Pure Mathematics 理论数学, 2023, 13(2), 364-374
Published Online February 2023 in Hans. https://www.hanspub.org/journal/pm
https://doi.org/10.12677/pm.2023.132040

Hans汉斯

基于SVM算法的中药材产地鉴别模型研究

杨环瑜[1]，陈　莹[2]，丘彤彤[1]，张子荣[1]，罗伟军[1]

[1]湛江幼儿师范专科学校，广东 湛江
[2]湛江机电学校，广东 湛江

收稿日期：2023年1月22日；录用日期：2023年2月21日；发布日期：2023年2月28日

摘　要

中药材在大辞典中已记载有12,000多种，而不同类中药材又分布在众多的产地中，因此鉴别中药材产地的任务极为艰巨。本文以2021年“高教社杯”全国大学生数学建模竞赛E题“附件3”所提供的近红外光谱数据与中红外光谱数据为研究样本，先对所提供数据进行预处理、提取特征向量、降维处理，然后通过支持向量机(SVM)算法进行求解并运用评价指标在训练集上达到了83.8%的正确率，在测试集上达到98.7%的正确率。

关键词

中药材产地，支持向量机，红外光谱，机器学习

Study on Identification Model of Chinese Medicinal Herbs Based on Support Vector Machine Algorithm

Huanyu Yang[1], Ying Chen[2], Tongtong Qiu[1], Zirong Zhang[1], Weijun Luo[1]

[1]Zhanjiang Preschool Normal College, Zhanjiang Guangdong
[2]Zhanjiang Mechanical and Electrical School, Zhanjiang Guangdong

Received: Jan. 22nd, 2023; accepted: Feb. 21st, 2023; published: Feb. 28th, 2023

Abstract

There are more than 12,000 kinds of traditional Chinese medicine recorded in the dictionary, and different kinds of traditional Chinese medicine are distributed in many places of origin, so the task of identifying the place of origin of traditional Chinese medicine is extremely difficult. In this paper, the near-infrared spectral data and mid infrared spectral data provided by the 2021 “Higher Education Society Cup” National Undergraduate Mathematical Modeling Contest E question “Ap-

文章引用: 杨环瑜, 陈莹, 丘彤彤, 张子荣, 罗伟军. 基于 SVM 算法的中药材产地鉴别模型研究[J]. 理论数学, 2023, 13(2): 364-374. DOI: 10.12677/pm.2023.132040

pendix 3" are taken as research samples. First, the data provided are preprocessed, feature vectors are extracted, and dimensions are reduced. Then, support vector machine (SVM) algorithm is used to solve the problem and evaluation indicators are used to achieve 83.8% accuracy in training sets and 98.7% accuracy in test sets.

Keywords

Origin of Chinese Medicinal Herbs, Support Vector Machine, Infrared Spectrum, Machine Learning

1. 引言

中药材的产地对于衡量药材的品质尤为重要，是大众比较关心的问题。由于中药材的种类较多且不同地区名称也可能不同，导致中药材的产地难以鉴别。传统中药材鉴别方法很多，如来源鉴别、性状鉴别、显微鉴别、理化鉴别等，这些鉴别方法效率较低，鉴别准确度不高。DNA 分子和色谱鉴别对中药材的鉴别准确率极高，但存在预处理较为复杂且分析的时间长、成本较高、操作繁琐，快速鉴别较难等不足。现如今的研究有：文献[1]将系统聚类方法运用至光谱分析；文献[2]采用监督分析法构建分类模型；文献[3]利用光谱指纹图谱对僵蚕进行鉴定；文献[4]采用质谱数据对中药材进行鉴定。本文主要研究根据中红外光谱数据鉴别中药材种类的分类模型的基础上进一步考虑提取近红外光谱数据的不同产地中药材在光谱数据中的差异性特征，提高中药材产地鉴别分类的精准度。对此，本文的研究有利于进一步贡献丰富中药材鉴定的研究结果、研究方法、理论基础。

2. 数据来源与分析

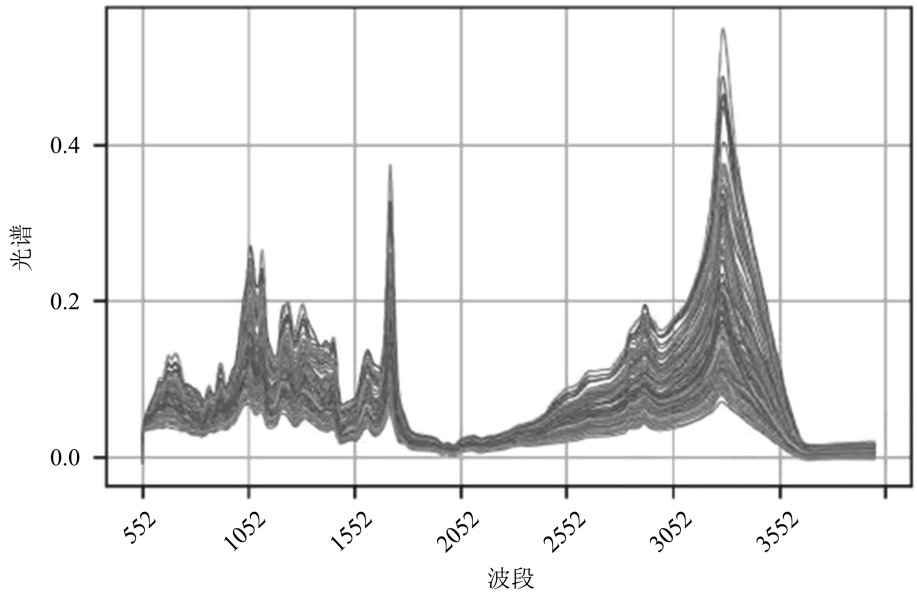

Figure 1. Mid-infrared spectral data plot
图 1. 中红外光谱数据图

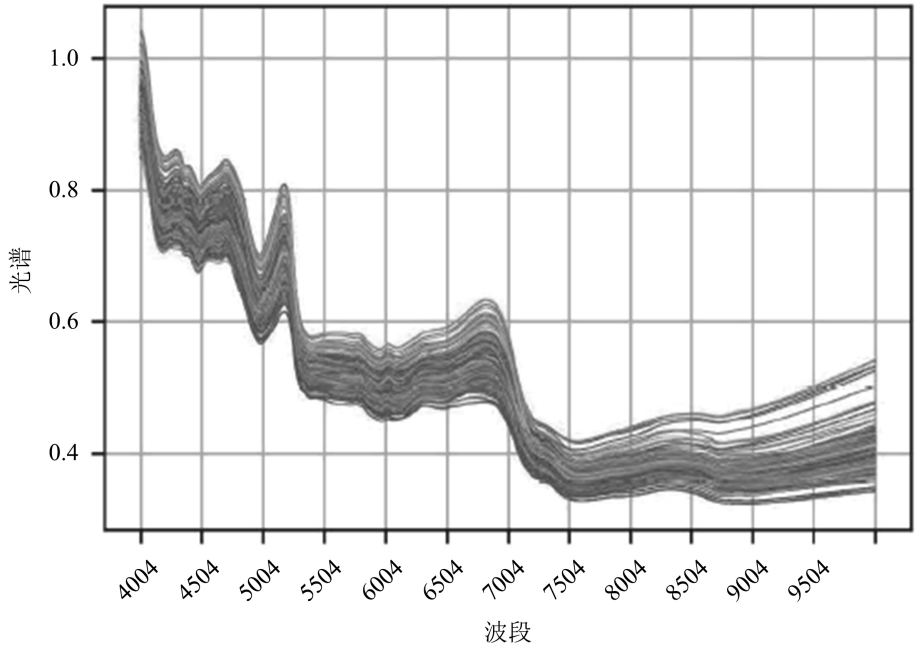

Figure 2. Near-infrared spectral data plot
图 2. 近红外光谱数据图

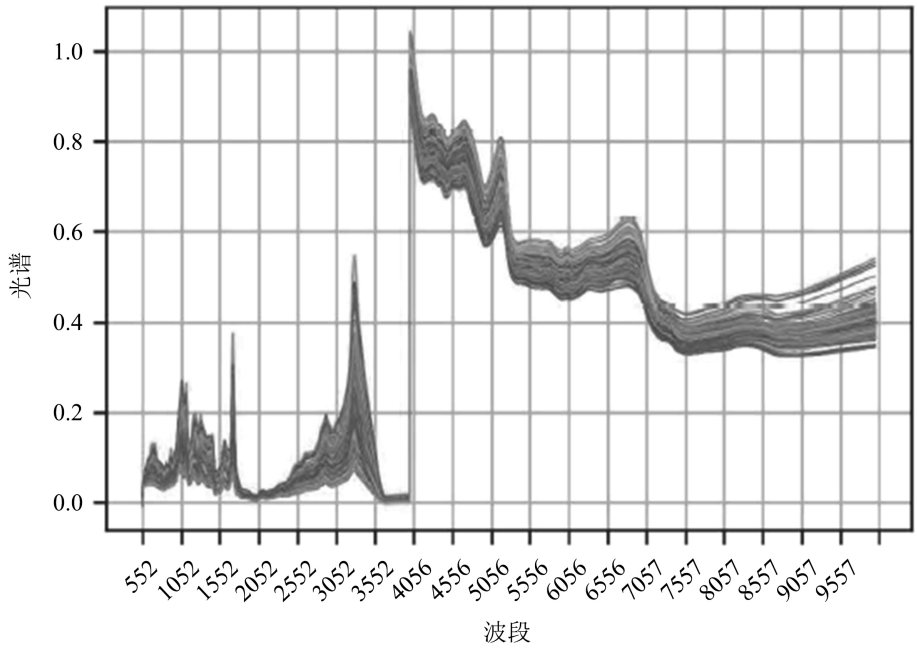

Figure 3. Combined mid-infrared spectra
图 3. 中近红外光谱合并图

本文以 2021 年“高教社杯”全国大学生数学建模竞赛 E 题“附件 3”的基础数据为研究对象，“附件 3”中提供了 255 个中药材基础样本数据，其中 No 列为药材的编号，OP 的名单显示了这种药材的来源。其它列第一行中的数据是光谱的波数(单位 cm^{-1})，第二行后的数据表示暴露于相应波段的编号药材的光谱吸光度(552~3999 cm^{-1} 4004~10,000 cm^{-1}) (见图 1~3)。其中，245 个样品为已知来源(编号为 1

至 17)，10 个样品为未知来源。对附件 3 中的数据进行预处理，检查是否存在缺失值、是否存在异常值、是否存在大量重复值，通过处理明确了每条记录有光谱波数 552 cm^{-1} 至 3999 cm^{-1} 共 878,985 个数据项，光谱波数 4004 cm^{-1} 至 10,000 cm^{-1} 共 1,528,980 个数据项，均未发现大量重复数据和异常数据，光谱波数列中未发现缺失数据。表示产地的 OP 列除去需鉴别场地的 17 个缺失项外也无缺失项，数据完整性良好。

3. 分类模型

3.1. 欧式距离

欧几里得度量也称为欧氏距离[5]，是一种常用的距离定义和最常见的距离测量，它测量多维空间中点之间的绝对距离。它是指 m 维空间中两点之间的实际距离，或向量的自然长度(即点到原点的距离)。二维和三维空间中的欧氏距离是两点之间的实际距离。在计算相似性的场景中(例如人脸识别)，欧氏距离是一种更直观和常见的相似性算法。欧氏距离越小，相似度越大。欧氏距离越大，相似度就越小。本文主要使用其算法来识别中红外光谱中药基础数据的相似度。欧氏距离的数学公式如下：

3.2. 支持向量机模型(SVM)

由于支持向量机[6]能够适应小样本的分类，因此分类速度快，其性能不低于人工神经网络[7]。因此，人们将 SVM 应用到各个领域。大量使用 SVM 模型的论文不断涌现，包括国内外基于固体统计理论的支持向量机。它是所有已知数据挖掘算法中最精确的方法之一，具有良好的学习能力和泛化能力。因此，使用支持向量机求解未知来源药材记录的模型是一种合适的方法。支持向量机的主要思想是找到一个超平面使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。假设给定一个特征空间上的训练数据集

$$T=\left\{\left(x_1,y_1\right),\left(x_2,y_2\right),\left(x_3,y_3\right),\cdots,\left(x_1,y_1\right)\right\}$$

其中，$x_i\in X=R^n$，$y_i\in Y=\left\{+1,-1\right\}$，$i=1,2,3,\cdots,N$，

N 表示为有 N 个样本实例，X_i 则表示为第 i 个特征向量，

该假设的目标是要从中提取出一个分离超平面，并将正负类分别分在该超平面的两侧。

分离超平面的对应方程可写为：

$$\omega\cdot x+b=0$$

当给定的训练数据值处于线性可分的状态时，存在无数个这样的分离超平面，感知机利用误分类的点来求解，则有无数个解。

SVM 可以通过最大化间隔来获得最优超平面和唯一解。

假设分类决策函数为：

$$f\left(x\right)=sign\left(\omega\cdot x+b\right)$$

4. 结果分析

4.1. 欧式距离结果

从图 1~3 可知来自 17 个产地中药材的中近红外光谱数据图整体趋势差别不大，不同产地的药材吸光度差异较小，其数据曲线贴合度较高。本文对附件 3 中近红外光谱数据(见图 4)、中红外光谱数据(见图 5)、中与近红外光谱数据(见图 6)进行一阶平滑处理，处理后图像如下。

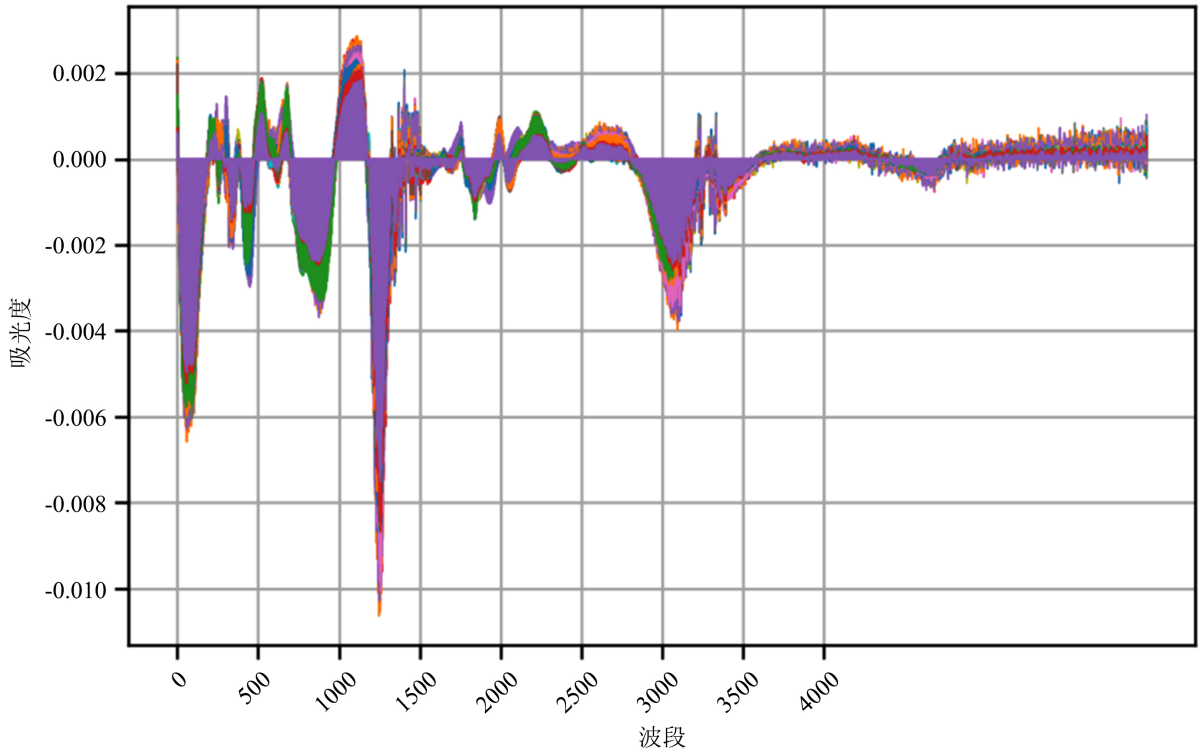

Figure 4. First-order smoothing of near-infrared spectral data
图 4. 近红外光谱数据一阶平滑处理后

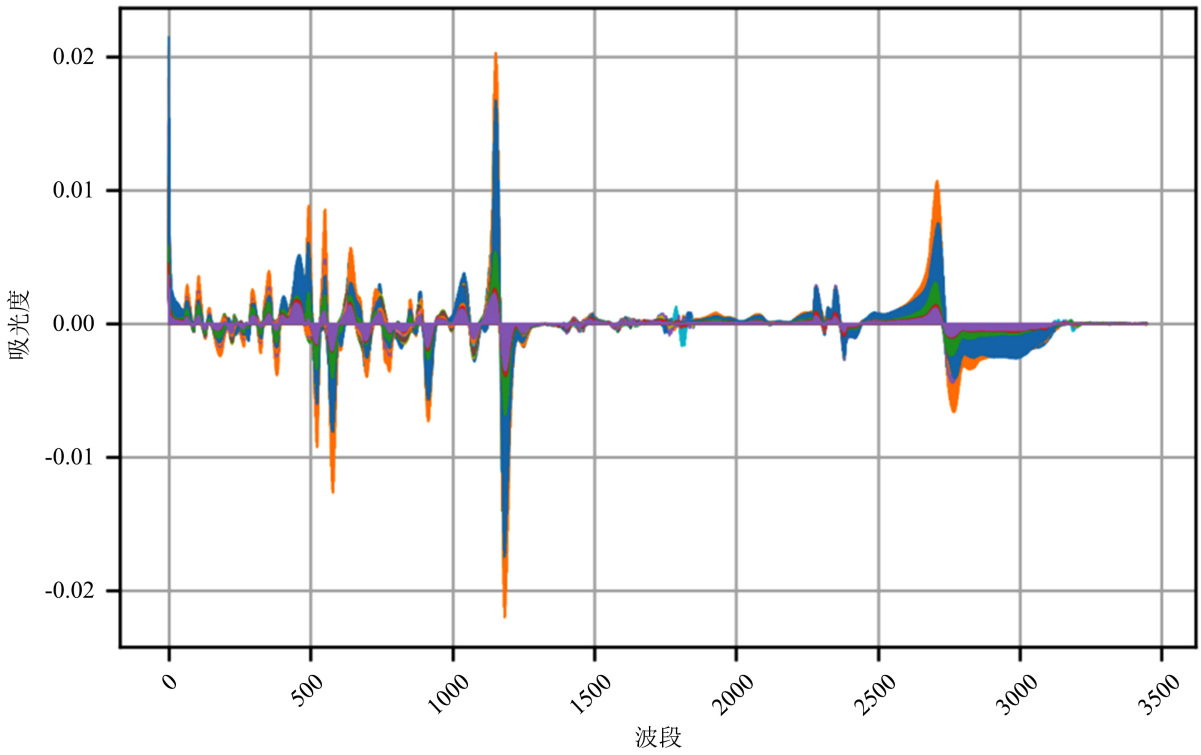

Figure 5. Mid-infrared spectral data after first-order smoothing
图 5. 中红外光谱数据一阶平滑处理后

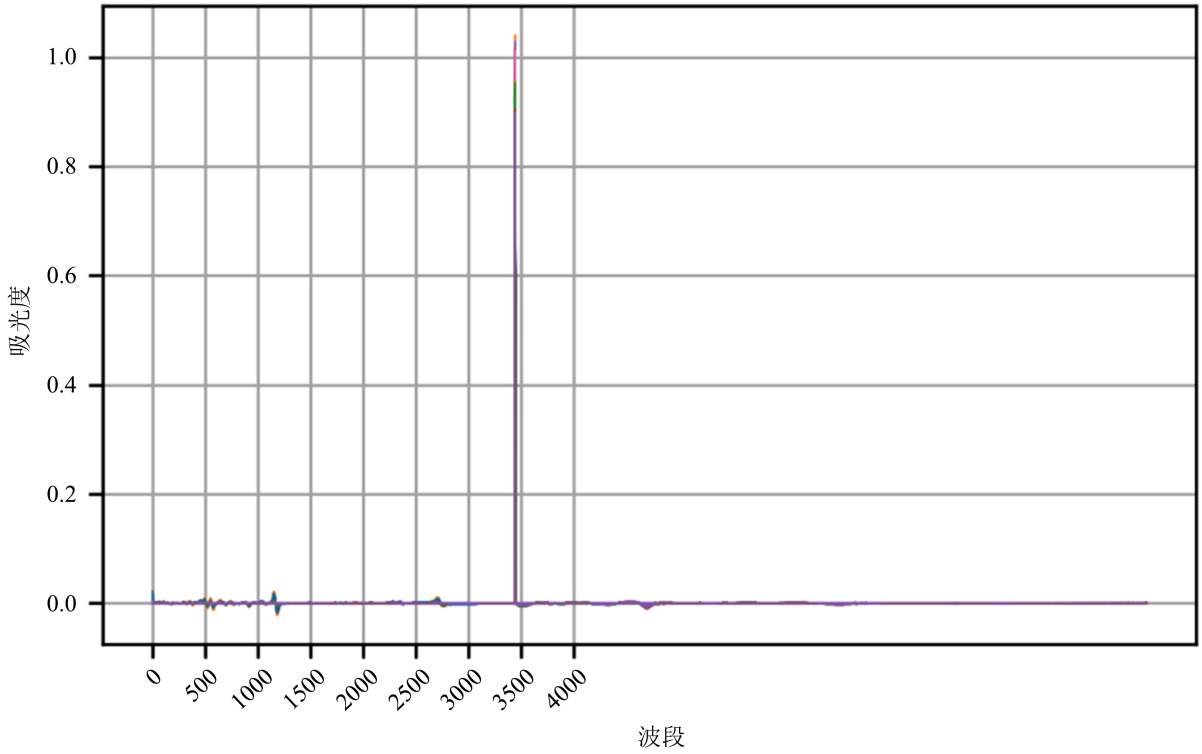

Figure 6. Near-infrared spectral data after first-order smoothing
图 6. 中近红外光谱数据一阶平滑处理后

我们定义，当一个光谱波数区间$[m,n]$上每个光谱波数$V_i(m \le i \le n)$在所有药材记录中的吸光度A_i最大值减去最小值(即极差)大千所有波数吸光度的平均极差。则称该区间为特征区间[7]。记为$T[m,n]$，即：

$$A_{i\max} - A_{i\min} > A_{\text{mean}} \quad m \le i \le n$$

通过 Python 所计算出的结果，显示在附件 3 光谱波数区间共有六个特征区间，总长度为 3005，即包含 3005 个不同且连续的光谱波数(区间与区间之间波数不连续)。在寻找到特征区间之后，以光谱波数作为 x 坐标，以吸光度作为 y 坐标。把我们将要分类的 10 条药材记录与已分类的 245 条药材记录在特征区间上计算曲线间的欧式距离，X_{ij}, $1 \le i \le 10$, $1 \le j \le 245$

$$X_{ij} = \frac{1}{n}\sum_{k=1}^{n}\sqrt{\left(x_{ki} - x_{kj}\right)^2 + \left(y_{ki} - y_{kj}\right)^2} \quad 1 \le i \le 10,\ 1 \le j \le 245,\ n = 3005$$

使用 Python 计算出未分类的 10 条药材记录与已分类的 245 条记录之间的距离矩阵如下(见表 1~3)。

Table 1. Near-infrared spectroscopy
表 1. 近红外光谱矩阵

No	OP	4004	4005	4006	4007	...	9997	9998	9999	10,000
1	14	0.869187	0.869187	0.869769	0.869769	...	0.420397	0.420397	0.420397	0.420764
2	13	0.944521	0.944521	0.944225	0.944225	...	0.382062	0.382062	0.382062	0.382062
3	17	0.959754	0.959754	0.961532	0.961532	...	0.401349	0.401349	0.401349	0.40136

Continued

4	0	0.869418	0.869418	0.87161	0.87161	...	0.405511	0.405511	0.405511	0.40605
5	16	0.913759	0.913759	0.915178	0.915178	...	0.417069	0.417069	0.417069	0.418042
6	17	0.949069	0.949069	0.950782	0.950782	...	0.412635	0.412635	0.412635	0.412939
7	5	1.01663	1.01663	1.0169	1.0169	...	0.458986	0.458986	0.458986	0.459253
...	...	...	...	...	...	...	...	...	...	...
249	7	0.915109	0.915109	0.916006	0.916006	...	0.390883	0.390883	0.390883	0.391525
250	7	0.941788	0.941788	0.942736	0.942736	...	0.391661	0.391661	0.391661	0.392434
251	16	0.952433	0.952433	0.953022	0.953022	...	0.405308	0.405308	0.405308	0.405596
252	2	0.959031	0.959031	0.959407	0.959407	...	0.402501	0.402501	0.402501	0.402977
253	17	0.955752	0.955752	0.957249	0.957249	...	0.397983	0.397983	0.397983	0.398028
254	11	0.909227	0.909227	0.909961	0.909961	...	0.4059	0.4059	0.4059	0.406435
255	6	0.902235	0.902235	0.902844	0.902844	...	0.386275	0.386275	0.386275	0.386464

Table 2. Mid-infrared spectral matrix
表 2. 中红外光谱矩阵

No	OP	552	553	554	555	...	3996	3997	3998	3999
1	14	0.026602	0.031197	0.031197	0.034768	...	0.006093	0.006118	0.006118	0.006148
2	13	0.027282	0.033386	0.033386	0.038157	...	0.006356	0.006385	0.006385	0.006415
3	17	0.019621	0.028953	0.028953	0.036467	...	0.012187	0.012229	0.012229	0.012268
4	0	0.004992	0.019785	0.019785	0.031172	...	0.006427	0.006437	0.006437	0.006427
5	16	0.017975	0.028234	0.028234	0.036466	...	0.015257	0.015259	0.015259	0.015276
6	17	0.020617	0.02996	0.02996	0.037593	...	0.009919	0.009931	0.009931	0.009933
7	5	0.028366	0.034168	0.034168	0.0387	...	−0.000311	−0.00317	−0.000317	−0.000321
...	...	...	...	...	...	...	...	...	...	...
249	7	0.027861	0.032238	0.032238	0.035482	...	0.001628	0.001621	0.001621	0.001604
250	7	0.025559	0.028484	0.028484	0.030757	...	−0.000159	−0.00181	−0.000181	−0.000217
251	16	−0.00863	0.012856	0.012856	0.02829	...	0.01229	0.010264	0.010264	0.010315
252	2	0.026924	0.030733	0.030733	0.03356	...	0.003821	0.0038	0.0038	0.003755
253	17	0.027081	0.032876	0.032876	0.037407	...	0.003574	0.003578	0.003578	0.003574
254	11	0.027094	0.031609	0.031609	0.034978	...	0.00566	0.005663	0.005663	0.005671
255	6	0.026076	0.029686	0.029686	0.032365	...	0.002604	0.002613	0.002613	0.002619

Table 3. Near-mid-infrared spectral matrix
表 3. 近中红外光谱矩阵

No	OP	552	553	555	556	...	9997	9998	9999	10,000
1	14	0.026602	0.031197	0.034768	0.034768	...	0.420397	0.420397	0.420397	0.420764
2	13	0.027282	0.033386	0.038157	0.038157	...	0.382062	0.382062	0.382062	0.382062
3	17	0.019621	0.028953	0.036467	0.036467	...	0.401349	0.401349	0.401349	0.40136
4	0	0.004992	0.019785	0.031172	0.031172	...	0.405511	0.405511	0.405511	0.40605
5	16	0.017975	0.028234	0.036466	0.036466	...	0.417069	0.417069	0.417069	0.418042
6	17	0.020617	0.02996	0.037593	0.037593	...	0.412635	0.412635	0.412635	0.412939
7	5	0.028366	0.034168	0.0387	0.0387	...	0.458986	0.458986	0.458986	0.459253
...	...	...	...	...		...	...	...	...	...
249	7	0.027861	0.032238	0.032238	0.035482	...	0.390883	0.390883	0.390883	0.391525
250	7	0.025559	0.028484	0.028484	0.030757	...	0.391661	0.391661	0.391661	0.392434
251	16	−0.008635	0.012856	0.012856	0.02829	...	0.405308	0.405308	0.405308	0.405596
252	2	0.026924	0.030733	0.030733	0.03356	...	0.402501	0.402501	0.402501	0.402977
253	17	0.027081	0.032876	0.032876	0.037407	...	0.397983	0.397983	0.397983	0.398028
254	11	0.027094	0.031609	0.031609	0.034978	...	0.4059	0.4059	0.4059	0.406435
255	6	0.026076	0.029686	0.029686	0.032365	...	0.386275	0.386275	0.386275	0.386464

将 10 条未分类药材记录与已分类的药材记录对比，若两者之间的欧式距离最小，则可以认为二者的光谱波数数据曲线在特征区间上重合度较高，即二者极大可能产自同一产地，光谱波数数据(见表 4、表 6、表 8)。

筛选近红外光谱数据得到重合度最高的光谱数据与初步得出对应的产地为下(见表 5、表 7、表 9)。

Table 4. Symmetry curve of coincidence degree is obtained from near-infrared spectral data
表 4. 近红外光谱数据得到重合度对称曲线

No	4	15	22	30	34	45	74	114	170	209
曲线	24	141	52	160	75	134	171	161	232	39

Table 5. Preliminary origin identification based on near-infrared spectral data
表 5. 根据近红外光谱数据初步产地鉴别

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

筛选中红外光谱数据得到重合度最高的光谱数据与初步得出对应的产地为下。

Table 6. The symmetry curve of coincidence degree is obtained from the mid-infrared spectral data
表 6. 中红外光谱数据得到重合度对称曲线

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

Table 7. Preliminary origin identification based on mid-infrared spectral data
表 7. 根据中红外光谱数据初步产地鉴别

No	4	15	22	30	34	45	74	114	170	209
曲线	24	141	52	160	75	134	171	161	232	39

筛选近中红外光谱数据得到重合度最高的光谱数据与初步得出对应的产地为下。

Table 8. Symmetry curve of coincidence degree is obtained from the near-mid-infrared spectral data
表 8. 近中红外光谱数据得到重合度对称曲线

No	4	15	22	30	34	45	74	114	170	209
曲线	225	141	214	98	251	182	19	149	230	369

Table 9. Preliminary origin identification based on near-mid-infrared spectral data
表 9. 根据近中红外光谱数据初步产地鉴别

No	4	15	22	30	34	45	74	114	170	209
OP	4	6	3	9	16	3	12	6	9	15

4.2. 支持向量机(SVM)结果

本章节为作者提供“资助信息”的示例。通过欧式距离可以初步确定 10 种未分类药材记录的来源。为了提高识别的准确性，使用支持向量机求解来源不明的药材记录模型。

我们选择 LLE 降维方法[8]，局部线性降维。经过反复的参数调整，我们将数据维数降低到 35 维。这不仅保留了原始数据的主要特征，而且充分减少了计算量。

对于输入空间中的非线性分类问题，可以通过非线性变换将其转化为维度特征空间中的线性分类问题，并且可以在高维特征空间中学习线性支持向量机[9]。

输入训练数据集[10]

$$T=\left\{\left(x_1,y_1\right),\left(x_2,y_2\right),\cdots,\left(x_N,y_N\right)\right\}\ \text{其中}x_i\in X=R^n,\ y_i\in\left\{+1,-1\right\},\ i=1,2,\cdots,N;$$

输出分类决策函数；

选择适当的核函数 $K\left(x,z\right)$ 和惩罚参数 $C>0$ 来构造和求解凸二次规划问题：

$$\min_{\alpha}\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j K\left(x_i,x_j\right)-\sum_{i=1}^{n}\alpha_i$$

使得：

$$\sum_{i=1}^{n}\alpha_i y_i = 0 \quad 0 \le \alpha_i \le C,\ i = 1,2,\cdots,N$$

得到最优解 $\alpha^* = \left(\alpha_1^*, \alpha_2^*, \cdots, \alpha_N^*\right)^{\mathrm{T}}$

计算，选择 α^* 的一个分量 α_j^* 满足条件 $0 < \alpha_j^* < C$，

计算 $b^* = y_j - \sum_{i=1}^{n}\alpha_i^* y_j K\left(x_i, x_j\right)$

分类决策函数： $f(x) = sign\left(\sum_{i=1}^{n}\alpha_i^* y_j K\left(x_i, x_j\right) + b^*\right)$

调用 python 的 sklearn 库[11]中的函数 SVC 计算，将已知的数据分为训练集与测试集，通过反复调参，最终近红外光谱的模型在训练即使达到了 91.9%的正确率，而且在测试集上达到 98.0%的正确率。

中红外光谱的模型在训练集达到了 86.5%的正确率，而且在测试集上达到 98.3%的正确率。

近中红外光谱的模型在训练集达到了 83.8%的正确率，而且在测试集上达到 98.7%的正确率。鉴别出未知中药材产地如下(见表 10)所示。

Table 10. Preliminary identification results of unknown medicinal materials
表 10. 未知药材产地初步鉴别结果

No	4	15	22	30	34	45	74	114	170	209	备注
OP	1	11	1	9	4	2	16	4	16	10	近红外
OP	16	10	1	9	4	2	16	4	4	11	中红外
OP	2	11	1	9	4	2	16	4	16	14	近中红外

因为支持向量机(SVM)被用来发现在不同的数据集上获得的结果并不完全相同，并且其在再训练集和测试集上的准确性是相同的。最后，使用从特征间隔差异较大的中红外数据获得的结果来校正从近红外数据和连接表数据得到的结果。综合考虑，最终得出以下结果(见表 11)。

Table 11. Final identification results of unknown medicinal materials
表 11. 未知药材产地最终鉴别结果

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

5. 结论

传统中药材鉴别方法有很多，如来源鉴别、性状鉴别、显微鉴别、理化鉴别等，这些鉴别方法效率较低，鉴别准确度不高。DNA 分子和色谱鉴别对中药材的鉴别准确率极高，但存在预处理较为复杂且分析的时间长、成本较高、操作繁琐，快速鉴别较难等不足，而支持向量机(SVM)算法具有可用于线性或非线性分类，也可以用于回归，泛化错误率低，具有良好的学习能力且学到的结果具有很好的推广性。也就是说具有低成本、高效性且鉴别精度较高的特点。本文对原始数据进行了一阶平滑处理并且运用欧氏距离初步鉴别出中药材产地，之后进一步将原始数据进行降维处理，形成了包括原始数据在内的不同数据集利用机器学习中的支持向量机算法(SVM)进行鉴别分析，比较充分地衡量出不同中药材产地之间的差异性与区分度，为今后的中药材产地鉴别提供了一种新思路。

基金项目

ZY2022QNCX01 (青年创新人才项目)。

参考文献

[1] 刘艳, 司民真, 李家旺, 等. FTIR 结合聚类分析法在姜科植物物种分类及鉴别中的应用[J]. 光散射学报, 2016, 28(3): 252-258.

[2] 孙仁爽, 金哲雄, 张哲鹏, 许长华, 周群, 孙素琴. 牻牛儿苗科 11 种中药材红外光谱鉴定及聚类分析[J]. 光谱学与光谱分析, 2013, 33(2): 371-375.

[3] 赵建国, 曲伟红, 石向群. 傅里叶变换红外光谱法鉴定中药僵蚕[J]. 九江学院学报(自然科学版), 2016, 31(3): 98-100. https://doi.org/10.19717/j.cnki.jjun.2016.03.026

[4] 于秀丽, 贾湖. 基于模式识别技术的中成药分类与质量鉴定流程[J]. 科学技术与工程, 2012, 12(18): 4530-4534.

[5] 宗容, 施继红, 尉洪, 李海燕. 数学实验与数学建模[M]. 昆明: 云南大学出版社, 2009.

[6] 柳长源. 相关向量机多分类算法的研究与应用[D]: [博士学位论文]. 哈尔滨: 哈尔滨工程大学, 2013.

[7] 易芳吉, 钟丽莎, 李章勇. 晶千 SVM 分类器的癫抚脑电时空特征提取方法的研究[J]. 重庆邮电大学学报(自然科学版), 2022, 34(3): 444-450.

[8] 孙喜利. 高维数据的降维及聚类方法研究[D]: [硕士学位论文]. 兰州: 兰州大学, 2016.

[9] 王华军, 修乃华. 支持向量机损失函数分析[J]. 数学进展, 2021, 50(6): 801-828.

[10] 黄冠泽. 基于 LSTM-SVM 的卷对卷系统预测性维护模型[J]. 机电工程技术, 2020, 49(11): 112-115.

[11] 王娟, 华东, 罗建平. Python 编程基础与数据分析[M]. 南京: 南京大学出版社, 2019.