类结果见附录)。并且在分布图中我们可以判断经过 SG 一阶导数平滑后聚类的效果最好，经过 SNV 处理和 SG 平滑后的数据聚类效果最差。

所以使用经过 SG 一阶导数平滑处理后的数据进行分为 6 类，并对每一类的药材进行编号，将每一类药材抽取一个最接近于聚类中性点的特征明显的样本进行特征值分析，再将多个样本进行画图比较差异值分析，见图 9。

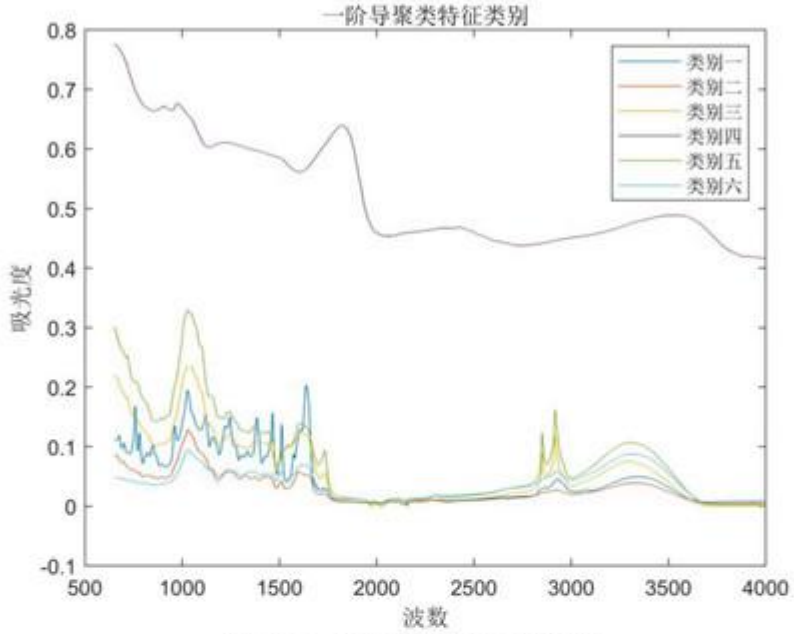

图 9 SG 一阶平滑导数聚类特征类别

根据上图观察，我们可以发现，在 SG 一阶平滑导数处理后的进行聚类的特征图中，第四类中药材的整体吸光度明显高于其他几类，特征最为明显；第六类中药材在 1600 波数出现最高吸收峰后，趋势变得平稳，类别二、三、五的大致趋势相似，但在峰差、峰宽及吸光度的大小是存在一定差异，可能是因为这几种中药材的生长环境相似。

除第四类与第六类药材外中药材外，其余药材的特征为：在波数为1100处出现最高吸收峰，且波数为 1700 之后，随波数增大，吸光度增幅逐渐放缓趋势，没有明显的“峰-谷”变化；波数在 2800 处时吸光度形成吸收谷且波数在 2900 时，存在一个弱反射峰。

表 2 不同种类的光谱差异表

	种类一	种类二	种类三	种类四	种类五	种类六
总峰数	1-2	1	2-3	1	2-3	1-2
最高峰位	1700-2000	1100-1300	1100-1300	652-700	1100-1300	1200-1400
峰强	较强	较弱	较强	较弱	强	弱
峰形	尖	钝	尖	钝	尖	钝

但还不能单纯地通过光谱图对中药材进行差异性判定，我们结合化学计量方法使用 SID 散度分析进一步分析六个中药材之间的差异性。

5.1.4 光谱信息散度 (SID)

SID 可用于表征不同样本光谱间相似性。不同样本光谱间的 SID 越小，这两

个样本的相似度越高，样本数据越接近,使用 SID 进行中药材的差异分析是非常合适的。

对于光谱 $x_1 = (x_{11}, x_{12}, ..., x_{1l})$ 和光谱 $x_2 = (x_{21}, x_{22}, ..., x_{2l})$ 可以得到两条光谱的概率向量分别是 $q = (q_1, q_2, ..., q_i)^T$ 和 $p = (p_1, p_2, ..., p_i)^T$。

其中 $q_i = x_{1i} / \left(\sum_{i=1}^{l} x_{1i} \right), p_i = x_{2i} / \left(\sum_{i=1} x_{2i} \right)$ 。

根据光谱信息理论，可以得到 x_1 和 x_2 的自信息为:

$$l_i(x_1) = -\log q_i \tag{8}$$

$$l_i(x_2) = -\log p_i \tag{9}$$

可以得到 x_2 关于 x_1 的相对熵：

$$D(x_1 \| x_2) = \sum_{i=1}^{l} q_i \log\left(\frac{q_i}{p_i}\right) \tag{10}$$

两者散度的计算公式如下：

$$SID(x_1, x_2) = D(x_1 \| x_2) + D(x_2 \| x_1) \tag{11}$$

通过计算，我们得到六个种类的之间相对的 SID 值，见下表：

表 3 六个种类的 SID 表

	种类一	种类二	种类三	种类四	种类五	种类六
种类一	0	0.0344	0.0362	0.2190	0.0460	0.1415
种类二	0.0344	0	0.0118	0.1571	0.0109	0.0907
种类三	0.0362	0.0118	0	0.2047	0.0028	0.1233
种类四	0.2190	0.1571	0.2047	0	0.1997	0.1271
种类五	0.0460	0.0109	0.0028	0.1997	0	0.1204
种类六	0.1415	0.0907	0.1233	0.1271	0.1204	0

由上表可知，种类一、二、三、五与种类四差异最大，数据最不接近，根据推测可能是因为地域原因，地理位置较远；种类一与种类六种类对种类二的数据 SID 值相对最小，分别为 0.0344 与 0.0907；而种类二与种类三对种类五 SID 值相对较小，分别为 0.0109 与 0.0028。所以总结：种类一、二、三、五、六地理位置应该相隔不远，环境气候相似，导致中药材的化学组成成分相差不大。

5.2 问题二模型建立与分析

问题二需要对某一种药材的产地进行鉴别，并分析不同产地的药材的差异和特征。附件2给出了问题二所需要的某一种药材的中红外光谱，我们需要先对中红外光谱数据进行去噪处理，根据问题一中的比较，我们得知使用标准正态变量变换（SNV）处理效果不够理想，所以参考问题一中的去噪过程选择采用SG与其一阶平滑导数和二阶平滑导数对附件2进行去噪。具体过程见流程图：

图 10 模型流程图

5.2.1 不同产地的中药材特征值和差异性分析

由于附件 2 给出的是某一种中药材不同产地的中红外光谱数据，所以在十一个地区随机选取了一个样本进行特征值比较与差异值分析，见下图：

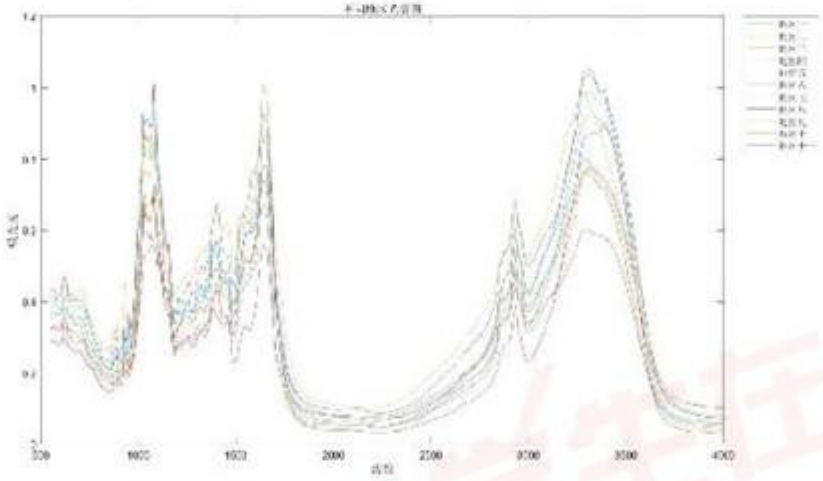

图 11 十一个产地药材的光谱特征图

根据上图观察，我们可以发现十一个产地药材的光谱特征图大致趋势是一致的，但是各药材的最大吸光度、峰高、峰宽，峰差不同，所有的药材分别在波数为1050、1600、3400左右出现三个明显峰。

表 4 不同地区药材的光谱差异表

地区	一	二	三	四	五	六	七	八	九	十	十一
峰数	3	3	3	3	3	3	3	3	3	3	3
峰位	3300-3400	1600-1700	3300-3400	3200-3300	3100-3400	1600-1700	1600-1700	3300-3400	1100-1200	3200-3300	1100-1200
峰强	较强	较弱	较弱	较弱	强	较强	较强	强	较弱	较强	强
峰形	尖	钝	钝	钝	尖	尖	尖	尖	钝	尖	钝

5.2.2 通过 SID 值对其进行定量分析

由SID表可知（见附录），地区一与地区十一的SID值最大，为0.0152数据最不接近，根据推测可能因为土壤、天气的原因造成药材的物理性质发生了变化；地区三与地区六SID值最小，为0.0004。因此可以推断：种类一、二、三、五、六地理位置应该相隔不远，环境气候相似，导致中药材的化学组成成分相差不大。

5.2.3 基于主成分分析（PCA）提取特征值

光谱数据去噪后，再采用主成分分析法（PCA）对中红外光谱进行降维，提取光谱特征，其贡献率作为代表性的光谱信息，见下图：

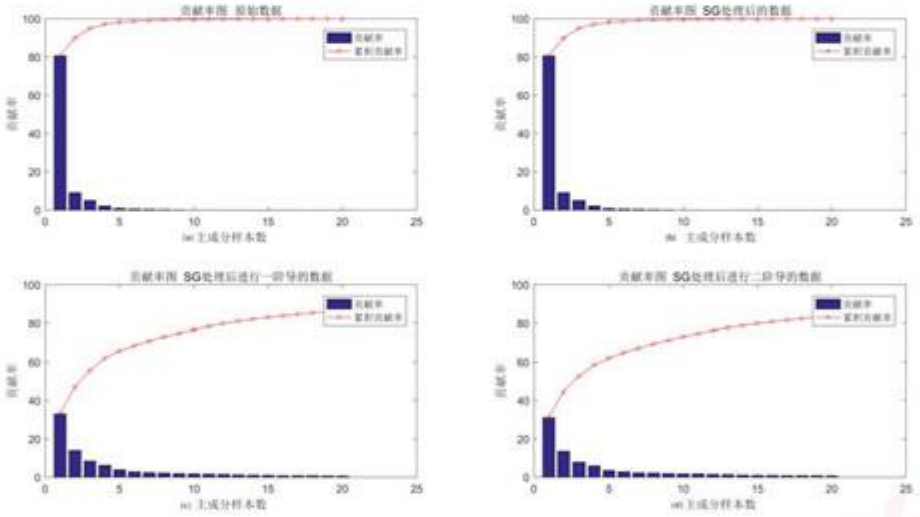

图 12 附件 2 中药材光谱数据进行 PCA 特征抽取之后的主成分贡献率

其中从左往右依次为：(a) 原始数据进行 PCA 特征抽取；(b) SG 平滑后的数据进行 PCA 特征抽取；(c) SG 平滑后一阶导数的数据进行 PCA 特征抽取；(d) SG 平滑后二阶导数的数据进行 PCA 特征抽取

观察我们发现，进行降维后，原始数据中主成分 1 达到了 80.70%，前三个主成分占据了 94.90%的信息量，在经过 SG 平滑后的数据提取主成分，前三个主成分占据了 94.89%的信息量，而在 SG 平滑后一阶导数和二阶导数的 PCA 特征抽取，前三个主成分分别占据了 55.52%及 52.54%。由此可知，光谱数据在经过一阶 SG 平滑导数与二阶 SG 平滑导数 PCA 降维后需要保留的主成分个数更多，见下表：

表 5 附件 2 进行 PCA 后主成分保留个数

	原始数据	SG 平滑算法	SG 一阶平滑	SG 二阶平滑

保存主成分的个数	3	3	14	17
累计贡献率	94.90%	94.89%	82.04%	81.51%

5.2.4 样本均衡性分析

为了防止出现过拟合现象，我们对样本的均衡度进行分析，结果见下表：

表 6 问题二的样本均衡度

OP	1	2	3	4	5	6	7	8	9	10	11
样本数	67	59	67	88	29	87	50	59	31	66	55

由于样本数相差不大，所以我们暂不考虑对样本数的统一。

5.2.5 鉴别分类器构建

在进行特征的选取后，我们需要利用分类器进行对选取的特征的学习，分类器可以构建出对中药材产地的鉴别模型，且不同的分类器具有不同的特性，即使采用的同一组数据进行训练，得到的结果也往往有所差异。因此我们主要采用数据挖掘中常见的贝叶斯分类器（NB）、决策树算法（DT）、K近邻分类器（KNN）、支持向量机（SVM）和线性判别分类器（LDA）来构造五个不同的鉴别分类器。使用多种分类器进行训练，将不同的结果进行对比，找到效果最好的分类器，从而更高效更准确地实现对药材产地的鉴别。

(1) 朴素贝叶斯分类器(NB)

朴素贝叶斯分类器(NB)是基于概率分布的分类器。贝叶斯决策理论是解决模式分类问题的基本途径之一，意为决策问题可以转化为概率分布的形式来表达，假设其相关的先验概率是已知的，根据贝叶斯公式，在已知其先验概率的前提下，我们可以推导出其后验概率，即

$$P(c_k|\ x)=P(c_k)\times P(x|\ c_k)\times P(x)^{-1} \tag{12}$$

上式中c_k表示第k个类别，即表示某一个中药材的产地。x表示位置的样本，$P(c_k)$表示该产地在样本中出现的概率，$P(c_k|\ x)$表示在该产地条件下，测试样本x的概率密度函数。通过公式（8），得到该测试的中药材样本属于某一产地的后验概率。通过选择最大的概率值，贝叶斯分类器将中药材样本鉴别为最可能的产地。

朴素贝叶斯分类的算法步骤：

1) 设$x=\{a_1,a_2,...,a_m\}$为一个待测样本集，a_i为x的第i个特征属性，x包含m个特征。
2) 设中药材产地类别集合$C=\{c_1,c_2,...,c_n\}$。
3) 通过公式（8），分别计算出$P(c_1|\ x)$，$P(c_2|\ x)$,...，$P(c_n|\ x)$。
4) 如果$p(c_k|\ x)=\max\{p(c_1|\ x),p(c_2|\ x),...,p(c_n|\ x)\}$，则$x\in c_k,1\le k\le n$。

(2) K近邻分类（KNN）

KNN的分类的过程主要是基于样本之间的距离比较即从训练样本中找出K个与其最相近的样本，通过比较K个样本中所属类别的个数多少，来判定测试样本的类别，在使用KNN进行分类时，选择的样本的个数不同，使用样本进行比较后，KNN算法需要我们人为决定K的取值，K的取值不同，也有可能会对识别结果造成

差异，在本文中我们主要选择欧式距离作为样本件距离的度量。

(3) **线性判别分析（LDA）**

LDA 其基本思想是寻找一投影方向，使训练样本投影到该方向时尽可能具有最大类间距离和最小类内距离，设训练样本中有c个类别，根据 Fisher's 线性判别准则，这需要$c-1$个判别函数。也就是将d维的特征空间向$c-1$维空间做投影。

(4) **决策树模型（DT）**

决策树模型采用树形结构，将特征属性和最终的决策值进行联系，每一个内部的节点表述一个输入变量，数分支代表一个测试输出，树叶代表分类的结果，根据输入，通过根到叶之间的路径判断得到一条分类规则。

决策树往往采用自顶向下的方式生成，从根节点开始，通过对原始的数据集进行逐个属性测试并分裂为子集得到。这一过程在每一个生成的子集上重复递归地进行，直到某一个节点上的所有样本都被划分到同一个类别，或者对原始数据不能再进行有效的分割时结束。

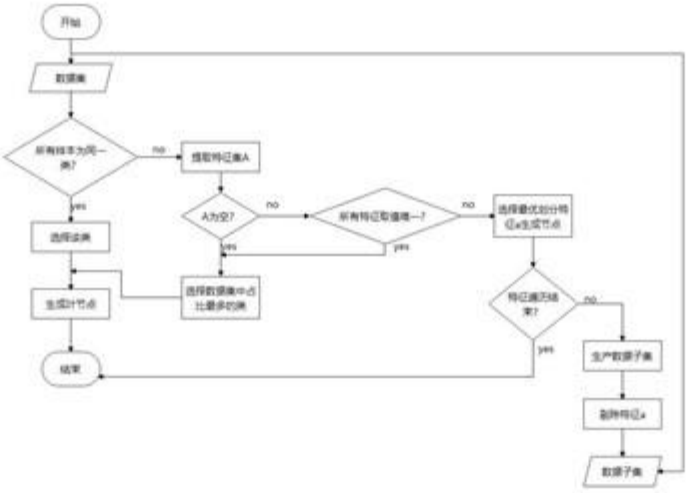

图 13 决策树模型流程图

(5) **支持向量机模型（SVM）**

支持向量机的主要思想是找到一个超平面是其尽可能多的将两类数据分开，同时使两类数据点距离分类面最短。

根据给定的训练集：

$$T=\{[a_1,y_1],[a_2,y_2],\cdots,[a_l,y_l]\}\in(\Omega\times Y)^l \tag{13}$$

上式中，$a_i\in\Omega=\mathbf{R}^n$；$\Omega$称为输入空间，输入空间中的每一个点a由n个属性特征组成；$y_i\in Y=\{-1,1\},i=1,2,\cdots,l$。

寻找$\mathbf{R}^n$上的一个实值函数$g(x)$，以便用分类函数

$$f(x)=sign(g(x)) \tag{14}$$

判断任意一个模式x相对应的y值的问题为分类问题。

5.2.6 基于分类器交叉验证与性能评价模型

为了挑选出最好的分类模型，我们需要测试分类器的性能，并使用一些指标进行性能的检验。

我们采用 $M \times N$ 交叉验证法。将数据集划分为10个等分，然后使用9个数据子集的数据进行训练，用剩下的一个子集进行测试。以上过程遍历10次以尽可能确保每一份数据集都参加了训练和测试，并且整个实验过程也要重复5次以确保样本划分的影响最小，最后测试集结束测试时，我们可以训练5×10个模型并获得50个测试结果，然后将测试结果的平均值作为次交叉验证模型的指标。进行50次不同的测试以获得最为精确的评价，从而减少样本划分引起的误差。

在性能指标的选取上我们通过正确率来评价鉴别模型的总的性能：

$$P_a = \frac{n_r}{N_t} \tag{15}$$

其中 n_r 是所有被正确分类的样本，N_t 为测试样本的总数。

由于在实践中样本的分布是未知的，所有我们还需要使用混淆矩阵来概括分类器的性能，并通过混淆矩阵得到另外几个不同的评价标准，计算步骤如下：

首先计算每个类别下的precision和recall：

$$precision_k = \frac{TP}{TP + FP} \tag{16}$$

$$recall_k = \frac{TP}{TP + FN} \tag{17}$$

准确率（accuracy）代表分类器对整个严格不能判断正确的比重：

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{18}$$

计算每个类别下的 F_1 分数：

$$F_{1k} = \frac{2 \cdot precision_k \cdot recall_k}{precision_k + recall_k} \tag{19}$$

对每个类别下的 F_1 分数求均值，得到最终测评结果：

$$score = \left(\frac{1}{n} \sum F1_k \right)^2 \tag{20}$$

上式中，TP代表预测答案正确；FP代表预测为本类；FN代表预测为其他类标。

当分类器能正确识别全部测试样本是时，F_1 分数达到最大值为1，反之为0。

5.2.7 分类器性能评价

各鉴别分类器分别进行5×10次交叉验证后得到的平均识别率见下表：

表 7 四个分类器的产地鉴别准确率

分类器	原始数据	SG平滑	SG一阶平滑导数	SG二阶平滑导数
DT	31.8%	28.3%	54.0%	56.8%
KNN	34.8%	34.8%	97.3%	97.3%
NB	32.7%	33.7%	91.2%	93.0%
LDA	**30.9%**	**30.9%**	**92.9%**	**98.3%**

SVM	34.3%	33.9%	96.7%	98.2%

从上表中可以看出，LDA 在 SG 二阶平滑导数数据集上识别率达到整个鉴别准确率的最高值，为 98%，其次 SVM 与 KNN 的表现最好，KNN 在 SG 一阶平滑导数据集上达到了 97.3%的准确率，且除 DT 外，所有的分类器在数据经过 SG 一阶及二阶平滑后，准确率提到了 90%以上。说明在经过 SG 一阶和二阶平滑导数数据后，KNN、NB、LDA 的性能都得到很大的提升，只有 DT 的性能提升不是理想，应该是因为对测试数据的分类没有那么准确，即出现过拟合现象。

所以综合看来，经过 SG 二阶平滑导数处理后的数据使用 LDA 分类器的产地鉴别准确率最高，为 98.3%，于是我们采用 LDA 分类器对中药材的产地进行鉴别。

5.2.8 LDA 分类器的性能评价

在选定分类器后，我们对分类器的性能进行评价，通过混淆矩阵进行判定分类错误的样本。见图 14。

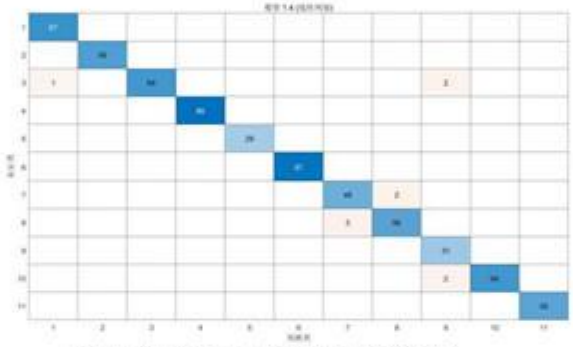

图 14 问题二 LDA 线性判别混淆矩阵

由图 14 我们可以看出，LDA 的判别中有 5 个地方判别错误，有一份样本应当属于第三类，却被判为第一类；有两份样本应当属于第三类、第七类、第十类，却被判为第九类第八类与第九类；有三份样本应当属于第八类，却被判为了第七类。

通过混淆矩阵得到的敏感度（*TPR*），特异性（*FPR*），F_1 分数的结果如下表所示：

表 8 各产地的评价指标

产地	1	2	3	4	5	6	7	8	9	10	11
TP	67	59	64	88	29	87	48	56	31	64	55
FP	1	0	0	0	0	0	3	2	4	2	0
FN	0	0	3	0	0	0	2	3	0	0	0
F_1 分数	0.99	1.00	0.98	1.00	1.00	1.00	0.95	0.96	0.94	0.98	1.00

通过对十一个产地的 F_1 分数进行求和取平均，计算得到 F_1 综合得分为0.98，说明药材产地的分类精确率很高。

最后计算，给出的药材产地鉴别为见下表：

表 9 问题二药材产地编号图

NO	3	14	38	48	58	71	79	86
OP	6	1	4	7	10	6	9	11

NO	89	110	134	152	227	331	618
OP	3	4	9	2	5	8	3

5.3 问题三的模型建立与求解

问题三需要我们根据同种药材的近红外数据与中红外数据，鉴别其产地并填写鉴别结果。

5.3.1 去噪处理与特征值提取

经过前两问我们发现原始数据和SG平滑后的数据分类效果相对于SG一阶平滑导数和二阶平滑导数不理想，于是我们仅采用SG一阶平滑导数和二阶平滑导数对近红外数据和中红外数据进行去噪。然后使用主成分分析（PCA）进行特征值的提取。在选取主成分时我们只采用指定解释方差达到95%的主成分，最后选取了54个主成分。

5.3.2 波段的选取与分类器的选择

因为药材会在近红外数据和中红外数据的光谱图上显示不同的效果，于是我们在使用以下三种方法选取波段：

1. 选取中红外光谱数据
2. 选取近红外光谱数据
3. 选取中红外光谱数据与近红外光谱数据

将选取的红外光谱数据进行数据集划分后利用分类器进行选取特征的学习，计算五种分类器的准确率，见下表：

表 10 五个分类器对不同波段红外数据分类后的产地鉴别准确率

分类器	近红外SG一阶平滑	近红外SG二阶平滑	中红外SG一阶平滑	中红外SG二阶平滑	整段红外SG一阶平滑	整段红外SG二阶平滑
DT	81.6%	78.4%	69.0%	70.6%	82.9%	73.9%
KNN	86.9%	85.3%	90.6%	90.6%	94.3%	92.2%
NB	83.3%	85.3%	87.8%	92.2%	91.0%	87.8%
LDA	**97.6%**	**97.6%**	**89.4%**	**92.2%**	**98.4%**	**97.1%**
SVM	91.0%	92.2%	91.0%	93.5%	91.3%	94.7%

从上表可知，近红外光谱数据经过 SG 一阶及二阶平滑后 LDA 分类器的准确率最高，都为 97.6%；中红外光谱数据经过 SG 一阶及二阶平滑后 SVM 分类器的准确率最高，分别为 91%与 93.5%；而选取的整段红外光谱数据经过 SG 一阶及二阶平滑后的准确率达到了 98.4%与 97.1%，综合所有的准确率来看，选取整段红外光谱进行训练后使用 LDA 分类器得到的产地鉴别准确率最高。所以我们选取经过 SG 一阶平滑处理后的近中红外结合数据使用 LDA 分类器进行中药材的产地鉴别。

5.3.3 样本均衡性分析

在使用分类器进行预测前，我们对样本数进行均衡性的检验

表 11 问题二各个产地的样本数

OP	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
近红外样本	14	14	14	14	15	15	15	15	14	14	14	15	15	14	15	14	14
中红外样本	14	14	14	14	15	15	15	15	14	14	14	15	15	14	15	14	14

由上表可得，每一个产地的近红外样本数与中红外样本数相差不超过一个样本数，说明均衡性非常好，可以直接进行预测。

5.3.4 基于LDA分类器的预测性能评价与鉴别

在选定分类器后，我们对分类器的性能进行评价，通过混淆矩阵进行判定分类错误的样本。

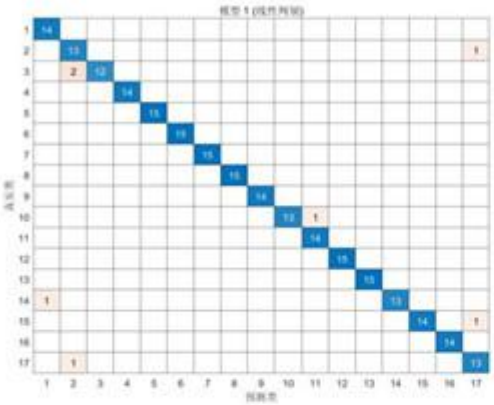

图 15 问题三 LDA 产地鉴别模型混淆矩阵

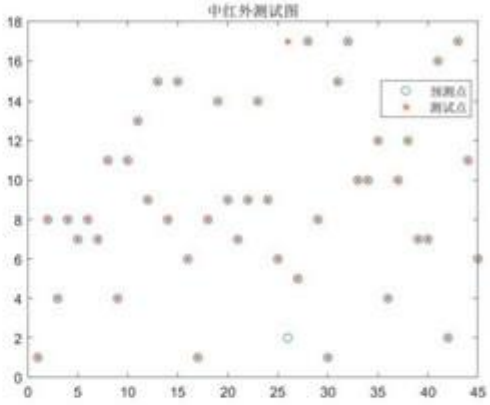

图 16 LDA 分类器分类器测试集仿真效果图

由图 15 的混淆矩阵我们可以看出，LDA 的判别中有 6 个地方判别错误，有一份样本应当属于第二类、第十四类、第十五类与第十七类，却被判到第十七类、第一类、第十七类与第二类；有两份样本应当属于第三类却被判为第二类。

由图 15的分类器预测效果图可知，将SG一阶平滑处理后的近中红外光谱数据作为测试点，除某个点的预测发生偏移外，其他测试点都在预测点的中心位置，使用此模型进行产地鉴别得到的预测点与测试点基本全部吻合，结果表明LDA分类器进行产地鉴别效果好且准确率高。

通过混淆矩阵得到的敏感度（TPR），特异性（FPR），F_1分数的结果如下表所示：

表 12 问题三的各评价指标

产地	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
TP	14	13	12	14	15	15	15	15	14	13	14	15	15	13	14
FP	1	3	0	0	0	0	0	0	0	0	1	0	0	0	0
FN	0	1	2	0	0	0	0	0	0	1	0	0	0	1	1
F_1分数	0.97	0.87	0.92	1	1	1	1	1	1	0.96	0.97	1	1	0.96	0.97

通过对十一个产地的F_1分数进行求和取平均，计算得到F_1综合得分为0.97，说明药材产地的分类精确率很高。

最后进行编号预测，给出的药材产地鉴别为见下表：

表 13 问题三的药材产地编号表

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

5.4 问题四的模型建立与求解

问题四要求我们根据附件4提供的几种药材的近红外光谱数据进行类别与产

地的鉴别。

由问题一和问题二提出的鉴别框架，分别对中药材的类别和产地进行建模预测。

5.4.1 基于近红外光谱的中药材类别鉴别模型：

分类模型准确率见下表：

表 14 对类别的预测准确率

分类器	SG一阶平滑导数	SG二阶平滑导数
DT	100%	100%
KNN	58.4%	100%
NB	100%	58.4%
LDA	100%	100%
SVM	**100%**	**100%**

由上表可知，在对类别的分类时，除KNN外，其他分类器的效果都达到了100%，而KNN预测准确率较低的原因可能是因为样本数据量少，不利于KNN分类器进行分类，因此我们随机选择SVM分类器对分类结果通过混淆矩阵与预测效果图进行数据可视化检验：

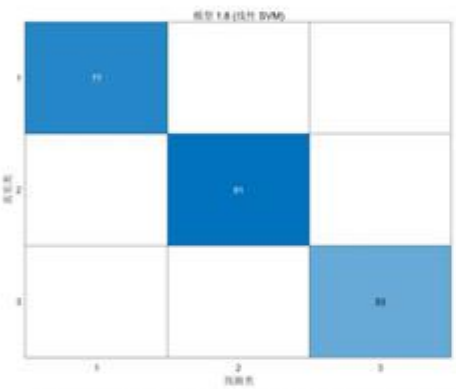

图 17 SVM 分类器的产地鉴别模型混淆矩阵

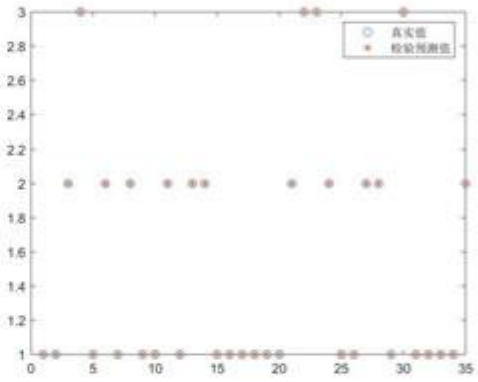

图 16 SVM 分类器测试集数据仿真结果

由上图可知，混淆矩阵中并无真实数据被错误预测，且每个真实值都处于预测值的预测中心，无任何偏差，说明SVM的预测准确率符合100%，分类效果已达到理想

未知中药材模型预测结果如下表：

表 15 未知中药材模型预测结果

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

5.4.2 产地鉴别模型

预测准确率见下表：

表 16 各种分类器对产地的预测准确率

分类器	SG一阶平滑导数	SG二阶平滑导数
DT	61.1%	56.4%
KNN	63.4%	58.0%
NB	54.8%	58.3%
LDA	**75.5%**	**65.0%**
SVM	49.0%	49.0%

由上表可知，在经过一阶平滑处理后的光谱数据经过LDA分类器进行预测的准确率较高，但仍不理想。

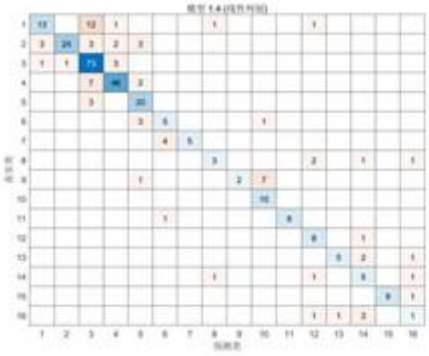

图 17 LDA 分类器模型的混淆矩阵

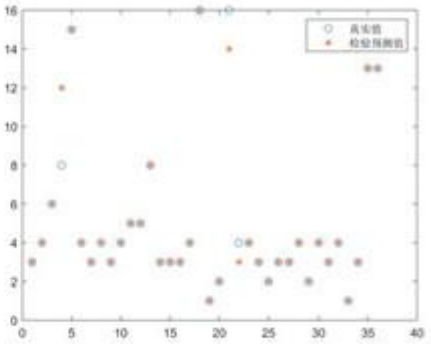

图 20 LDA 分类器测试集仿真效果

由上述结果，各种分类器基于近红外光谱数据建模时，模型的准确率不高。我们通过观察各样品的近红外光谱图，发现在近红外波段中药材的光吸收差异不大。

为探索原因，我们对光谱数据进行样本均衡性分析，统计了不同类别和不同产地的样本数量，结果如下表：

表 17 样本数量统计表

	A	B	C	未知	总计
1	10	5	4	11	30
2	4	8	14	12	38
3	30	5	16	37	**88**
4	24	6	9	26	65
5	12	4	0	9	25
6	0	8	0	2	10
7	0	7	0	2	9
8	0	3	0	6	9
9	0	7	0	3	10
10	0	6	0	4	10
11	0	5	0	4	9
12	0	5	0	4	9
13	0	4	0	6	10
14	0	4	0	4	**8**
15	0	6	0	5	11
16	0	7	0	1	**8**
未知	17	12	14	7	50
总计	97	102	57	143	399

由上表我们可以看出就产地而言样本数量存在较为严重的不均衡性，产地 3 的样本最多有 88 个，而产地 14、16 的样本只有 8 个。因此我们考虑建立基于中药材的类别的场地鉴别模型。

5.4.3 基于中药材的类别的产地鉴别模型

对三种中药材（A、B、C）分别进行产地鉴别模型结果如下表：

表 18 提升后的产地鉴别准确率

	SG一阶平滑			SG二阶平滑		
分类器	A	B	C	A	B	C
DT	62.5%	68.9%	90.7%	53.8%	62.2%	95.3%
LDA	**76.2%**	**88.9%**	**100%**	**72.5%**	**80.0%**	**100%**
NB	68.8%	73.3%	95.3%	61.3%	77.8%	95.3%
SVM	75.0%	74.4%	100%	72.5%	57.8%	100%
KNN	65.0%	80.0%	90.7%	72.5%	80.0%	95.3%

从上表可以发现相对于未进行药材类别分类的的识别模型，准确率有显著的提升。其中，经过SG一阶平滑后的光谱数据再经过LDA进行产地鉴别，鉴别准确率最高，三个分类分别达到了76.2%、88.9%、100%的鉴别准确率。

表 19 各类别产地鉴别模型的的 F_1 分数

产地	1	2	3	4	5	6	7	8
A	0.57	0.75	0.73	0.85	0.83	--	--	--
B	1	0.89	0.57	0.92	0.89	1	1	1
C	1	1	1	1	--	--	--	--
产地	9	10	11	12	13	14	15	16
A	--	--	--	--	--	--	--	--
B	0.83	0.92	1	1	0.66	0.5	1	0.83
C	--	--	--	--	--	--	--	--

注：--表示该类别没有此产地的样品

运用基于中药材类别的产地鉴别模型对未知样品进行鉴别结果如下：

表 20 问题四药材的鉴别与分类表

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

六、模型检验

由于在模型建立中，我们已经通过敏感度（*TPR*），特异性（*FPR*），F_1 分数对模型进行了检验，并且所有模型的正确率都达到了理想结果，所以在这里不做过多阐述。

七、模型评价

优点：

1、 在问题三中，我们将近红外光谱数据与中红外光谱数据进行合并后，产地鉴别的模型准确率有所提高。

2、 在问题四中，我们改进了机器学习的产地鉴别模型，建立了基于中药材类别的产地鉴别模型。

3、 结合机器学习、化学计量学等现代分析方法，建立了多种中药材类别与

产地的鉴别和分析模型，提高了中药材类别与产地的鉴别和分析的准确性。

缺点：

1、 在光谱特征提取时，可以使用更复杂的特征抽取算法，如流形学习、压缩感知等技术进行研究。

八、模型推广

本文建立的基于机器学习的中药材类别与产地鉴别模型，对其他植物的类别鉴别与产地鉴别都具有参考作用。

九、参考文献

[1] 但松健.基于 NIR 光谱分析的柑橘产地鉴别及品质检测技术研究[D].重庆大学.2017.

[2] 田园盛.黄土高原水蚀风蚀交错区生物土壤结皮高光谱特征研究[D].西北农林科技大学,2019.

[3] 鄢悦,张红光,卢建刚,施英姿,陈金水.基于光谱信息散度的近红外光谱局部偏最小二乘建模方法[J].计算机与应用化学.

[4] 黄艳华， 杜娟， 夏田， 刘斯佳， 李洪超， 张蕴薇． 近红外光谱在植物种及品种鉴定中的应用[J]. 中国农学通报. 2014, (6): 46-51.

十、附录

表 21 药材分类明细表

SG 处理后进行一阶导数据药材分类					
类别一	类别二	类别三	类别四	类别五	类别六
61	1	8	64	3	6
67	2	10	136	5	12
69	4	13	201	11	14
78	7	15		16	17
94	9	20		18	30
120	19	24		22	31
130	21	26		25	48
183	23	27		33	53
184	32	28		34	59
200	35	29		38	77
220	39	36		47	79
258	40	37		62	86
279	41	42		63	93
297	45	43		65	95
308	46	44		66	100
318	51	49		68	103

345	55	50		74	106
384	56	52		81	109
390	57	54		82	113
397	60	58		85	124
400	71	70		87	128
412	73	72		90	131
	75	84		99	133
	76	88		104	138
	80	96		116	145
	83	97		118	154
	89	98		126	167
	91	101		127	168
	92	105		129	170
	102	107		134	173
	111	108		137	174
	114	110		140	175
	121	112		143	179
	122	115		144	180
	132	117		152	187
	135	119		155	190
	139	125		164	192
	142	141		165	195
	146	147		171	208
	148	149		176	210
	151	150		181	216
	156	153		188	217
	157	158		196	236
	159	163		197	239
	160	172		198	242
	161	178		199	247
	162	182		203	251
	166	185		206	254
	169	186		213	256
	177	207		218	261
	189	211		221	262
	191	212		222	274
	193	214		225	276
	194	215		232	278
	202	219		237	287
	204	226		238	290
	205	229		240	292
	209	241		249	295
	223	243		253	304

	224	245		257	306
	227	248		259	311
	228	250		260	312
	230	252		270	314
	231	263		271	315
	233	264		272	316
	234	269		275	321
	235	273		281	327
	244	277		282	329
	246	283		288	332
	255	284		291	339
	265	301		293	340
	266	302		296	343
	267	317		298	344
	268	319		300	348
	280	322		303	352
	285	328		305	353
	286	349		309	355
	289	350		313	358
	294	356		323	360
	299	359		324	365
	307	363		325	374
	310	368		331	393
	320	370		334	395
	326	378		336	399
	330	379		347	405
	333	381		354	409
	335	383		361	414
	337	387		362	420
	338	389		366	424
	341	394		371	
	342	396		373	
	346	401		375	
	351	402		376	
	357	404		377	
	364	406		380	
	367	411		382	
	369	415		385	
	372	422		386	
	392	423		388	
	398			391	
	403			407	
	416			408	

	417			410	
	421			413	
	425			418	
	123			419	
类别一	类别二	类别三	类别四	类别五	类别六
308	193	356	136	272	95

表 22 问题二的 SID 表

地区	1	2	3	4	5
1	0.0000	0.0083	0.0043	0.0063	0.0075
2	0.0083	0.0000	0.0038	0.0029	0.0070
3	0.0043	0.0038	0.0000	0.0024	0.0086
4	0.0063	0.0029	0.0024	0.0000	0.0054
5	0.0075	0.0070	0.0086	0.0054	0.0000
6	0.0041	0.0034	0.0004	0.0020	0.0072
7	0.0122	0.0016	0.0049	0.0028	0.0080
8	0.0115	0.0067	0.0077	0.0025	0.0064
9	0.0095	0.0034	0.0030	0.0012	0.0066
10	0.0148	0.0100	0.0075	0.0036	0.0100
11	0.0152	0.0100	0.0046	0.0056	0.0142
地区	6	7	8	9	10
1	0.0041	0.0122	0.0115	0.0095	0.0148
2	0.0034	0.0016	0.0067	0.0034	0.0100
3	0.0004	0.0049	0.0077	0.0030	0.0075
4	0.0020	0.0028	0.0025	0.0012	0.0036
5	0.0072	0.0080	0.0064	0.0066	0.0100
6	0.0000	0.0043	0.0066	0.0023	0.0073
7	0.0043	0.0000	0.0059	0.0023	0.0083
8	0.0066	0.0059	0.0000	0.0033	0.0027
9	0.0023	0.0023	0.0033	0.0000	0.0033
10	0.0073	0.0083	0.0027	0.0033	0.0000
11	0.0045	0.0067	0.0084	0.0033	0.0051

问题 1

```
clc;clear
a = xlsread('附件 1.xlsx');
ques_1_1 = a(:,2:end);
%% 原数据
data = ques_1_1;
x_wn = data(1,:);
y_abb = data(2:end,:);
figure()
plot(x_wn,y_abb);%不同种类编号不同波数的吸光率
```

```
title('原始数据');xlabel('波数');ylabel('吸光度')
%%  SNV 标准正交变换
figure();
[xsnv] = snv(data);
plot(x_wn,xsnv)
title('SNV 光谱数据');xlabel('波数');ylabel('吸光度')
%%  SG 平滑算法
rd = 5;
fl = 11;
figure();
kmtlb = sgolayfilt(y_abb',rd,fl,kaiser(fl,38));
plot(x_wn,kmtlb)
title('SG 光谱数据');xlabel('波数');ylabel('吸光度')
%% 在 SG 平滑的基础上进行求导
figure();
k_deriv1 = diff(kmtlb,1);  %SG 数据的一阶求导
plot(x_wn(1:end-1),k_deriv1)
title('SG 求一阶导 光谱数据');xlabel('波数');ylabel('吸光度')
figure();
k_deriv2 = diff(kmtlb,2);  %SG 数据的二阶求导
plot(x_wn(1:end-2),k_deriv2)
title('SG 求二阶导 光谱数据');xlabel('波数');ylabel('吸光度')
%% 将平滑后的数据进行 PCA 分析降维度
x1 = y_abb;
[cum_contribution_rate1,F1,contribution_rate1]= PCA1(x1);
xsnv1 = xsnv(2:end,:);
x2 = xsnv1;
[cum_contribution_rate2,F2,contribution_rate2]= PCA1(x2);
x3 = kmtlb';
[cum_contribution_rate3,F3,contribution_rate3]= PCA1(x3);
x4 = k_deriv1';
[cum_contribution_rate4,F4,contribution_rate4]= PCA1(x4);
x5 = k_deriv2';
[cum_contribution_rate5,F5,contribution_rate5]= PCA1(x5);
%% 将 PCA 提取出贡献率与累计贡献率进行可视化
xx = 20;  % 看前 20 个特征的贡献率与累计贡献率
y1 = contribution_rate1(1:xx)*100;
figure;bar(y1)
hold on
yy1 = cum_contribution_rate1(1:xx)*100;
plot(1:xx,yy1,'r*-')  %未经处理数据
title('贡献率图  原始数据');
xlabel('主成分样本数');ylabel('贡献率')
legend('贡献率','累积贡献率')
```

```
54.
55. y2 = contribution_rate2(1:xx)*100;
56. figure;bar(y2)
57. hold on
58. yy2 = cum_contribution_rate2(1:xx)*100;
59. plot(1:xx,yy2,'r*-')  %SNV 处理后的数据
60. title('贡献率图  SNV 处理后的数据');
61. xlabel('主成分样本数');ylabel('贡献率')
62. legend('贡献率','累积贡献率')
63.
64.
65. y3 = contribution_rate3(1:xx)*100;
66. figure;bar(y3)
67. hold on
68. yy3 = cum_contribution_rate3(1:xx)*100;
69. plot(1:xx,yy3,'r*-')  %SG 处理后的数据
70. title('贡献率图  SG 处理后的数据');
71. xlabel('主成分样本数');ylabel('贡献率')
72. legend('贡献率','累积贡献率')
73.
74. y4 = contribution_rate4(1:xx)*100;
75. figure;bar(y4)
76. hold on
77. yy4 = cum_contribution_rate4(1:xx)*100;
78. plot(1:xx,yy4,'r*-')  %SG 处理后进行一阶导的数据
79. title('贡献率图  SG 处理后进行一阶导的数据');
80. xlabel('主成分样本数');ylabel('贡献率')
81. legend('贡献率','累积贡献率')
82.
83. y5 = contribution_rate5(1:xx)*100;
84. figure;bar(y5)
85. hold on
86. yy5 = cum_contribution_rate5(1:xx)*100;
87. plot(1:xx,yy5,'r*-')  %SG 处理后进行二阶导的数据
88. title('贡献率图  SG 处理后进行二阶导的数据');
89. xlabel('主成分样本数');ylabel('贡献率')
90. legend('贡献率','累积贡献率')
```

```
1.  function [cum_contribution_rate,F,contribution_rate]= PCA1(x)
2.  [n,p] = size(x);
3.  X=zscore(x);
4.  R = cov(X);
5.  R = corrcoef(x);
```

```
[V,D] = eig(R);
lambda1 = diag(D);
lambda1 = lambda1(end:-1:1);
contribution_rate = lambda1 / sum(lambda1);                 % 计算贡献率
cum_contribution_rate = cumsum(lambda1)/ sum(lambda1);      % 计算累计贡献率
V=rot90(V)';
%% 计算我们所需要的主成分的值
i = 1;
s = 0;
if cum_contribution_rate(i)>0.8
    m = 3;
else
    while s <0.8
        s = cum_contribution_rate(i);
        i = i + 1;
        m = i;
    end

end
F = zeros(n,m);
for i = 1:m
    ai = V(:,i)';   % 将第 i 个特征向量取出，并转置为行向量
    Ai = repmat(ai,n,1);
    F(:, i) = sum(Ai .* X, 2);
end
end
```

```
function [xsnv]=snv(data)
x=data;
[m,n]=size(x);
xsnv=(x-mean(x')'*ones(1,n))./(std(x')'*ones(1,n));
```

```
clc;clear
load('pca_p.mat', 'F1')
ppoint = F1;
k = 6;
 count = 100;                                % 定义光谱最大循环次数
[N,~] = size(ppoint);
center = ppoint(1:k,:);                      % 令前 k 个点为光谱初始中心
 distance_square = zeros(N,k);
while count~=0
```

```
    for i = 1:k
        distance_square(:,i) = sum((ppoint - repmat(center(i,:),N,1)).^2,2);

        str1 = ['Center',num2str(i),'=[];'];
        eval(str1);
    end

    for i = 1:N
        minposition = find(distance_square(i,:)==min(distance_square(i,:)));

        str1 = ['Center',num2str(minposition)];
        eval([str1,'=[',str1,';ppoint(i,:)];']);
    end

    for i = 1:k
        str1 = ['Center',num2str(i)];
        eval(['center_New(',num2str(i),',:) = mean(',str1,',1);']);
    end

    if sym(sum((center_New - center).^2)) == 0
        break
    else
        center = center_New;
    end

    count = count-1;
end

for i = 1:k
    I = num2str(i);
    disp(['第',I,'组聚类的点集：']);
    disp(eval(['Center',I]))
end                                                    % 把光谱聚类点显示出来

hold on
for i = 1:k
    str1 = ['Center',num2str(i)];
    plot(eval([str1,'(:,1)']),eval([str1,'(:,2)']),'.','Markersize',15,'colo
r',[rand rand rand]);
    eval(['kn = boundary(',str1,'(:,1),',str1,'(:,2),0.1);']);
    if isempty(kn)
        eval(['plot(',str1,'(:,1),',str1,'(:,2));']);
    else
        eval(['plot(',str1,'(kn,1),',str1,'(kn,2));']);
```

```
    end
    plot(center(:,1),center(:,2),'k+');
    title('原始数据药材分类');


end
```

```
clc;clear
load('pca_p.mat', 'F2')
ppoint = F2;
k = 6;

count = 100;                                % 定义光谱最大循环次数
[N,~] = size(ppoint);
center = ppoint(1:k,:);                     % 令光谱前 k 个点为初始中心
 distance_square = zeros(N,k);
while count~=0
    for i = 1:k
        distance_square(:,i) = sum((ppoint - repmat(center(i,:),N,1)).^2,2);

        str1 = ['Center',num2str(i),'=[];'];
        eval(str1);
    end

    for i = 1:N
        minposition = find(distance_square(i,:)==min(distance_square(i,:)));

        str1 = ['Center',num2str(minposition)];
        eval([str1,'=[',str1,';ppoint(i,:)];']);
    end

    for i = 1:k
        str1 = ['Center',num2str(i)];
        eval(['center_New(',num2str(i),',:) = mean(',str1,',1);']);
    end

    if sym(sum((center_New - center).^2)) == 0
        break
    else
        center = center_New;
    end

    count = count-1;
```

```
end
for i = 1:k
    I = num2str(i);
    disp(['第',I,'组聚类的点集为: ']);
    disp(eval(['Center',I]))
end                                          % 把光谱聚类点显示出来

hold on
for i = 1:k
    str1 = ['Center',num2str(i)];
    plot(eval([str1,'(:,1)']),eval([str1,'(:,2)']),'.','Markersize',15,'colo
r',[rand rand rand]);
    eval(['kn = boundary(',str1,'(:,1),',str1,'(:,2),0.1);']);
    if isempty(kn)
        eval(['plot(',str1,'(:,1),',str1,'(:,2));']);
    else
        eval(['plot(',str1,'(kn,1),',str1,'(kn,2));']);
    end

    plot(center(:,1),center(:,2),'k+');
    title('snv 据药材分类');
end
```

```
clc;clear
load('pca_p.mat', 'F3')
ppoint = F3;                                 % 输入点的光谱坐标
k = 6;
 count = 100;                                % 定义最大循环次数
[N,~] = size(ppoint);
center = ppoint(1:k,:);                      % 令前 k 个点为初始的光谱中心
 distance_square = zeros(N,k);
while count~=0
    for i = 1:k
        distance_square(:,i) = sum((ppoint - repmat(center(i,:),N,1)).^2,2);

        str1 = ['Center',num2str(i),'=[];'];
        eval(str1);
    end  % 计算到每个点到各个聚类中心的距离

    for i = 1:N
        minposition = find(distance_square(i,:)==min(distance_square(i,:)));

        str1 = ['Center',num2str(minposition)];
```

```
        eval([str1,'=[',str1,';ppoint(i,:)];']);
    end

    for i = 1:k
        str1 = ['Center',num2str(i)];
        eval(['center_New(',num2str(i),',:) = mean(',str1,',1);']);
    end

    if sym(sum((center_New - center).^2)) == 0
        break
    else
        center = center_New;
    end

    count = count-1;
end
for i = 1:k
    I = num2str(i);
    disp(['第',I,'组聚类的点集为：']);
    disp(eval(['Center',I]))
end                                          % 把光谱聚类点显示出来
hold on
for i = 1:k
    str1 = ['Center',num2str(i)];
    plot(eval([str1,'(:,1)']),eval([str1,'(:,2)']),'.','Markersize',15,'colo
r',[rand rand rand]);
    eval(['kn = boundary(',str1,'(:,1),',str1,'(:,2),0.1);']);
    if isempty(kn)
        eval(['plot(',str1,'(:,1),',str1,'(:,2));']);
    else
        eval(['plot(',str1,'(kn,1),',str1,'(kn,2));']);
    end
    plot(center(:,1),center(:,2),'k+');
    title('SG 数据药材分类');
end
```

```
clc;clear
load('pca_p.mat', 'F4')
ppoint = F4;                                 % 输入光谱的坐标
k = 6;
 count = 100;
[N,~] = size(ppoint);
center = ppoint(1:k,:);                      % 令光谱前 k 个点为初始中心
```

```
8.   distance_square = zeros(N,k);
9.  while count~=0
10.     for i = 1:k
11.         distance_square(:,i) = sum((ppoint - repmat(center(i,:),N,1)).^2,2);

12.         str1 = ['Center',num2str(i),'=[];'];
13.         eval(str1);
14.     end  % 计算到每个点到各个聚类中心的距离
15.
16.     for i = 1:N
17.         minposition = find(distance_square(i,:)==min(distance_square(i,:)));

18.         str1 = ['Center',num2str(minposition)];
19.         eval([str1,'=[',str1,';ppoint(i,:)];']);
20.     end
21.
22.     for i = 1:k
23.         str1 = ['Center',num2str(i)];
24.         eval(['center_New(',num2str(i),',:) = mean(',str1,',1);']);
25.     end
26.
27.     if sym(sum((center_New - center).^2)) == 0
28.         break
29.     else
30.         center = center_New;
31.     end
32.     count = count-1;
33. end
34. for i = 1:k
35.     I = num2str(i);
36.     disp(['第',I,'组聚类的点集为：']);
37.     disp(eval(['Center',I]))
38. end                                  %  把光谱聚类点显示出来
39.
40. hold on
41. for i = 1:k
42.     str1 = ['Center',num2str(i)];
43.     plot(eval([str1,'(:,1)']),eval([str1,'(:,2)']),'.','Markersize',15,'colo
    r',[rand rand rand]);
44.     eval(['kn = boundary(',str1,'(:,1),',str1,'(:,2),0.1);']);
45.     if isempty(kn)
46.         eval(['plot(',str1,'(:,1),',str1,'(:,2));']);
47.     else
48.         eval(['plot(',str1,'(kn,1),',str1,'(kn,2));']);
```

```
    end
    plot(center(:,1),center(:,2),'k+');

     title('SG处理后进行一阶导数据药材分类');
end
a1 = zeros(size(Center1,1),1);
for ii = 1:size(Center1,1)
    cc1 = find(F4(:,1)==Center1(ii,1));
    a1(ii) = cc1;
end
a2 = zeros(size(Center2,1),1);
for ii1 = 1:size(Center2,1)
    cc2 = find(F4(:,2)==Center2(ii1,2));
    if length(cc2)==1
    a2(ii1) = cc2;
    else
    a2(ii1) = cc2(1);
    aaa1 =cc2(2);
    end
end
a2=unique(a2);a2=[a2;aaa1];
a3 = zeros(size(Center3,1),1);
for ii2 = 1:size(Center3,1)
    cc3 = find(F4(:,1)==Center3(ii2,1));
    a3(ii2) = cc3;
end
a4 = zeros(size(Center4,1),1);
for ii3 = 1:size(Center4,1)
    cc4 = find(F4(:,1)==Center4(ii3,1));
    a4(ii3) = cc4;
end
a5 = zeros(size(Center5,1),1);
for ii4 = 1:size(Center5,1)
    cc5 = find(F4(:,1)==Center5(ii4,1));
    a5(ii4) = cc5;
end
a6 = zeros(size(F4,1),1);
aa = [a1;a2;a3;a4;a5];
aa = sort(aa,1,'descend');
for j = 1:length(aa)
    a6(aa(j)) = 1;
end

aa1 = find(a6 == 0);
```

```
93. a6 = aa1;
```

```
1.  clc;clear
2.  load('pca_p.mat', 'F5')
3.  ppoint = F5;
4.  k = 6;                                  % 输入光谱聚类组数
5.
6.  count = 100;
7.  [N,~] = size(ppoint);
8.  center = ppoint(1:k,:);                 % 令光谱前 k 个点为初始中心
9.
10. distance_square = zeros(N,k);
11. while count~=0
12.     for i = 1:k
13.         distance_square(:,i) = sum((ppoint - repmat(center(i,:),N,1)).^2,2);
14.         str1 = ['Center',num2str(i),'=[];'];
15.         eval(str1);
16.     end
17.
18.     for i = 1:N
19.         minposition = find(distance_square(i,:)==min(distance_square(i,:)));
20.         str1 = ['Center',num2str(minposition)];
21.         eval([str1,'=[',str1,';ppoint(i,:)];']);
22.     end
23.
24.     for i = 1:k
25.         str1 = ['Center',num2str(i)];
26.         eval(['center_New(',num2str(i),',:) = mean(',str1,',1);']);
27.     end
28.
29.     if sym(sum((center_New - center).^2)) == 0
30.         break
31.     else
32.         center = center_New;
33.     end
34.     count = count-1;
35. end
36. for i = 1:k
37.     I = num2str(i);
38.     disp(['第',I,'组聚类点集为：']);
39.     disp(eval(['Center',I]))
```

```
end  % 把聚类点显示出来

hold on
for i = 1:k
    str1 = ['Center',num2str(i)];
    plot(eval([str1,'(:,1)']),eval([str1,'(:,2)']),'.','Markersize',15,'colo
r',[rand rand rand]);
    eval(['kn = boundary(',str1,'(:,1),',str1,'(:,2),0.1);']);
    if isempty(kn)
        eval(['plot(',str1,'(:,1),',str1,'(:,2));']);
    else
        eval(['plot(',str1,'(kn,1),',str1,'(kn,2));']);
    end
    plot(center(:,1),center(:,2),'k+');
       title('SG 处理后进行二阶导数据药材分类');
end
a1 = zeros(size(Center1,1),1);
for ii = 1:size(Center1,1)
    cc1 = find(F5(:,1)==Center1(ii,1));
    a1(ii) = cc1;
end
a2 = zeros(size(Center2,1),1);
for ii1 = 1:size(Center2,1)
    cc2 = find(F5(:,2)==Center2(ii1,2));
    a2(ii1) = cc2;
end
a3 = zeros(size(Center3,1),1);
for ii2 = 1:size(Center3,1)
    cc3 = find(F5(:,1)==Center3(ii2,1));
    a3(ii2) = cc3;
end
a4 = zeros(size(Center4,1),1);
for ii3 = 1:size(Center4,1)
    cc4 = find(F5(:,2)==Center4(ii3,2));
    if length(cc4)==1
    a4(ii3) = cc4;
    else
    a4(ii3) = cc4(1);
    aaa1 =cc4(2);
    end
end
a4=unique(a4);a4=[a4;aaa1];
a5 = zeros(size(Center5,1),1);
for ii4 = 1:size(Center5,1)
```

```
    cc5 = find(F5(:,1)==Center5(ii4,1));
    a5(ii4) = cc5;
end
a6 = zeros(size(F5,1),1);
aa = [a1;a2;a3;a4;a5];
aa = sort(aa,1,'descend');
for j = 1:length(aa)
    a6(aa(j)) = 1;
end
aa1 = find(a6 == 0);
a6 = aa1;
```

```
clc;clear
a = xlsread('附件 1.xlsx');
ques_1_1 = a(:,2:end);
data2 = [308 193 356 136 272 95;130 268 356 191 272 95];
a1 = data2(1,:);
data1 = ques_1_1(2:end,:);
b1 = data1(a1,:);
x_wn = ques_1_1(1,:);
plot(x_wn,b1);   %SG 处理后进行一阶导数据药材分类后提取特征数据
title('一阶导聚类特征类别')
xlabel('波数');ylabel('吸光度')
legend('类别一','类别二','类别三','类别四','类别五','类别六')


ll = zeros(length(data2(1,:)));
for i = 1:length(data2(1,:))
    for j = 1:length(data2(1,:))
        g1 = b1(i,:);g2 = b1(j,:);
        p1 = g1/sum(g1);q1 = g2/sum(g2);
        D12 = sum(p1.*(log10(p1/q1)));
        D21 = sum(q1.*(log10(q1/p1)));
        sid = D12 + D21;
        ll(i,j) = sid;
    end
end
```

问题二：

```
clc;clear
%% 将问题二数据提取到工作区
a = xlsread('附件 2.xlsx');
```

```
ques_2_1 = a(:,2:end);
ques_2_1(isnan(ques_2_1(:,1)),1)=0;
data = ques_2_1(2:end,2:end);
data1 = ques_2_1(2:end,1);
data2 = ques_2_1(1,2:end);
%% 数据预处理 SG SG-1 SG-2
rd = 5;
fl = 11;
kmtlb = sgolayfilt(data',rd,fl,kaiser(fl,38));  % SG
k_deriv1 = diff(kmtlb,1);                       % SG-1
k_deriv2 = diff(kmtlb,2);                       % SG-2
%% 提取主成分 PCA
x1 = data;
[cum_contribution_rate1,F1,contribution_rate1]= PCA1(x1);
x2 = kmtlb';
[cum_contribution_rate2,F2,contribution_rate2]= PCA1(x2);
x3 = k_deriv1';
[cum_contribution_rate3,F3,contribution_rate3]= PCA1(x3);
x4 = k_deriv2';
[cum_contribution_rate4,F4,contribution_rate4]= PCA1(x4);
%% 画出贡献率图
xx = 20;  % 看前 20 个特征的贡献率与累计贡献率
y1 = contribution_rate1(1:xx)*100;
figure;bar(y1)
hold on
yy1 = cum_contribution_rate1(1:xx)*100;
plot(1:xx,yy1,'r*-')  %未经处理数据
title('贡献率图  原始数据');
xlabel('主成分样本数');ylabel('贡献率')
legend('贡献率','累积贡献率')


y2 = contribution_rate2(1:xx)*100;
figure;bar(y2)
hold on
yy2 = cum_contribution_rate2(1:xx)*100;
plot(1:xx,yy2,'r*-')  %SG 处理后的数据
title('贡献率图  SG 处理后的数据');
xlabel('主成分样本数');ylabel('贡献率')
legend('贡献率','累积贡献率')

y3 = contribution_rate3(1:xx)*100;
figure;bar(y3)
hold on
```

```
yy3 = cum_contribution_rate3(1:xx)*100;
plot(1:xx,yy3,'r*-')  %SG 处理后进行一阶导的数据
title('贡献率图  SG 处理后进行一阶导的数据');
xlabel('主成分样本数');ylabel('贡献率')
legend('贡献率','累积贡献率')

y4 = contribution_rate4(1:xx)*100;
figure;bar(y4)
hold on
yy4 = cum_contribution_rate4(1:xx)*100;
plot(1:xx,yy4,'r*-')  %SG 处理后进行二阶导的数据
title('贡献率图  SG 处理后进行二阶导的数据');
xlabel('主成分样本数');ylabel('贡献率')
legend('贡献率','累积贡献率')
```

```
clc;clear
a = xlsread('附件 2.xlsx');
ques_2_1 = a(:,2:end);
ques_2_1(isnan(ques_2_1(:,1)),1)=0;
data1 = ques_2_1(2:end,1);
st = data1;
st1 = unique(st);st1(1)=[];
st_1 = sum(st ==st1(1));st_ind_1 = find(st ==st1(1));
st_2=sum(st ==st1(2));st_ind_2 = find(st ==st1(2));
st_3=sum(st ==st1(3));st_ind_3 = find(st ==st1(3));
st_4=sum(st ==st1(4));st_ind_4 = find(st ==st1(4));
st_5=sum(st ==st1(5));st_ind_5 = find(st ==st1(5));
st_6=sum(st ==st1(6));st_ind_6 = find(st ==st1(6));
st_7=sum(st ==st1(7));st_ind_7 = find(st ==st1(7));
st_8=sum(st ==st1(8));st_ind_8 = find(st ==st1(8));
st_9=sum(st ==st1(9));st_ind_9 = find(st ==st1(9));
st_10=sum(st ==st1(10));st_ind_10 = find(st ==st1(10));
st_11=sum(st ==st1(11));st_ind_11 = find(st ==st1(11));
st_0=sum(st ==0);st_ind_0 = find(st ==0);

randv1 = randperm(st_1);randv3 = randperm(st_3);
randv4 = randperm(st_4); randv6 = randperm(st_6);
randv10 = randperm(st_10); randv2 = randperm(st_2);
randv7 = randperm(st_7);randv8 = randperm(st_8);
randv11 = randperm(st_11); randv5 = randperm(st_5);
randv9 = randperm(st_9);
```

```
data2 = [st_ind_1(randv1(1)),st_ind_2(randv2(1)),st_ind_3(randv3(1)),st_ind_4(randv4(1)),...
    st_ind_5(randv5(1)),st_ind_6(randv6(1)),st_ind_7(randv7(1)),st_ind_8(randv8(1)),...
    st_ind_9(randv9(1)),st_ind_10(randv10(1)),st_ind_11(randv11(1))];
a1 = data2;
data1 = ques_2_1(2:end,2:end);
b1 = data1(a1,:);
x_wn = ques_2_1(1,2:end);
plot(x_wn,b1);
title('不同地区光谱图')
xlabel('波数');ylabel('吸光度')
legend('地区一','地区二','地区三','地区四','地区五','地区六','地区七','地区八'...
,'地区九','地区十','地区十一','Location','northeastoutside');


ll = zeros(length(data2(1,:)));
for i = 1:length(data2(1,:))
    for j = 1:length(data2(1,:))
        g1 = b1(i,:);g2 = b1(j,:);
        p1 = g1/sum(g1);q1 = g2/sum(g2);
        D12 = sum(p1.*(log10(p1/q1)));
        D21 = sum(q1.*(log10(q1/p1)));
        sid = D12 + D21;
        ll(i,j) = sid;
    end
end
```

```
%%GSdiff1
clc;clear
a = xlsread('附件 2.xlsx');
a = a';x = a(3:end,1);y = a(2:end,2:end);
%提取需要鉴别药材
tiqu_column = [3 14 38 48 58 71 79 86 89 ...
               110 134 152 227 331 618];
op = y(1,:);
op(:,tiqu_column) = [];
B = y(2:end,:);
order = 5;
framelen = 11;
sgf = sgolayfilt(B,order,framelen);     %SG
k_deriv2 = diff(sgf,1);                 %SG 数据的一阶求导
```

```
sgf1 = k_deriv2;
sgf1 = mapstd(sgf1);                    %数据标准化
y_tiqu = sgf1(:,tiqu_column);
sgf1(:,tiqu_column) = [];
sgf1 = [sgf1;op];
% classificationLearner                  %此处调用分类器工具箱
```

```
%%GSdiff2
clc;clear
a = xlsread('附件 2.xlsx');
a = a';x = a(3:end,1);y = a(2:end,2:end);
%提取需要鉴别药材
tiqu_column = [3 14 38 48 58 71 79 86 89 ...
               110 134 152 227 331 618];
op = y(1,:);
op(:,tiqu_column) = [];
B = y(2:end,:);
order = 5;
framelen = 11;
sgf = sgolayfilt(B,order,framelen);     %SG
k_deriv2 = diff(sgf,2);                 %SG 数据的二阶求导
sgf1 = k_deriv2;
sgf1 = mapstd(sgf1);                    %数据标准化
y_tiqu = sgf1(:,tiqu_column);
sgf1(:,tiqu_column) = [];
sgf1 = [sgf1;op];
% classificationLearner                  %此处调用分类器工具箱
% save('data2WEN.mat','trainedModel')
load data2WEN
yuce = trainedModel.predictFcn(y_tiqu);
OP = [6 1 4 7 10 6 9 11 3 4 9 2 5 8 3]  %得到编号所在的 OP
```

```
%%GSsoonth
clc;clear
a = xlsread('附件 2.xlsx');
a = a';x = a(3:end,1);y = a(2:end,2:end);
%提取需要鉴别药材
tiqu_column = [3 14 38 48 58 71 79 86 89 ...
               110 134 152 227 331 618];
op = y(1,:);
op(:,tiqu_column) = [];
```

```
B = y(2:end,:);
order = 5;
framelen = 11;
sgf = sgolayfilt(B,order,framelen);      %SG
sgf1 = mapstd(sgf);                      %数据标准化
y_tiqu = sgf1(:,tiqu_column);
sgf1(:,tiqu_column) = [];
sgf1 = [sgf1;op];
% classificationLearner                  %此处调用分类器工具箱
```

```
%%原始数据
clc;clear
a = xlsread('附件 2.xlsx');
a = a';x = a(3:end,1);y = a(2:end,2:end);
%提取需要鉴别药材
tiqu_column = [3 14 38 48 58 71 79 86 89 ...
               110 134 152 227 331 618];
op = y(1,:);
op(:,tiqu_column) = [];
B = y(2:end,:);
B1 = mapstd(B);                          %数据标准化
y_tiqu = B1(:,tiqu_column);
B1(:,tiqu_column) = [];
B1 = [B1;op];
% classificationLearner                  %此处调用分类器工具箱
```

问题三

```
%%进红外 SGdiff1
clc;clear
a = xlsread('附件 3.xlsx','近红外');
a = a';x = a(3:end,1);y = a(2:end,2:end);
%提取需要鉴别药材
tiqu_column = [4 15 22 30 34 45 74 114 170 209];
op = y(1,:);
op(:,tiqu_column) = [];
B = y(2:end,:);
order = 5;
framelen = 11;
sgf = sgolayfilt(B,order,framelen);      %SG
k_deriv1 = diff(sgf,1);                  %SG 数据的一阶求导
sgf1 = k_deriv1;
sgf1 = mapstd(sgf1);                     %数据标准化
```

```
y_tiqu = sgf1(:,tiqu_column);
sgf1(:,tiqu_column) = [];
sgf1 = [sgf1;op];
% classificationLearner                    %此处调用分类器工具箱
% save('data22.mat','trainedModel')
```

```
%%进红外 SGdiff2
clc;clear
a = xlsread('附件 3.xlsx','近红外');
a = a';x = a(3:end,1);y = a(2:end,2:end);
%提取需要鉴别药材
tiqu_column = [4 15 22 30 34 45 74 114 170 209];
op = y(1,:);
op(:,tiqu_column) = [];
B = y(2:end,:);
order = 5;
framelen = 11;
sgf = sgolayfilt(B,order,framelen);      %SG
k_deriv1 = diff(sgf,1);                  %SG 数据的一阶求导
sgf1 = k_deriv1;
sgf1 = mapstd(sgf1);                     %数据标准化
y_tiqu = sgf1(:,tiqu_column);
sgf1(:,tiqu_column) = [];
sgf1 = [sgf1;op];
% classificationLearner                    %此处调用分类器工具箱
% save('data21.mat','trainedModel')
load data21
yuce = trainedModel.predictFcn(y_tiqu);
yuce1 = [17;11;1;2;16;3;4;10;9;14]       %这是得到问题三的 OP
```

```
%%检验中红外
clc;clear
a = xlsread('附件 3.xlsx','中红外');
a = a';x = a(3:end,1);y = a(2:end,2:end);
%提取需要鉴别药材
tiqu_column = [4 15 22 30 34 45 74 114 170 209];
op = y(1,:);
op(:,tiqu_column) = [];
B = y(2:end,:);
order = 5;
framelen = 11;
```

```
sgf = sgolayfilt(B,order,framelen);      %SG
k_deriv2 = diff(sgf,2);                  %SG 数据的一阶求导
sgf1 = k_deriv2;
sgf1 = mapstd(sgf1);                     %数据标准化
y_tiqu = sgf1(:,tiqu_column);
sgf1(:,tiqu_column) = [];
sgf1 = [sgf1;op];
train = sgf1(:,1:200);                   %选择两百个做训练集
test = sgf1(:,201:end);                  %选择剩下的做测试集
sim = test(end,:);                       %检验集
test(end,:) = [];
% classificationLearner                  %此处调用分类器工具箱
% save('data2.mat','trainedModel1')
load data2
yuce = trainedModel1.predictFcn(test);
yuce1 = [1;8;4;8;7;8;7;11;4;11;2;9;15;8;...%这是用模型得到的预测 OP
    15;6;1;8;14;9;7;9;14;9;6;2;5;17;8;1;...
    15;17;10;10;12;4;7;12;7;7;16;2;17;11;6]
plot(1:45,yuce,'o')
hold on
plot(1:45,sim,'*')                       %画出的检验图显然正确率很高
```

```
%%检验近红外
clc;clear
a = xlsread('附件 3.xlsx','近红外');
a = a';x = a(3:end,1);y = a(2:end,2:end);
%提取需要鉴别药材
tiqu_column = [4 15 22 30 34 45 74 114 170 209];
op = y(1,:);
op(:,tiqu_column) = [];
B = y(2:end,:);
order = 5;
framelen = 11;
sgf = sgolayfilt(B,order,framelen);      %SG
k_deriv2 = diff(sgf,2);                  %SG 数据的一阶求导
sgf1 = k_deriv2;
sgf1 = mapstd(sgf1);                     %数据标准化
y_tiqu = sgf1(:,tiqu_column);
sgf1(:,tiqu_column) = [];
sgf1 = [sgf1;op];
train = sgf1(:,1:200);                   %选择两百个做训练集
test = sgf1(:,201:end);                  %选择剩下的做测试集
sim = test(end,:);                       %检验集
```

```
test(end,:) = [];
% classificationLearner                    %此处调用分类器工具箱
% save('data22.mat','trainedModel')
load data22
yuce = trainedModel.predictFcn(test);
yuce1 = [1,8,4,8,7,8,7,11,4,11,13,9,...%这是用模型得到的预测 OP
    15,8,15,6,1,8,14,9,7,9,14,9,6,2,5,...
    17,8,1,15,17,10,10,12,4,10,12,7,7,16,2,17,11,6]
plot(1:45,yuce,'o')
hold on
plot(1:45,sim,'*')                          %画出检验图显然正确率很高
```

```
clc;clear
a1 = xlsread('附件 3.xlsx','近红外');
a2 = xlsread('附件 3.xlsx','中红外');
a1 = a1(:,3:end);
a = [a2,a1];
a = a';x = a(3:end,1);y = a(2:end,2:end);
%提取需要鉴别药材
tiqu_column = [4 15 22 30 34 45 74 114 170 209];
op = y(1,:);
op(:,tiqu_column) = [];
B = y(2:end,:);
order = 5;
framelen = 11;
sgf = sgolayfilt(B,order,framelen);     %SG
k_deriv1 = diff(sgf,1);                 %SG 数据的一阶求导
sgf1 = k_deriv1;
sgf1 = mapstd(sgf1);                    %数据标准化
y_tiqu = sgf1(:,tiqu_column);
sgf1(:,tiqu_column) = [];
sgf1 = [sgf1;op];
% classificationLearner                    %此处调用分类器工具箱
```

```
clc;clear
a1 = xlsread('附件 3.xlsx','近红外');
a2 = xlsread('附件 3.xlsx','中红外');
a1 = a1(:,3:end);
a = [a2,a1];
a = a';x = a(3:end,1);y = a(2:end,2:end);
%提取需要鉴别药材
```

```
tiqu_column = [4 15 22 30 34 45 74 114 170 209];
op = y(1,:);
op(:,tiqu_column) = [];
B = y(2:end,:);
order = 5;
framelen = 11;
sgf = sgolayfilt(B,order,framelen);      %SG
k_deriv2 = diff(sgf,2);                  %SG 数据的一阶求导
sgf1 = k_deriv2;
sgf1 = mapstd(sgf1);                     %数据标准化
y_tiqu = sgf1(:,tiqu_column);
sgf1(:,tiqu_column) = [];
sgf1 = [sgf1;op];
%classificationLearner                    %此处调用分类器工具箱
```

```
%%中红外 SGdiff1
clc;clear
a = xlsread('附件 3.xlsx','中红外');
a = a';x = a(3:end,1);y = a(2:end,2:end);
%提取需要鉴别药材
tiqu_column = [4 15 22 30 34 45 74 114 170 209];
op = y(1,:);
op(:,tiqu_column) = [];
B = y(2:end,:);
order = 5;
framelen = 11;
sgf = sgolayfilt(B,order,framelen);      %SG
k_deriv1 = diff(sgf,1);                  %SG 数据的一阶求导
sgf1 = k_deriv1;
sgf1 = mapstd(sgf1);                     %数据标准化
y_tiqu = sgf1(:,tiqu_column);
sgf1(:,tiqu_column) = [];
sgf1 = [sgf1;op];
% classificationLearner                    %此处调用分类器工具箱
```

```
%%中红外 SGdiff2
clc;clear
a = xlsread('附件 3.xlsx','中红外');
a = a';x = a(3:end,1);y = a(2:end,2:end);
%提取需要鉴别药材
tiqu_column = [4 15 22 30 34 45 74 114 170 209];
```

```
op = y(1,:);
op(:,tiqu_column) = [];
B = y(2:end,:);
order = 5;
framelen = 11;
sgf = sgolayfilt(B,order,framelen);      %SG
k_deriv2 = diff(sgf,2);                  %SG 数据的二阶求导
sgf1 = k_deriv2;
sgf1 = mapstd(sgf1);                     %数据标准化
y_tiqu = sgf1(:,tiqu_column);
sgf1(:,tiqu_column) = [];
sgf1 = [sgf1;op];
% classificationLearner                  %此处调用分类器工具箱
```

问题四：将附件四 class 数据（A、B、C 分别转换成 1，2，3）

```
%将附件四 class 数据（A、B、C 分别转换成 1，2，3）
clc;clear
a = xlsread('附件 4.xlsx');
x = a(2:end,3);y = a(2:end,4:end);
%提取需要鉴别药材
tiqu_column = find(isnan(x));
tiqu_column1 = find(~isnan(x));
op = x;
op (tiqu_column)=[];
order = 5;
framelen = 11;
sgf = sgolayfilt(y',order,framelen);      %SG
k_deriv2 = diff(sgf,1);                   %SG 数据的一阶求导
sgf1 = k_deriv2;
sgf1 = mapstd(sgf1);                      %数据标准化
y_tiqu = sgf1(:,tiqu_column1);
y_tiqu = y_tiqu';
sgf2 = [op,y_tiqu];
%classificationLearner                    %此处调用分类器工具箱
% save('trainedModel3.mat','trainedModel3')
train = sgf2(1:314,:);      % 训练集
load trainedModel3
test = sgf2(314:end,:);     % 检验集
test1 = test(:,1);test2 = test(:,2:end);
yfit = trainedModel3.predictFcn(test2);
xx =1:length(test1);
plot(xx,test1,'o'); hold on
plot(xx,yfit,'*')
```

```
legend('真实值','检验预测值')%检验集可视化
sgf1 = sgf1';
sim = sgf1([94 109 140 278 308 330 347],:);
yfit1 = trainedModel3.predictFcn(sim);
```

```
%将附件四 class 数据（A、B、C 分别转换成 1，2，3）
clc;clear
a = xlsread('附件 4.xlsx');
data_opandc1 = a(2:end,2:end);
data_opandc1(isnan(data_opandc1(:,1)),:)=[];
data_opandc1(isnan(data_opandc1(:,2)),:)=[];
st_ind_1 = find(data_opandc1(:,1) ==1);
class_1 = (data_opandc1(data_opandc1(:,1)==1,2));
l=data_opandc1(st_ind_1,3:end);
order = 5;
framelen = 11;
sgf = sgolayfilt(l',order,framelen);      %SG
k_deriv1 = diff(sgf,1);                   %SG 数据的一阶求导
class_1_all =[class_1,k_deriv1'];
ll = a([94,109,140]+1,4:end);       %预测
sgf1 = sgolayfilt(ll',order,framelen);      %SG
k_deriv11 = diff(sgf1,1);                   %SG 数据的一阶求导
load trainedModel6
yfit1 = trainedModel6.predictFcn(k_deriv11');

st_ind_2 = find(data_opandc1(:,1) ==2);
class_2 = (data_opandc1(data_opandc1(:,1)==2,2));
l2=data_opandc1(st_ind_2,3:end);
order = 5;
framelen = 11;
sgff = sgolayfilt(l2',order,framelen);      %SG
k_deriv2 = diff(sgff,2);                   %SG 数据的一阶求导
class_2_all =[class_2,k_deriv2'];
ll2 = a([347]+1,4:end);       %预测
sgf2 = sgolayfilt(ll2',order,framelen);      %SG
k_deriv11 = diff(sgf2,1);                   %SG 数据的一阶求导
load trainedModel7
yfit2 = trainedModel7.predictFcn(k_deriv11');

st_ind_3 = find(data_opandc1(:,1) ==3);
class_3 = (data_opandc1(data_opandc1(:,1)==3,2));
l3=data_opandc1(st_ind_3,3:end);
order = 5;
```

```
framelen = 11;
sgff3 = sgolayfilt(l3',order,framelen);      %SG
k_deriv3 = diff(sgff3,1);                    %SG 数据的一阶求导
class_3_all =[class_3,k_deriv3'];
l13 = a([278 308 330]+1,4:end);       %预测
sgf3 = sgolayfilt(l13',order,framelen);      %SG
k_deriv13 = diff(sgf3,1);                    %SG 数据的一阶求导
load trainedModel8
yfit3 = trainedModel8.predictFcn(k_deriv13');
```

```
%将附件四 class 数据（A、B、C 分别转换成 1，2，3）
clc;clear
a = xlsread('附件 4.xlsx');
x = a(2:end,2);y = a(2:end,4:end);
%提取需要鉴别药材
tiqu_column = find(isnan(x));
tiqu_column1 = find(~isnan(x));
op = x;
op (tiqu_column)=[];
order = 5;
framelen = 11;
sgf = sgolayfilt(y',order,framelen);      %SG
k_deriv2 = diff(sgf,1);                   %SG 数据的一阶求导
sgf1 = k_deriv2;
sgf1 = mapstd(sgf1);                      %数据标准化
y_tiqu = sgf1(:,tiqu_column1);
y_tiqu = y_tiqu';
sgf2 = [op,y_tiqu];
%classificationLearner                    %此处调用分类器工具箱
% save('trainedModel.mat','trainedModel')
train = sgf2(1:221,:);       % 训练集
load trainedModel
test = sgf2(222:end,:);      % 检验集
test1 = test(:,1);test2 = test(:,2:end);
yfit = trainedModel.predictFcn(test2);
xx =1:length(test1);
plot(xx,test1,'o'); hold on
plot(xx,yfit,'*')
legend('真实值','检验预测值')%检验集可视化
sgf1 = sgf1';
sim = sgf1([94 109 140 278 308 330 347],:);
yfit1 = trainedModel.predictFcn(sim);
```

```
%将附件四 class 数据（A、B、C 分别转换成 1，2，3）
clc;clear
a = xlsread('附件 4.xlsx');
x = a(2:end,3);y = a(2:end,4:end);
%提取需要鉴别药材
tiqu_column = find(isnan(x));
tiqu_column1 = find(~isnan(x));
op = x;
op (tiqu_column)=[];
order = 5;
framelen = 11;
sgf = sgolayfilt(y',order,framelen);      %SG
k_deriv2 = diff(sgf,2);                   %SG 数据的二阶求导
sgf1 = k_deriv2;
sgf1 = mapstd(sgf1);                      %数据标准化
y_tiqu = sgf1(:,tiqu_column1);
y_tiqu = y_tiqu';
sgf2 = [op,y_tiqu];
%classificationLearner                    %此处调用分类器工具箱
% save('trainedModel2.mat','trainedModel2')
train = sgf2(1:314,:);      % 训练集
load trainedModel2
test = sgf2(314:end,:);     % 检验集
test1 = test(:,1);test2 = test(:,2:end);
yfit = trainedModel2.predictFcn(test2);
xx =1:length(test1);
plot(xx,test1,'o'); hold on
plot(xx,yfit,'*')
legend('真实值','检验预测值')%检验集可视化
sgf1 = sgf1';
sim = sgf1([94 109 140 278 308 330 347],:);
yfit1 = trainedModel2.predictFcn(sim);
```

```
%将附件四 class 数据（A、B、C 分别转换成 1，2，3）
clc;clear
a = xlsread('附件 4.xlsx');
x = a(2:end,2);y = a(2:end,4:end);
%提取需要鉴别药材
tiqu_column = find(isnan(x));
tiqu_column1 = find(~isnan(x));
op = x;
```

```
op (tiqu_column)=[];
order = 5;
framelen = 11;
sgf = sgolayfilt(y',order,framelen);      %SG
k_deriv2 = diff(sgf,2);                   %SG 数据的一阶求导
sgf1 = k_deriv2;
sgf1 = mapstd(sgf1);                      %数据标准化
y_tiqu = sgf1(:,tiqu_column1);
y_tiqu = y_tiqu';
sgf2 = [op,y_tiqu];
%classificationLearner                    %此处调用分类器工具箱
% save('trainedModel1.mat','trainedModel1')
train = sgf2(1:221,:);      % 训练集
load trainedModel1
test = sgf2(222:end,:);     % 检验集
test1 = test(:,1);test2 = test(:,2:end);
yfit = trainedModel1.predictFcn(test2);
xx =1:length(test1);
plot(xx,test1,'o'); hold on
plot(xx,yfit,'*')
legend('真实值','检验预测值')%检验集可视化
sgf1 = sgf1';
sim = sgf1([94 109 140 278 308 330 347],:);
yfit1 = trainedModel1.predictFcn(sim);
```

基于机器学习的中药材类别与产地鉴别

摘要

本文通过建立基于机器学习的五种分类器，分析了近中红外数据及其光谱图，对药材实现种类与产地的鉴别。

针对问题一，主要解决两个问题：一是需要研究不同种类药材的特征和差异性，二是要鉴别药材的种类。鉴于原始光谱数据可能会出现基线偏移和重叠，所以我们使用Savitzky-Golay卷积平滑法与其一阶及二阶平滑导数和标准正态变换（SNV）来对光谱数据进行预处理；运用主成分分析（PCA）对原始光谱数据进行降维，抽取12个主成分，进行K-means聚类，将药材划分为六类；对每一类药材进行药材特征及差异性的分析及SID值的比较，发现第四类药材（只有3个样本）的光谱图明显不同于其他种类药材，而其他种类药材的SID值相互间差异不大，推测除第四类药材的产地外，其他类药材的产地及生长环境相差不大。

针对问题二，我们提出了相对完整的中药产地和类型鉴别框架：包括光谱数据预处理、特征提取、特征选择、分类模型构建及性能测试。采用了SG平滑法结合一阶和二阶导数法进行预处理，用PCA降维提取特征，使用决策树、K-邻近、朴素贝叶斯和线性判别分析的分类器，对中药光谱数据进行类型鉴别，结果表明：SG平滑结合二阶导数能增加大部分分类器的识别效果，在不同的分类器中，LDA的性能最稳定，最佳识别准确率为98.3%，F_1分数为0.98。利用鉴别分类器可提取对未知样品的鉴别结果为：

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

针对问题三，利用第二问的识别框架，我们分别基于近红外光谱数据、中红外光谱数据及两者合并的光谱数据分别建立分类学习模型，结果在产地鉴别效果中，近红外光谱不如中红外光谱不如二者合并的综合光谱。其中的LDA识别效果最好，识别率为98.4%，F_1分数为0.97，对未知产地进行鉴别，结果为：

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

针对问题四，基于二问中的识别框架，利用附件四中的近红外数据对药材类型进行鉴别，识别准确率达到100%，但是对于鉴别产地时，效果不理想，于是我们建立基于药材类型的产地识别模型，在识别了药材类型的基础上，对不同药材的产地建立识别模型，分别提升准确率达到72.6%，88.9%与100%，最终结果为:

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

本文研究发现光谱数据预处理、特征抽取和选择算法可以有效的提高光谱分析的准确率。就本题问题而言SG的一阶导数预处理，PCA提取特征、LDA学习的准确率最稳定

关键词：红外光谱 中药材鉴别 机器学习

一、问题重述

1.1 问题背景

道地药材是指经过中医临床长期应用优选出来的，产在特定地域，与其他地区所产同种中药材相比，品质和疗效更好，且质量稳定，具有较高知名度的中药材。中药材的道地性以产地为主要指标，产地的鉴别对于药材品质鉴别尤为重要。很多中药材形态相似，但其化学成分、性味、毒性、用量、药理作用和功能等方面均不完全相同。我国传统医学对中药材鉴别一直采用性状鉴别，基源鉴别，显微鉴别及理化鉴别等方法。但是对于药材外观损坏，或残缺不全的中药材应用传统鉴别方法有一定的困难。采用红外光谱法，根据其微观特征，可以准确地进行中药材真伪及混乱品种的鉴别。

1.2 问题的提出

附件 1 至附件 4 是一些中药材的近红外或中红外光谱数据，其中 No 列为药材的编号，Class 列为中药材的类别，OP 列为该种药材的产地，其余各列第一行的数据为光谱的波数，第二行以后的数据表示该行编号的药材在对应波段光谱照射下的吸光度，但由于该吸光度为仪器矫正后的值，所以可能存在负值。

建立数学模型，研究解决以下问题：

(1) 根据附件 1 中几种药材的中红外光谱数据，研究不同种类药材的特征和差异性，并鉴别药材的种类。

(2) 根据附件 2 中某一种药材的中红外光谱数据，分析不同产地药材的特征和差异性,鉴别该种药材的产地，并将药材产地的鉴别结果填入编号表格中。

(3) 根据附件 3 中某一种药材的近红外数据和中红外数据，鉴别该种药材的产地，并将所给出编号的药材产地的鉴别结果填入编号表格中。

(4) 根据附件 4 中几种药材的近红外光谱数据，鉴别药材的类别与产地，并将所给出编号的药材类别与药材产地的鉴别结果填入表格中。

二、问题分析

2.1 数据预处理

由于在对中药材进行光谱采集的过程中，会受到颗粒大小、高频随机噪声、样品背景、散射光、仪器响应速度及外界环境等多种因素的干扰，因此获取的光谱数据除含有样本本身的大量信息以外，往往还含有与测试样本无关的成分，从而导致光谱出现了基线偏移和重叠。这些因素会严重影响到所建模型的稳定性和准确度，所以为减弱或者消除这些无关的非目标因素从而获取一个稳定、可靠和准确的校正模型，我们必须要对光谱数据进行去噪，可以使用 Savitzky-Golay 卷积平滑法和标准正态变量变换（SNV）对光谱数据进行预处理。

2.2 问题一的分析

问题一要求根据附件 1 中几种药材的中红外光谱数据，研究不同种类药材的特征和药材之间的差异性，并鉴别药材的种类。在对附件 1 光谱数据预处理后，使用主成分分析（PCA）将高维度数据降维到适当的维度，再基于主成分分析进行特征抽取，选取经过主成分分析后累计贡献率超过 80%的特征值所对应的主成分，使用 K-means 聚类方法对选取的主成分进行聚类，并通过 SID 光谱散度分析探究分类的合理性，最后通过对不同种类药材的光谱数据进行作图，通过峰度、峰形及峰强进行分析几种药材之间不同的特征及差异性。

2.3 问题二的分析

问题二要求根据附件 2 某一种药材的中红外光谱数据分析不同产地药材的特征和差异性，并鉴别药材的产地。在对附件 2 光谱数据预处理后，使用主成分分析（PCA）法抽取特征值，并对数据集进行划分，用特征值进行机器学习训练，然后鉴别出药材的产地，最后对药材进行特征的分析及差异性的比较。

2.4 问题三的分析

问题三要求根据附件 3 中某一种药材的近红外数据和中红外数据对药材产地进行鉴别。分别对附件 3 中近红外与中红外的光谱数据预处理后，使用分类器模型对药材的产地进行鉴别。

2.5 问题四的分析

问题四要求根据附件 4 中几种药材的近红外光谱数据鉴别药材的类别与产地。在对附件 4 中的近红外光谱数据预处理后，我们利用问题二的分类器模型将药材的类别进行分类与产地进行鉴别。

图 1 基于机器学习的光谱产地识别框架

三、模型假设与约定

1. 假设所有的近红外波长都在 0.75 μm 到 3 μm 范围内，中红外波长都在为 3 μm 到 30 μm 范围内。
2. 各个附件的药材数据真实可靠。
3. 不考虑外界因素比如空气湿度，光照强度等对数据进行干扰。

四、符号说明及名词定义

符号	说明
x_i	第 i 样品光谱的平均值
m	波长点数
n	校正样品数
λ_m	特征值
a_m	特征向量
$P(c_k)$	产地在样本中出现的概率
$P(c_k \mid x)$	概率密度函数
Ω	输入空间

$precision_k$	查准率
$recall_k$	查全率
$accuracy$	准确率

五、模型建立与求解

5.1 问题一的模型建立与分析

5.1.1 光谱去噪

对光谱数据进行预处理，采用Savitzky-Golay卷积平滑法和标准正态变量变换（SNV）来对光谱数据进行预处理。

(1) Savitzky-Golay卷积平滑法

Savitzky-Golay卷积是利用局部滑动窗口来实现的一种多项式回归算法，其原理为：将原始的受干扰的光谱数值替换为平均值，该均值实质是通过局部的移动窗口对要替换的数值进行最近似于真实信号的多项式拟合，从而使得原来受到干扰的光谱值可以更正为更为合理的信号值，其近似过程实质上可以看做是一种加权的平均方法。

设滤波窗口的宽度 $n=2m+1$，各测量点为 $x=\{-m,-m+2,..,0,1,\ldots,m-1,m\}$，我们采用二次多项式对窗口内的数据进行拟合，其表达式为：

$$y=a_0+a_1x+a_2x^2 \tag{1}$$

其中 a_0,a_1 和 a_2 为二次多项式的系数，y 为多项式拟合后的值。由公式（1）可以得到 n 个类似的方程，从而构成一个三元线性方程组。

通过公式（1），我们可以得到拟合后的 SG 平滑光谱值。

一阶 SG 平滑导数为：

$$\left.\frac{dy}{dx}\right|_{x=0}=a_1 \tag{2}$$

二阶 SG 平滑导数为：

$$\left.\frac{d^2y}{dx^2}\right|_{x=0}=a_2 \tag{3}$$

(2) 标准正态变量变换(SNV)

标准正态变量变换(SNV)主要是用来消除固体颗粒大小、表面散射以及光程变化对漫反射光谱的影响。对需SNV变换的光谱按下式计算：

$$X_{i,SNV}=\frac{x_{i,k}-x_i}{\frac{m_{k=i}\left(x_{i,k}-x_i\right)^2}{(m-1)}} \tag{4}$$

式（5）中 x_i 是第 i 样品光谱的平均值，$k=1,2,\cdots,m$，m 为波长点数；$i=1,2,\cdots n$，n 为校正样品数；$X_{i,SNV}$ 是变换后的光谱。

本次处理使用四种去噪的算法：*SG* 平滑法，一阶 *SG* 平滑导数和二阶 *SG* 平滑导数，来获得近似去噪后的中药材光谱数据，使鉴别模型更为精确地表达中药

材的光谱特征。

使用光谱去噪模型，对附件1中的中药材光谱图进行处理，结果如图 2图 3 所示。原始的几种中药材光谱吸收率显示在（图 3）中，使用原始SG平滑及在此基础上衍生出的一阶SG平滑导数和二阶SG平滑导数的去噪效果图见图 3 (b)-(d)中。将（图 2）与（图 3）进行对比，我们可以发现经过SG平滑后原始光谱图变得更加光滑。在进行一阶导数运算后，光谱范围从[-0.1, 0.9]压缩[-0.06, 0.12]，同时光谱信号得到更进一步的平滑，消除了基线漂移，使原始光谱中的波峰变为0在进行二阶导数运算后，范围缩小到[-0.08]到[0.06]，二阶导数光谱的平滑效果跟一阶导数接近，但数据得到更进一步压缩，消除了基线漂移和光谱倾斜，增强了光谱特征。而经过SNV预处理后的光谱图，却保留了原光谱数据的值，且吸光度放大后，可以非常直观地看出各数据之间的差异，其得注意的是，通过图 3 可以看出，中药材样本的光谱具有很大的重叠性。

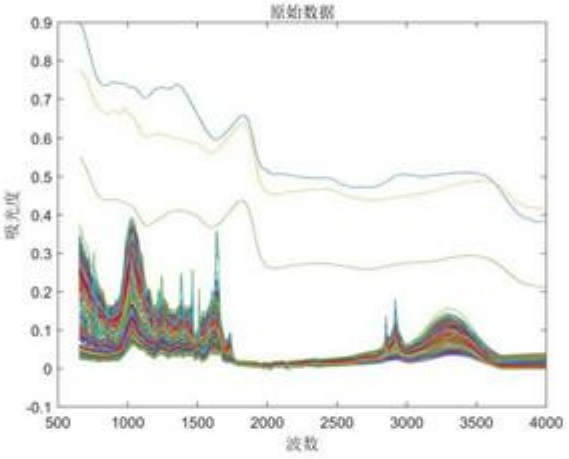

图 2 原始数据光谱图

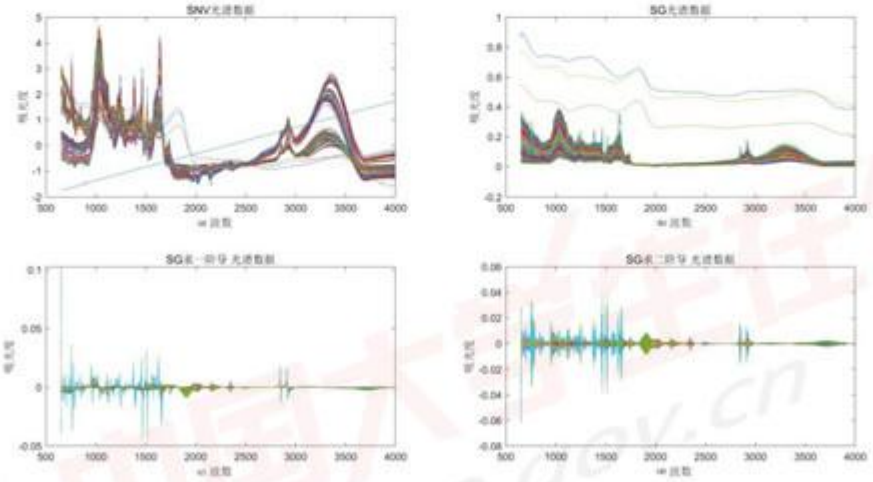

图 3 中药材原始光谱去噪后效果图

其中从左往右依次为：(a) SNV 光谱图；(b) SG 平滑后的效果图；(c) SG 平滑后一阶导数的效果图；(d) SG 平滑后二阶导数的效果图

在对光谱去噪后，由于数据量过多，因此我们需要使用主成分分析选取特征值对其进行降维。

5.1.2 基于主成分分析（PCA）选取特征值

由于中药材的数据量太大，我们需要将上千维中药材的原始光谱数据用少量的数据主成分来代替，所以首先用 PCA 主成分分析方法降维选取中药材的特征值，然后利用特征值使用K-means 将其进行聚类。

主成分分析（PCA）是常见的用于高维数据中的特征抽取方法之一， PCA（主成分分析法）是对高维样本空间引入随机变量，将多个变量通过线性变换选出较少重要变量的一种多元统计方法。

原始样本数可以表达为：$\mathrm{X}=\left[\mathrm{x}^{1},\mathrm{x}^{2},\ldots,\mathrm{x}^{\mathrm{m}}\right]$，m为原始光谱分布图的维数。

为了使光谱数据集能更适合PCA的运算，所以需要先将数据进行归一化处理，然后计算标准化和偶样本的协方差矩阵R，与R的特征值 λ_{m} 和特征向量 a_{m}。

$$\text{贡献率}=\frac{\lambda_{i}}{\sum_{k=1}^{p}\lambda_{k}}(i=1,2,\cdots,p) \tag{5}$$

$$\text{累计贡献率}=\frac{\sum_{k=1}^{i}\lambda_{k}}{\sum_{k=1}^{p}\lambda_{k}}(i=1,2,\cdots,p) \tag{6}$$

取贡献率达到80%的特征值对应的第一至第m(m≤p)个主成分，利用公式

$$F_{\mathrm{i}}=a_{1i}X_{x1}+a_{2i}X_{x2}+\cdots+a_{pi}X_{p}(\mathrm{i}=1,2\cdots,\mathrm{m}) \tag{7}$$

其中 $a_{1i},a_{2i}\ldots a_{pi}$ 为 X 的协方差阵 Σ 的特征值所对应的特征向量，X_{x1}，$X_{x2},\cdots$，a_{pi} 是原始变量标准化后的值。

当特征向量的值越大时，表示其含有更多的原始信息或能量，对该主成分的影响越大，具有更大的重要性。

因为没有足够的证据表明某一段光谱具有很强的区分度，因此对附件1中整个光谱段 $652-3999(cm^{-1})$ 进行主成分提取以得到最具代表性的光谱信息，以主成分的贡献度排序得到的结果如图 4与图 5所示。

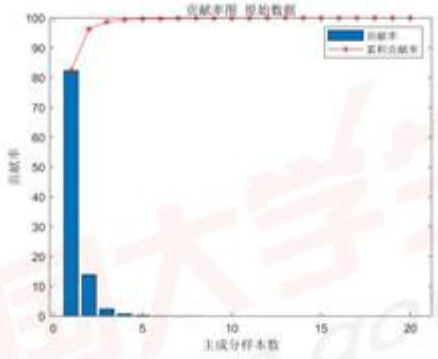

图 4 原始贡献率

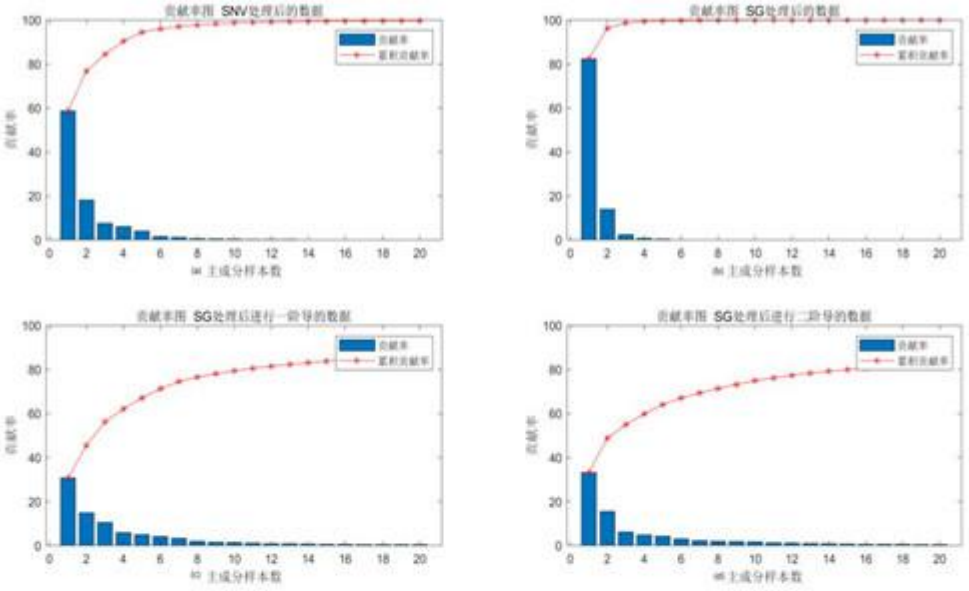

图 5　附件 1 中药材光谱数据进行 PCA 特征抽取之后的主成分贡献率

其中从左往右依次为：（a）SNV 平滑后的数据进行 PCA 特征抽取；（b）SG 平滑后的数据进行 PCA 特征抽取；（c）SG 平滑后一阶导数的数据进行 PCA 特征抽取；（d）SG 平滑后二阶导数的数据进行 PCA 特征抽取

观察图 4 与图 5，我们发现，进行降维后，原始数据中主成分 1 达到了 82.37%，前三个主成分占据了 98.67%的信息量，在经过 SNV 平滑后的主成分贡献率前三个主成分占据了 84.38%的信息量，在对 SG 平滑后的数据提取主成分，前三个主成分占据了 98.67%的信息量，而对一阶导数和二阶导数后的平滑数据，前三个主成分分别占据了 56.19%及 54.93%。由此可知，光谱数据在经过一阶 SG 平滑导数与二阶 SG 平滑导数 PCA 降维后累计贡献率较低，因此需要提取的样本数相较于使用其他方式进行数据处理后的样本数更多。

提取每一个处理方式处理后的光谱数据使用主成分分析后累计贡献率超过 80%的特征值所对应的主成分，保存的主成分个数见下表：

表 1　附件 1 进行 PCA 主成分保留个数

	原始数据	SNV 标准正交变换	SG 平滑算法	SG 一阶平滑	SG 二阶平滑
主成分个数	3	4	3	12	16
累计贡献率	98.67%	90.47%	98.67%	80.56%	80.00%

5.1.3　基于K-means聚类算法进行药材分类

K-means 聚类算法需要先指定需要划分的类的个数 k 值，然后随机地选择 K 个数据对象作为初始的聚类中心，并且计算其余的各个数据对象到这 K 个初始聚类中心的距离，把数据对象划归到距离它最近的那个中心所处在的簇类中：同一聚类中的对象相似度较高；而不同聚类中的对象相似度较低。聚类相似度是利用各聚类中对象的均值所获得一个“中心对象（引力中心）来进行计算的。

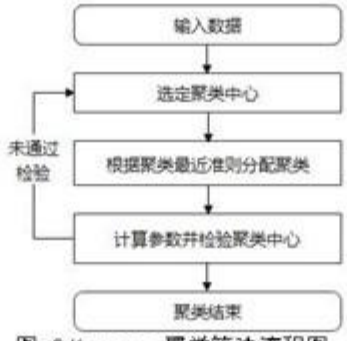

图 6 K-means 聚类算法流程图

由于无法得知最好聚类效果的聚类中心设定值，所以我们将聚类中心分别设置为 4 和 6 进行聚类测试，并将聚类结果可视化，结果见图 7 和图 8。

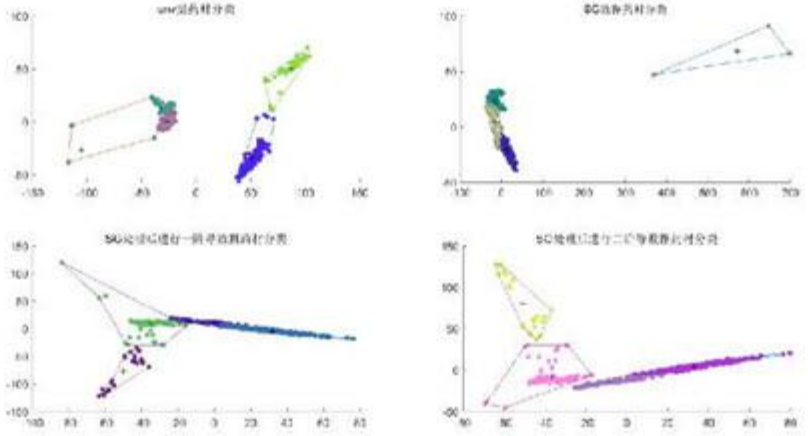

图 7 聚类中心为 4 的聚类效果图

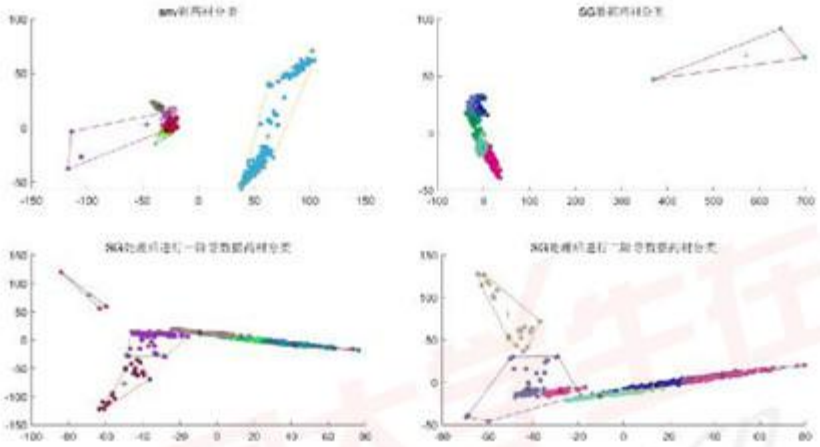

图 8 聚类中心为 6 的聚类效果图

在原始光谱和 SG 平滑后的光谱数据上进行 PCA 降维和 K-means 聚类后，不同种类之间的中药材分布具有一定的区分度。

根据观察，我们可以得知设定聚类中心为 6 的聚类效果比设定聚类中心为 4 的聚类效果更好，所以我们将聚类中心定为 6，将中药材分为六个类别（具体分