中药材的鉴别

摘要： 本文主要利用Excel对数据的处理、Matlab中的APP机器学习，聚类分析、主成分分析和线性判断等理论，分析和解决了基于红外光谱数据，对中药材种类及产地的鉴别。

在问题1中:通过Excel分析附件1中不同种类药材的特征性和差异性，利用特征性和差异性去对药材进行鉴别，由于数据很多，应当进行主成分分析，提炼出最有影响力的数据，进行聚类分析，也为后面几问打下基础。

在问题2中：先对附件2中的数据利用Excel的数据筛选功能，选择出需要识别的药材，通过Matlab绘图观察已知种类药材的数据是否有异样进行特征提取，通过Matlab中app的机器学习把需要识别的药材进行识别，最后通过准确率最高的模型来识别结果。

在问题3中：利用Excel对附件3已给出药材近红外和中红外光谱数据进行筛选，首先将近红外和中红外数据依照问题2中的处理方法分别进行处理，将处理后的数据利用Matlab绘图观察数据是否有异常。然后利用Matlab机器学习训练模型，再将未知药材产地近红外和中红外数据分别带入后进行分类，比较近红外和中红外的结果，选择差异最小的，得出最终结果。

在问题4中：利用Excel对附件4中的数据首先按照产地进行筛选，提取出已知产地和未知产地的数据，将已知产地数据导入Matlab机器学习中训练模型，选择准确率最高的模型进行识别未知产地的数据，得到结果。同理，再对附件4中的数据按照类别筛选，得到结果。最后进行汇总。

关键词： Excel数据筛选 Matlab机器学习 红外光谱 分类 识别

一、 问题重述

1.1背景知识

从古至今中药为我国人民健康做出了重要贡献甚至造福了全人类。古典书籍对中药材的记载更是数不胜数，不仅使其发展为中国传统医药宝库的重要组成部分还为中华民族优秀文化奠定了基础。随着中医药文化走向全世界，鉴别中药材的种类和产地成为了我国研究的重要方向。其中红外光谱分析可以作为鉴别中药材种类及产地的重要特征。中药材的道地性以产地为主要指标，相同种类的药材在不同产地的红外光谱数据在同一光谱内比较接近，不容易鉴别，可以通过近红外、中红外光谱数据相互验证来进行鉴别。

1.2问题重述

问题1：不同种类的中药材它们的中红外光谱数据具有一定的特征性和差异性，根据附件1提供的数据来分析他们的特征性和差异性，从而鉴别药材的种类。

问题2：同一种药材在不同产地的中红外光谱数据会有不同，根据附件2提供的数据分析不同产地同种药材的特征性和差异性，来鉴别同种药材的产地。

问题3：不同产地的同种药材有的在近红外区数据差异明显，有的在中红外区数据差异明显，根据附件3提供的数据分析该药材近红外区和中红外区的特征性和差异性，两者相互验证，鉴别药材的产地。

问题4：附件4提供了几种药材的近红外光谱数据，且有已知药材的产地和类别，通过产地和类别的相互分析，鉴别未知药材的类别和产地。

二、问题分析

2.1预备知识

红外光谱图：纵坐标为吸收强度，横坐标为波数。

红外光波数的划分：近红外（14000-4000cm^{-1}），中红外（4000-400cm^{-1}），远红外（400-10cm^{-1}）。

2.2对问题 1 的分析：

根据中红外光谱数据，来鉴别药材的种类。首先先对附件1中的数据进行处理，探究不同种类药材的差异性和特征性，由于数据太多，我们用主成分分析法，寻找影响最大的一部分数据，然后利用这一部分的数据进行特征的寻找，最后根据特征性进行聚类分析。

2.3对问题 2 的分析：

利用Excel中的筛选功能，对附件2 的数据进行筛选，选出已知产地的药材和未知产地的药材；然后用已知产地的数据，进行机器学习，用确率最高的模型来识别未知产地的药材。

2.4对问题 3 的分析：

不同产地的同一种药材在同一波段内的光谱比较接近，使得光谱鉴别的误差较大。另外，有些中药材的近红外区别比较明显，而有些药材的中红外区别比较明显。根据题目中所给的图2和图3来看不同产地的同种药材的鉴别不能只单单依赖于近红外或者中红外，需要两者的数据相互验证来鉴别药材产地。利用Excel对近红外的数据筛选，选出已知产地和未知产地的数据，同问题2的分析，进行机器学习，训练模型来识别未知产地的数据。同理，可以根据中红外数据求出产地。两者结果结合，相互验证。

2.5对问题4的分析：

问题四要求根据数据分别求出产地和类别，根据问题3的分析步骤，我们首先按照产地这一指标，将附件4的数据分为未知产地的和已知产地的，然后在Matlab中进行机器学习，训练出模型来识别未知产地的药材。同理，将种类作为指标，即可得到未知种类药材的种类。两者再结合相互验证。

三、问题假设

1.假设所有附件中的数据全部准确无误。

2.假设数据中的红外光谱特征能够完全反应药材的产地。

3.假设药材红外光谱特征完全能够从数据中反映出来药材类别。

4.假设数据分析时的误差极小不影响最终结果。

四、符号说明

符号	名称
NO	编号
Class	产地
OP	种类

五、模型建立与求解

问题1 模型建立与求解

特征波长的提取：

以652～3999cm^{-1}波数为样本，药材的编号1～425为指标进行主成分的分析。得到如图1所示。X轴表示主成分，y轴表示所占的百分比

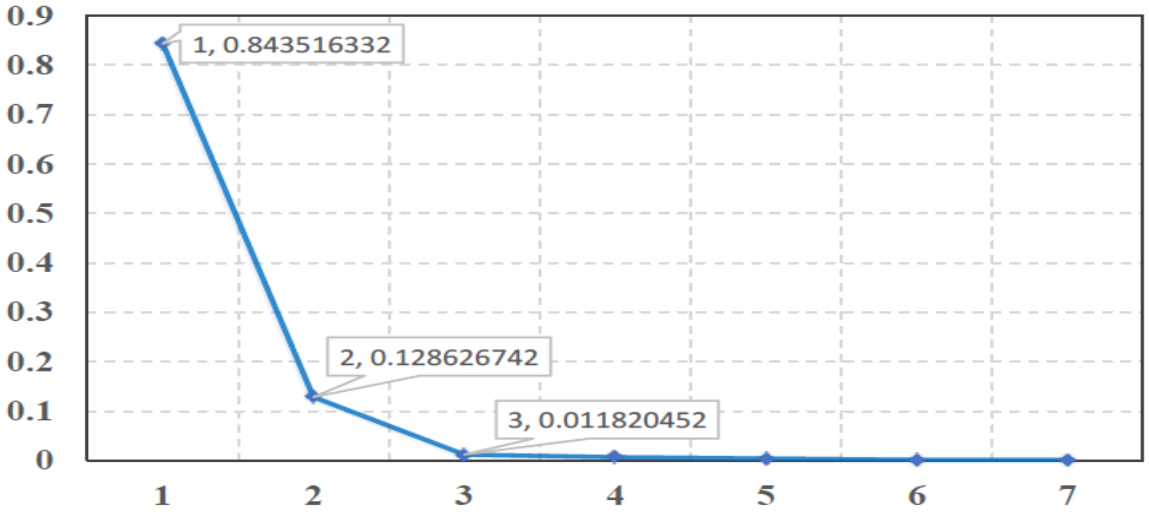

图1 特征的提取

由此看出前面三个成分的占比率极大，直接取这三个主成分，他们可以很好的反映出三个特征值。然后在以652～3999cm^{-1}波数为样本，三个主成分值为指标分析如图2所示。

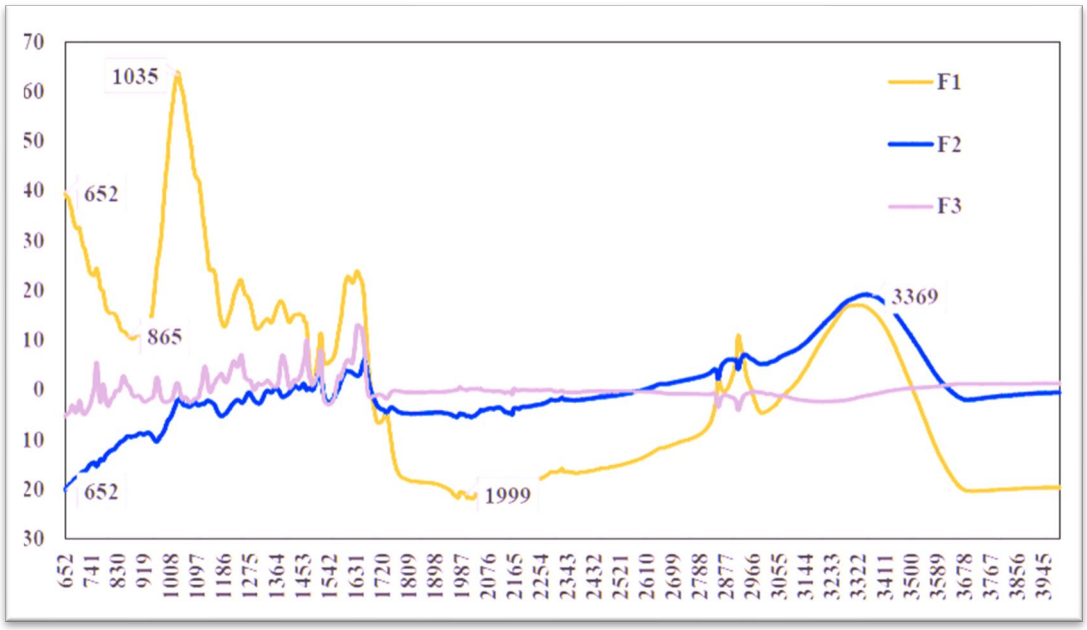

图2 占比最大的三个主成分

由图2可知，根据F1曲线选择特征为652、865、1035、1999、3369；根据F2曲线选择特征为652、3369。F3曲线比较平滑，无法选出特征来。根据三个主成分综合起来选择652、865、1035、1999、3369为特征。

聚类分析：

首先对数据中的异常值进行剔除，如图3所示，其中编号64、136、201的药材应被剔除。

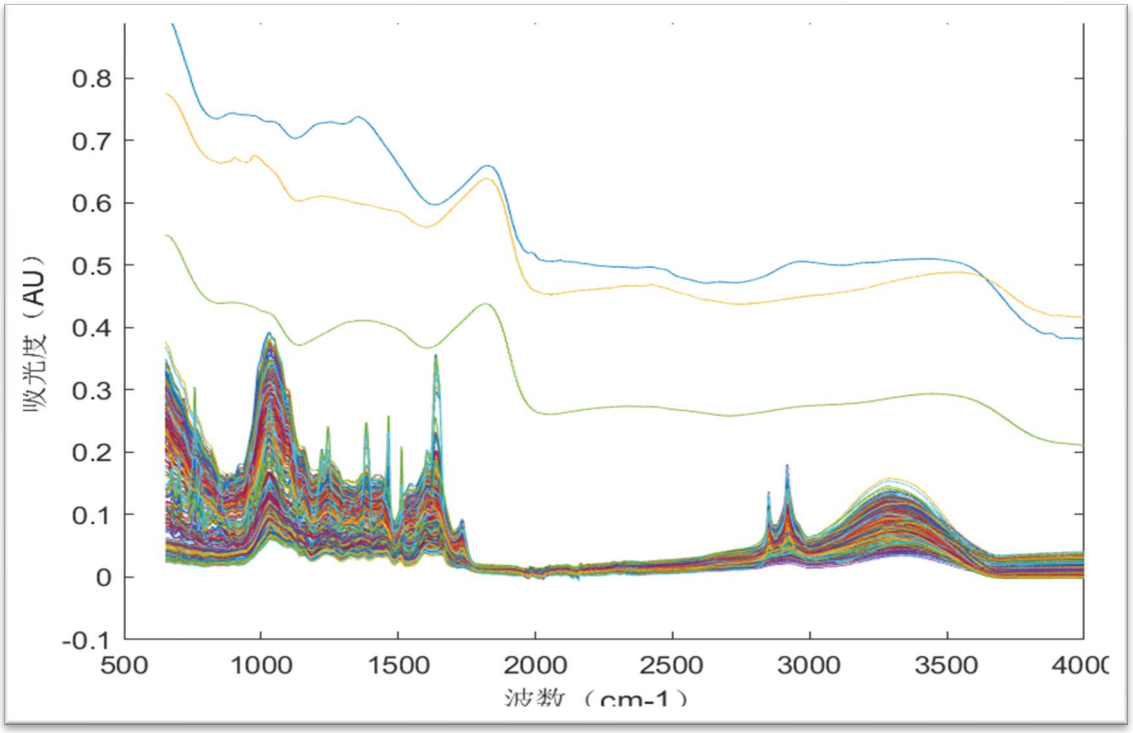

图3 异常数据剔除

然后进行聚类分析，将64、136、201药材归为一类。其他药材的分析图，如图4所示

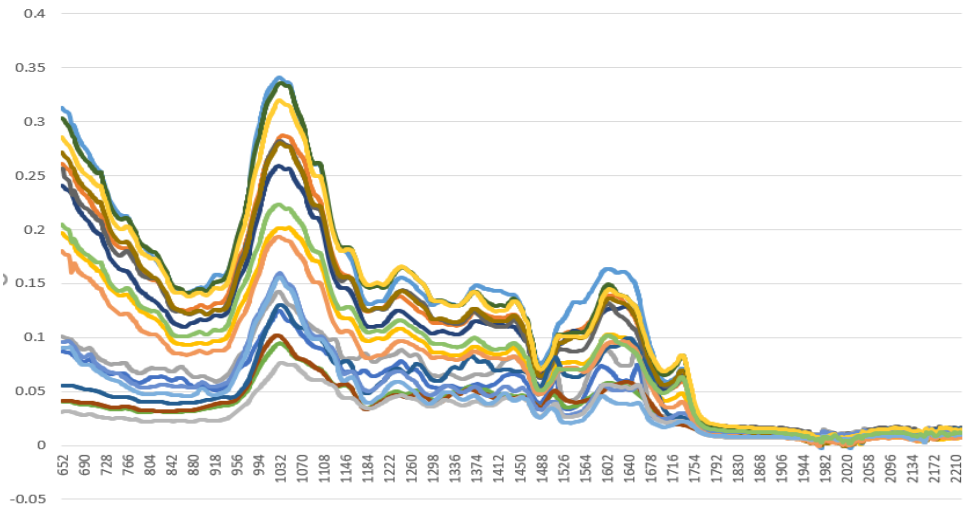

图4 药材的分类

最终得到如下四个分类，如图5所示：

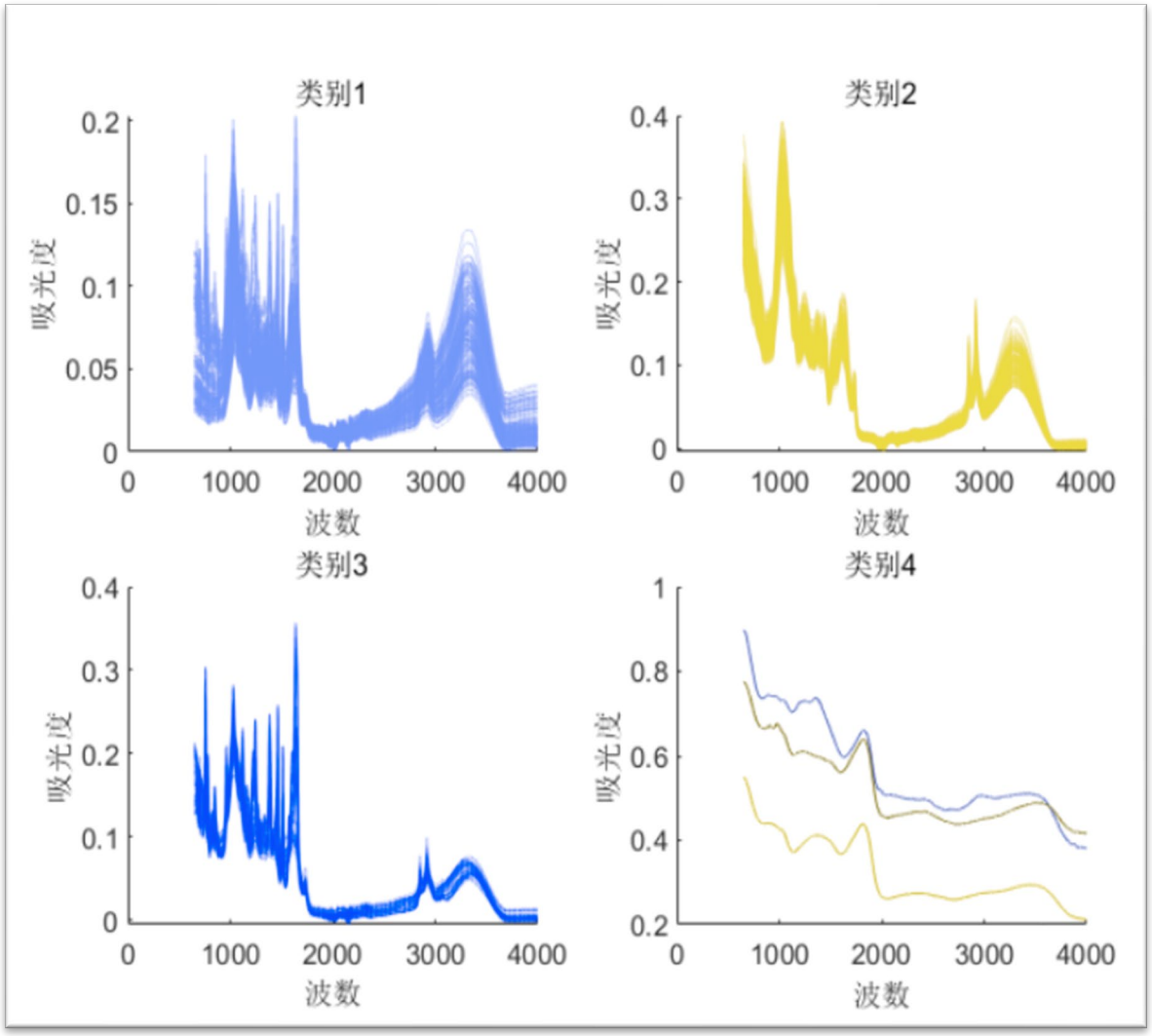

图5 最终分类结果

问题2 模型建立与求解

数据的分析处理：

利用Excel 中的 筛选功能，先选出未进行识别产地的药材并剔除出来，然后根据剩余数据在matlab中绘图，如图6所示：

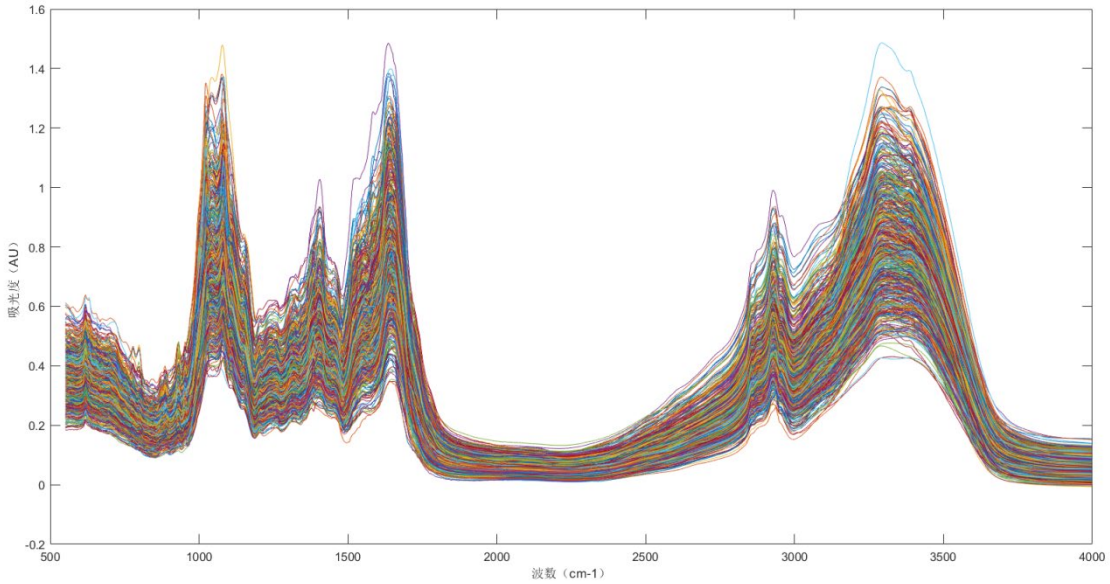

图6 观察数据是否有异常

根据图像发现该组数据中并没有需要剔除的数据，将该数据带入matlab机器学习的模型中进行训练，得到图7

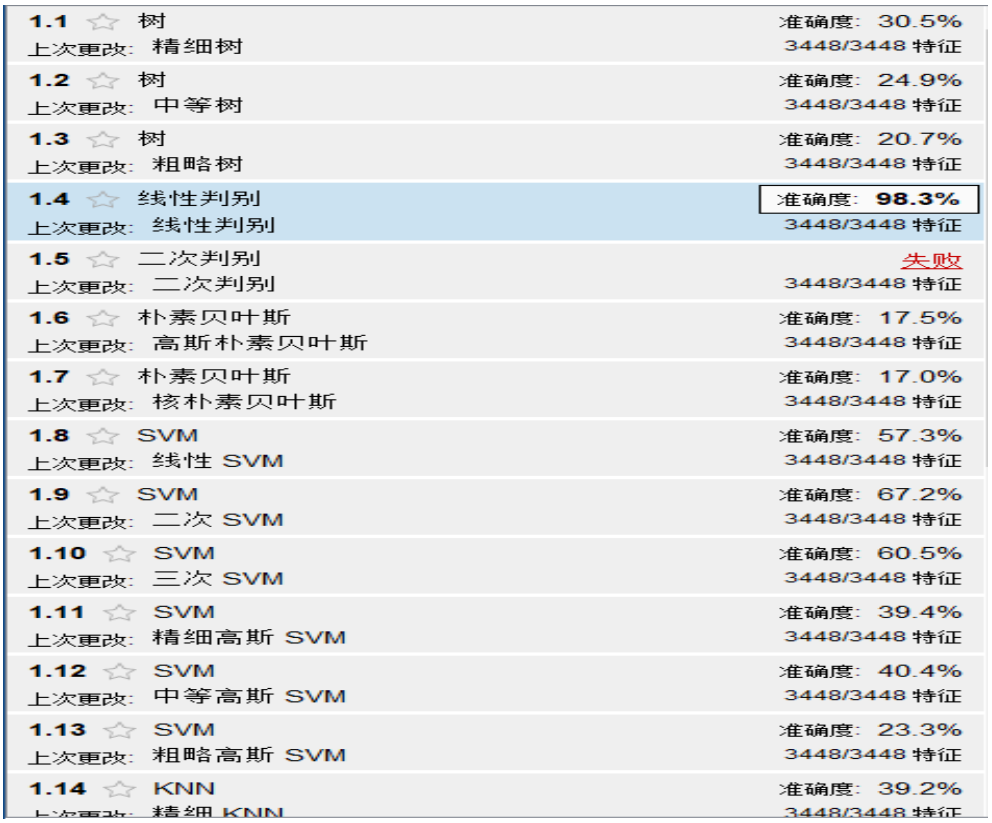

图7 模型的选择

我们选择准确率最高的模型线性判别。将模型导出，在Matlab命令行窗口中导入需要识别的数据，最终得出结果如下表1所示

No	3	14	38	48	58	71	79	86	89	110	134	152	227	331	618
OP	6	1	4	7	10	6	9	11	3	4	9	2	5	8	3

表1

问题3 模型建立与求解

首先将近红外和中红外的数据依照问题二中对数据处理的方法进行处理，然后将处理过后的数据在matlab中进行绘图，得到图8和图9

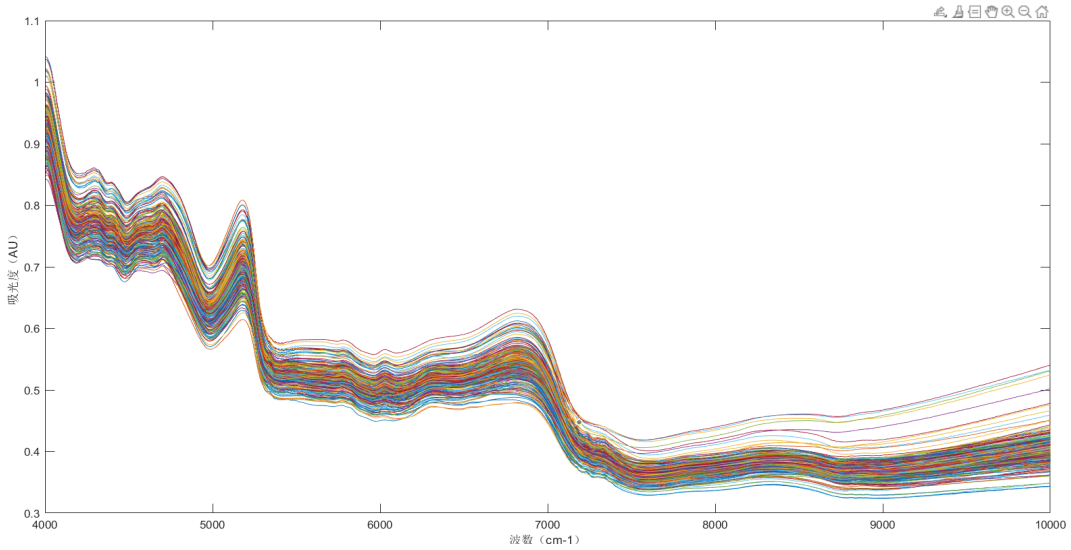

图8近红外光谱

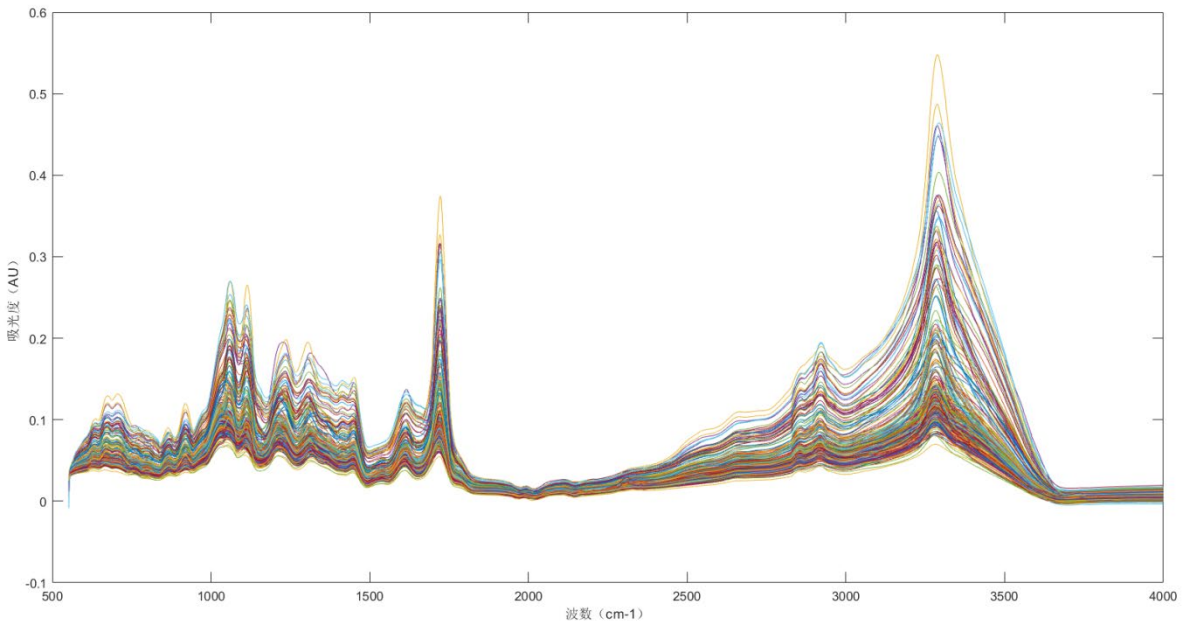

图9中红外光谱

先将已知产地的药材的近红外的数据带入matlab中机器学习的模型中，再将未知产地的药材的近红外数据带入分类，得到如下表2所示：

NO	4	15	22	30	34	45	74	114	170	209
OP	17	11	1	2	16	3	4	10	9	14

表2

同理将中红外的数据进行分类，得到如下表3所示

No	4	15	22	30	34	45	74	114	170	209
OP	17	11	1	2	16	3	4	10	9	14

表3

根据两个表中的数据，近红外和中红外的分类结果最终一致，最终结果如下表4

No	4	15	22	30	34	45	74	114	170	209
OP	17	11	1	2	16	3	4	10	9	14

表4

问题4 模型建立与求解

首先利用Excel表格对附件4数据按产地进行筛选，分成未知产地和已知产地两个表格，然后将未知产地与已知产地的表格数据导入到Matlab中，利用机器学习小程序进行产地模型的训练，

然后在将未知的产地的数据进行分类识别，即可得到未知产地药材的产地。如下表5所示：

No	94	109	140	278	308	330	347
Class	A	A	A	C	C	C	B

表5

同理，用按种类筛选的数据进行机器学习，训练模型得到结果如下表6所示：

No	94	109	140	278	308	330	347
OP	5	3	1	1	4	5	11

表6

整理两个表格中的数据得到最终结果如表7所示：

No	94	109	140	278	308	330	347
Class	A	A	A	C	C	C	B
OP	5	3	1	1	4	5	11

表7

六、模型的评价与推广

6.1模型的评价

本次构建的模型，运用了Excel对数据的处理和筛选，利用Matlab对数据进行绘图，使数据变得可视化，建立相关模型，得到红外光谱数据对不同种类药材，不同产地药材的影响因素，通过机器学习建立识别模型，可以根据红外光谱数据判断药材的种类及产地。图形与模型结合，形象生动，建模过程明了，对实际应用具有一定指导意义.

6.2模型的优点

我们多方面考虑实际制作模型，通过分析各个附件中大量的样本数据进行大数据分析，使得模型更加形象，更灵活，且尽可能缩小误差对数据不产生影响;在建立模型过程中，我们会运用数学软件进行运算，如：MATLAB、Excel等进行数据拟合，这样使得我们运算误差更小，反应更直观，过程更清晰、结果更明确、效果更理想化；充分运用附件中的各项数据,重复分析，不仅能计算速度快还对模型参数有动态确定的能力，精度较好。

6.3模型的不足之处

由于数据量过于庞大，进行处理时难免会造成一些误差，从而容易对最终结论造成影响，但建立的模型具有一定的普适性，可以解决文中的问题。

6.4模型的推广

在人体健康领域，中药材发挥着它独特的魅力和价值。对用光谱特征对药材鉴别的研究与我们的身体健康可以说是紧密相关，不仅是鉴别药材的重要方向更是推动中药发展的动力和根本而且还通过本次构建模型的过程，为鉴别中药材的类别和产地的问题找到了独到的方法，且该方法对其他数学问题及模型仍可用所以说本文的模型建立具备实用性和普遍性。

七、模型的改进

由于所处理数据过于庞大且所用工具对处理数据的能力有限所以在进行数据筛选和处理时出现一定的不足之处，故应采用更专业化和处理数据能力强的软件进行数据分析。同时对庞大的数据进行计算时不能准确无误可能有差值出现的问题，对此应选用更专业处理数据能力更强的软件进行处理并多次分析重复检测直至误差最小化，从而达到理想效果的数据。因为存在理想化假设，对数据的真实性难免会造成影响，可以将假设做的更合理更客观，尽可能的减少次要因素对模型的影响。

参考文献

[1] 何勇，基于中红外光谱和化学计量学算法鉴别核桃产地及品种，浙江大学生物系统工程与食品科学学院，2019年9月。

[2] 李波霞，红外光谱技术及化学计量学在党参、当归定性定量模型研究中的应用，兰州大学，药物信息技术，2009年5月。

[3] 阿培丁，机器学习导论，机械工业出版社，2009年6月。

[4] 姜启源，谢金星，叶俊，数学模型，北京：高等教育出版社，2005

附录

附录1：见支撑材料。

附录2：Matlab绘图程序：

```
load("X.mat");    读取文件名
[a,b]=size(X);    设置矩阵大小
figure(1);
for i=2:a
    plot(X(1,2:b),X(i,2:b));
    hold on;
end
xlabel("波数（cm-1）");
ylabel("吸光度（AU）");
```

Matlab中机器学习识别数据程序：

>>yfit=num2.predictFcn() 括号中填需要识别的数据