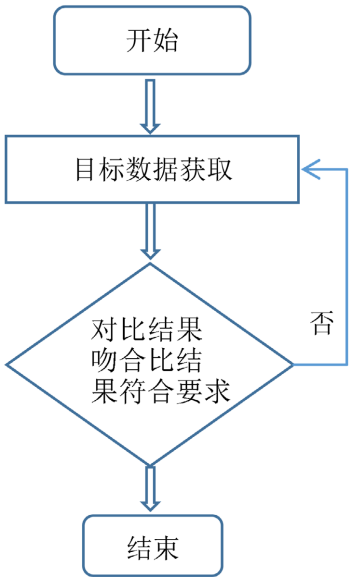



**Figure 2.** Flow chart of violence enumeration method
**图 2.** 暴力枚举法流程图

根据问题要求，根据给出的几种药材的近红外光谱数据判断出其种类和产地。首先以药材种类所在列为主要关键字，药材产地所在列为次要关键字，对数据进行排列，并利用暴力枚举法对数据进行分析。

## 4.2. 药材种类分析及结果

首先以药材种类作为分类，在 C++中进行循环遍历，如果目标在某一个光数处不满足数据所给出的范围，那么返回 false，如果满足所给范围，则循环至下一个光数处，直至遍历整个数组。若一直能满足数组范围，说明目标在每一处光数处都能满足此种类的吸光度，所以返回 true，证明此目标满此子种类。得到以下布尔型值(表 3)：

**Table 3.** Values of Boolean types
**表 3.** 布尔型的值

| NO | A | B | C |
|---|---|---|---|
| 94 | true | false | false |
| 109 | true | false | false |
| 140 | true | false | false |
| 278 | false | false | true |
| 308 | false | false | true |
| 330 | false | false | true |
| 347 | false | true | false |

即，A 种类药材为 NO94、109、140；B 种类药材有 NO278、308、330；C 种类药材为 NO347。

### 4.3. 药材产地分析及结果

在同种药材类别下对产地进行分类，按照附件所给数据，A 种类共分为 5 个产地，为 $A_a$ $(a=1,\cdots,5)$；B 种类共分为 16 个产地，为 $B_b$ $(b=1,\cdots,16)$；C 种类共分为 4 个产地，为 $C_c$ $(c=1,\cdots,4)$。将目标药材循环遍历数组，根据程序运行返回 false 或者 true，得到产地初步分类(表 4)：

**Table 4.** Preliminary classification of origin
**表 4.** 产地初步分类

| NO | 94 | 109 | 140 | 278 | 308 | 330 | 347 |
|---|---|---|---|---|---|---|---|
| OP | $A_3$、$A_4$ | $A_4$、$A_5$ | $A_4$、$A_5$ | $C_1$ | $C_3$ | $C_4$ | $B_{10}$、$B_{11}$、$B_{12}$、$B_{10}$、$B_{10}$、 |

由于 NO94、109、140、347 出现多个产地的判定结果，因此为了精确判断出产地，求出该目标在每一处光数处与数组中相应产地的中位数的方差，最小方差则为对应产地，得到药材最后的产地分类(表 5)：

**Table 5.** Final classification of origin
**表 5.** 产地最终分类

| NO | 94 | 109 | 140 | 278 | 308 | 330 | 347 |
|---|---|---|---|---|---|---|---|
| OP | $A_3$ | $A_4$ | $A_4$ | $C_1$ | $C_3$ | $C_4$ | $B_{10}$ |

## 5. 总结

利用 K-means 聚类分析法对数据进行迭代处理，可以按照特征对数据进行分类，同时利用方差公式对种类进行吻合分析，最后利用暴力枚举法以及方差估计进一步得到种类和产地的最优匹配结果。在对未知药材的种类进行判定时，对相关数据进行了均值简化处理，虽不失精确性，但在接下来的改进中我们可以对给出的全体样本数据直接进行方差处理，这样能够更加精确地判断出产地和种类。

## 基金项目

2021 年度安徽高校自然科学基金(KJ2021A1447)。

## 参考文献

[1] 姜启源, 谢金星, 叶俊. 数学建模[M]. 第三版. 北京: 高等教育出版社, 2005.
[2] 向学洪. 中药鉴别依据及方法[J]. 中医临床研究, 2016, 8(15): 145-146.
[3] 张岳, 罗文汇, 孙冬梅. 红外光谱技术在中药配方颗粒中的研究进展[J]. 中医药导报, 2014, 20(2): 99-101.
[4] 薛巍敏. 标准化让“道地药材”更地道[N]. 甘肃经济日报, 2021-09-07(001).