以药品监管理念为导向的基于近红外光谱与机器学习联用的艾叶真伪判别研究

张凯笑[1,2]，熊婧[3]，郭涛[1*]，王晓伟[2]，王海波[2]，李彦超[2]，张文静[2]，石岩[3*]（1. 河南中医药大学药学院，郑州 450046；2. 河南省药品医疗器械检验院（河南省疫苗批签中心），国家药品监督管理局中药材及饮片质量控制重点实验室，郑州 450008；3. 中国食品药品检定研究院，北京 102629）

摘要：目的 以药品监管理念为导向，建立基于近红外光谱与机器学习联用的艾叶真伪品判别的方法。**方法** 使用近红外光谱仪测定艾叶真伪样品的近红外光谱，并采用特征工程中的特征筛选、特征衍生等相关技术对实验数据进行处理。随机划分训练集和测试集，使用训练集样品数据对机器学习领域经典的逻辑回归模型进行2分类模式训练，测试集样品进行模型的评估。**结果** 使用逻辑回归模型对测试集样品的判别准确率为97%，其他各项评价指标也均在92%以上。同时，对于正品与伪品混合样品的判别也较准确。相较于传统化学计量学方法，判别准确率更高。**结论** 本研究所建立的逻辑回归模型可以实现鉴别艾叶的真伪，对药品监管工作具有技术支撑作用。

关键词：艾叶；近红外光谱；机器学习；逻辑回归；特征工程

doi：10.11669/cpj.2024.13.009　　**中图分类号：**R917　　**文献标志码：**A　　**文章编号：**1001－2494（2024）13－1238－08

A Study Guided by Drug Regulatory Philosophy on the Authenticity Discrimination of Artemisiae Argyi Folium Based on the Combination of Near-Infrared Spectroscopy and Machine Learning

ZHANG Kaixiao[1,2], XIONG Jing[3], GUO Tao[1*], WANG Xiaowei[2], WANG Haibo[2], LI Yanchao[2], ZHANG Wenjing[2], SHI Yan[3*] (1. *College of Medicine, Henan University of Chinese Medicine, Zhengzhou* 450046, *China*; 2. *NMPA Key Laboratory for Quality Control of Traditional Chinese Medicine (Chinese Materia Medica and prepared slices), Henan Institute for Drug and Medical Device Inspection (Henan Vaccine Issuance Center), Zhengzhou* 450008, *China*; 3. *National Institutes for Food and Drug Control, Beijing* 102629, *China*)

ABSTRACT: OBJECTIVE To establish a method for identifying the authenticity of Artemisiae Argyi Folium suitable for use in drug regulatory work. **METHODS** The near-infrared spectra of samples of Artemisiae Argyi Folium and counterfeit were determined, and the experimental data was processed using feature engineering related techniques, such as feature screening and feature derivation. The training set and test set were divided randomly, and the logistic regression model, a classic model in the field of machine learning, was trained in 2-class mode and evaluated with the training set data and the test set data used, respectively. **RESULTS** The discrimination accuracy of the samples in the test set was 97%, and the other evaluation indicators were also above 92% with the logistic regression model. In addition, the results of discrimination between genuine and counterfeit mixed samples were also relatively accurate. Compared with traditional chemometrics methods, the machine learning used in the study had higher discrimination accuracy. **CONCLUSION** The logistic regression model established in this study can achieve the authenticity identification of Artemisiae Argyi Folium, providing technical support for actual drug regulatory work.

KEY WORDS: Artemisiae Argyi Folium; near-infrared spectrum; machine learning; logistic regression; feature engineering

中药材艾叶是菊科植物艾（*Artemisia argyi* Lévl. et Vant.）的干燥叶，是我国常用中药材之一，主要含有挥发油类、黄酮类、萜类等化学成分，具有抗菌、抗病毒、抗炎、止血等药理活性，多用于温经止血，散寒止痛等症的治疗[1-5]。据《中国植物志》记载[6]，艾叶所属菊科蒿属植物在我国有186个种

基金项目：河南省科技厅科技攻关项目资助（222102310110）；中国药品监管科学行动计划第二批重点项目资助（NMPAJGKX-2023-030）；河南省高层次人才国际化项目资助（2021-72）；国家药品监督管理局药品监管科学体系建设重点项目“新技术新方法在中药质量控制中的应用”资助（RS2024Z006）

作者简介：张凯笑，女，硕士研究生　　研究方向：中药鉴定、资源与评价；熊婧，女，硕士，研究员　　研究方向：药品质量分析、评价与控制。张凯笑与熊婧为共同第一作者　　*** 通讯作者：**石岩，男，博士，研究员　　研究方向：中药质量评价与控制　　Tel：（010）53852081；郭涛，男，博士，教授　　研究方向：中药新药开发

及44个变种，其中不乏有难辨真伪的艾叶混伪品。课题组在药材种植基地的实地调研中发现，艾叶在生长过程中常伴生有蒙古蒿、水蒿、小艾叶等艾叶混伪品，采收时极有可能混入。此外，在药材市场的实地调研中，也发现不少以艾叶近缘种植物混作艾叶药材售卖的情况。

基于艾叶药材存在的上述问题，从药品监管工作实际情况出发，本研究开展了艾叶及其伪品的近红外光谱研究。具体方法是：以来源可靠的艾叶药材作为正品类，以种植基地和市场上常见的几种艾叶的伪品作为伪品类，这样将艾叶药材及其各种混伪品的判定归为2分类问题，并且采用不同比例正品与伪品混合模拟的形式，更贴近于实际监管工作中的检验判定。然而由于伪品类样品来源繁杂，缺乏规律可循，出现传统的化学计量学技术判别结果不准的情况。为解决这一问题，本研究采用机器学习中经典的逻辑回归算法模型，通过特征筛选和特征数据增强技术，实现艾叶正伪品近红外光谱的准确判别。考虑到药品监管工作的实际应用情况，在研究过程中使用了正品与不同伪品以及不同品种的伪品之间相混合的样品，另外在仪器参数相同的条件下，不同样品近红外光谱的测定分布在夏冬两个温湿度差异较大的季节。本研究的思路、方法与结果，可以为该品种药材的监管工作提供有力的技术支撑与有益的参考。

1 仪器与试剂

Bruker傅里叶变换近红外光谱仪（美国Bruker公司）；OPUS光谱分析软件（美国Bruker公司）；Python计算机编程语言（美国Python Software Foundation，版本：3.8.8）。本研究使用的机器学习技术与算法，所涉及到的随机种子均设置为837。

正品艾叶药材样品51批，不同种伪品艾叶药材45批（含不同种伪品之间任意比例混合样品），正品与伪品不同比例混合样品14批。样品信息见表1。

2 方法与结果

2.1 近红外光谱数据采集

取样品粉末约4 g，置于石英样品杯中，振摇样品杯使样品分布均匀，以内置背景为参比，采用积分球漫反射方式扫描。扫描范围12 000～4 000 $\cdot cm^{-1}$，扫描次数为32次，分辨率为8 $\cdot cm^{-1}$，每个样品重复测定6次，取平均光谱作为样品的最终近红外光谱数据。所有样品近红外光谱图见图1。

2.2 数据初步分析

2.2.1 基于主成分分析的数据空间分布 将1～96号样品按照各波长进行数据标准化处理，然后进行主成分分析，按照累计方差不低于95.0%为限，可将数据降维至3个主成分（实际累计方差为98.3%）数学空间，1～96号样品在3个主成分空间的分布见图2。

2.2.2 聚类分析 将1～96号样品以各批样品自身数据进行标准化处理，选择使用ward聚类算法，欧式距离度量簇间距离，结果见图3。

2.2.3 相关性分析 以近红外光谱数据计算1～96号样品间的Pearson相关系数，并绘制样品间Pearson相关热图，见图4。

2.2.4 相似度分析 将正品艾叶（1～51号）样品的近红外光谱数据求均值，得到正品艾叶的平均光谱数据，然后计算各批伪品艾叶（52～96号）样品与正品艾叶平均光谱的夹角余弦相似度，结果各批伪品艾叶与正品艾叶的相似度均达到0.991以上，将相似度绘制散点图见图5。

2.3 基于逻辑回归模型算法的正伪品判别

2.3.1 样品的数据集划分 将1～96号样品划分为2类，即艾叶正品为1类，其他艾叶的各种伪品统一归为另1类。按此划分，艾叶正品和伪品各有51批和45批，分类别打乱排序按照7∶3比例随机划分模型的训练集和测试集，结果67批样品划为训练集，29批样品划为测试集，训练集中正品和伪品批数分别为34批和33批，测试集中正品和伪品批数分别为17批和12批。

2.3.2 数据的特征筛选、衍生及预处理 样品近红外光谱采样波长点共计1 899个，通过嵌入法使用逻辑回归算法对1 899个波长特征数据进行筛选，以0.091 2作为阈值筛选出391个波长特征。然后使用二阶多项式对这391个特征进行衍生，最终可得到77 028个特征。将经过数据筛选和衍生后的训练集样品数据进行标准化预处理，用于后续机器学习建模。同法得到的测试集样品数据用于模型判别效果的测试验证。

2.3.3 逻辑回归模型算法的建立 逻辑回归模型的正则化惩罚项选择L2法，正则化强度倒数C为0.9，求解器为“liblinear”，多分类选择自动模式。使用67批训练集样品数据训练模型，29批测试集样品对完成训练的模型进行测试验证。

结果表明，模型对于训练集的67批样品的艾叶正品与伪品判别完全正确，对于29批测试集样品的

表 1 正品与伪品艾叶样品信息表

Tab. 1 Sample information of genuine and counterfeit Artemisiae Argyi Folium

Code	Origins (in Chinese)	Producer/Collection place (in Chinese)	Note
1	Artemisiae Argyi Folium(艾叶)	Anhui(安徽)	
2	Artemisiae Argyi Folium(艾叶)	Anhui(安徽)	
3	Artemisiae Argyi Folium(艾叶)	Anhui(安徽)	
4	Artemisiae Argyi Folium(艾叶)	Henan(河南)	
5	Artemisiae Argyi Folium(艾叶)	Henan(河南)	
6	Artemisiae Argyi Folium(艾叶)	Guangdong(广东)	
7	Artemisiae Argyi Folium(艾叶)	Gansu(甘肃)	
8	Artemisiae Argyi Folium(艾叶)	Gansu(甘肃)	
9	Artemisiae Argyi Folium(艾叶)	Hainan(海南)	
10	Artemisiae Argyi Folium(艾叶)	Heilongjiang(黑龙江)	
11	Artemisiae Argyi Folium(艾叶)	Hubei(湖北)	
12	Artemisiae Argyi Folium(艾叶)	Hubei(湖北)	
13	Artemisiae Argyi Folium(艾叶)	Hubei(湖北)	
14	Artemisiae Argyi Folium(艾叶)	Henan(河南)	
15	Artemisiae Argyi Folium(艾叶)	Jiangxi(江西)	
16	Artemisiae Argyi Folium(艾叶)	Yunnan(云南)	
17	Artemisiae Argyi Folium(艾叶)	Henan(河南)	
18	Artemisiae Argyi Folium(艾叶)	Neimenggu(内蒙古)	
19	Artemisiae Argyi Folium(艾叶)	Ningxia(宁夏)	
20	Artemisiae Argyi Folium(艾叶)	Qinghai(青海)	
21	Artemisiae Argyi Folium(艾叶)	Sichuan(四川)	
22	Artemisiae Argyi Folium(艾叶)	Henan(河南)	
23	Artemisiae Argyi Folium(艾叶)	Henan(河南)	
24	Artemisiae Argyi Folium(艾叶)	Henan(河南)	
25	Artemisiae Argyi Folium(艾叶)	Shanghai(上海)	
26	Artemisiae Argyi Folium(艾叶)	Yunnan(云南)	
27	Artemisiae Argyi Folium(艾叶)	Chongqing(重庆)	
28	Artemisiae Argyi Folium(艾叶)	Anhui(安徽)	
29	Artemisiae Argyi Folium(艾叶)	Anhui(安徽)	
30	Artemisiae Argyi Folium(艾叶)	Anhui(安徽)	
31	Artemisiae Argyi Folium(艾叶)	Anhui(安徽)	
32	Artemisiae Argyi Folium(艾叶)	Anhui(安徽)	
33	Artemisiae Argyi Folium(艾叶)	Anhui(安徽)	
34	Artemisiae Argyi Folium(艾叶)	Anhui(安徽)	
35	Artemisiae Argyi Folium(艾叶)	Anhui(安徽)	
36	Artemisiae Argyi Folium(艾叶)	Anhui(安徽)	
37	Artemisiae Argyi Folium(艾叶)	Anhui(安徽)	
38	Artemisiae Argyi Folium(艾叶)	Anhui(安徽)	
39	Artemisiae Argyi Folium(艾叶)	Xinjiang(新疆)	
40	Artemisiae Argyi Folium(艾叶)	Anhui(安徽)	
41	Artemisiae Argyi Folium(艾叶)	Hubei(湖北)	
42	Artemisiae Argyi Folium(艾叶)	Gansu(甘肃)	
43	Artemisiae Argyi Folium(艾叶)	Zhejiang(浙江)	
44	Artemisiae Argyi Folium(艾叶)	Jiangsu(江苏)	
45	Artemisiae Argyi Folium(艾叶)	Zhejiang(浙江)	
46	Artemisiae Argyi Folium(艾叶)	Guizhou(贵州)	
47	Artemisiae Argyi Folium(艾叶)	Beijing(北京)	
48	Artemisiae Argyi Folium(艾叶)	Zhejiang(浙江)	
49	Artemisiae Argyi Folium(艾叶)	Fujian(福建)	
50	Artemisiae Argyi Folium(艾叶)	Zhejiang(浙江)	
51	Artemisiae Argyi Folium(艾叶)	Henan(河南)	
52	Artemisiae Mongolica Folium(蒙古蒿叶)	Henan(河南)	
53	Artemisiae Lavandulaefolia Folium(野艾蒿叶)	Gansu(甘肃)	
54	Artemisiae Mongolica Folium(蒙古蒿叶)	Henan(河南)	
55	Artemisiae Mongolica Folium(蒙古蒿叶)	Henan(河南)	
56	Artemisiae Mongolica Folium(蒙古蒿叶)	Henan(河南)	
57	Artemisiae Mongolica Folium(蒙古蒿叶)	Henan(河南)	
58	Artemisiae Mongolica Folium(蒙古蒿叶)	Henan(河南)	
59	Artemisiae Mongolica Folium(蒙古蒿叶)	Henan(河南)	
60	Artemisiae Mongolica Folium(蒙古蒿叶)	Shandong(山东)	
61	Counterfeit indeterminate species(未定种伪品)	Shaanxi(陕西)	
62	Artemisiae Mongolica Folium(蒙古蒿叶)	Henan(河南)	
63	Counterfeit indeterminate species(未定种伪品)	Henan(河南)	
64	Counterfeit indeterminate species(未定种伪品)	Henan(河南)	
65	Artemisiae Mongolica Folium(蒙古蒿叶)	Henan(河南)	
66	Counterfeit indeterminate species(未定种伪品)	Henan(河南)	
67	Artemisiae Mongolica Folium(蒙古蒿叶)	Henan(河南)	
68	Artemisiae Mongolica Folium(蒙古蒿叶)	Henan(河南)	
69	Artemisiae Mongolica Folium(蒙古蒿叶)	Henan(河南)	
70	Counterfeit indeterminate species(未定种伪品)	Anhui(安徽)	
71	Counterfeit indeterminate species(未定种伪品)	Shandong(山东)	
72	Counterfeit indeterminate species(未定种伪品)	Henan(河南)	
73	Counterfeit indeterminate species(未定种伪品)	Henan(河南)	
74	Counterfeit indeterminate species(未定种伪品)	Henan(河南)	
75			mix 52 & 61
76			mix 53& 66
77			mix 55 & 66
78			mix 55, 61 & 69
79			mix 54 & 69
80			mix 56, 57 & 58
81			mix 58 & 59
82			mix 58 & 60
83			mix 59 & 60
84			mix 64 & 65
85			mix 63 & 64
86			mix 63 & 65
87			mix 62 & 67
88			mix 62 & 68
89			mix 67 & 68
90			mix 70 & 74
91			mix 70 & 71
92			mix 71 & 74
93			mix 72 & 73
94			mix 73 & 74
95			mix 56 & 74
96			mix 56 & 72
t1			mix 13 & 74(1∶9)
t2			mix 13 & 74(2∶8)
t3			mix 13 & 74(4∶6)
t4			mix 13 & 74(1∶1)
t5			mix 13 & 74(8∶2)
t6			mix 13 & 74(9∶1)
t7			mix 13 & 56(1∶9)
t8			mix 13 & 56(2∶8)
t9			mix 13 & 56(3∶7)
t10			mix 13 & 56(9∶1)
t11			mix 17 & 52(1∶1)
t12			mix 23 & 52(1∶1)
t13			mix 17 & 53(1∶1)
t14			mix 23& 53(1∶1)

注:1 ~ 51 为正品艾叶样品;52 ~ 74 为伪品艾叶样品;75 ~ 96 为伪品艾叶的任意混合样品;t1 ~ t14 为正品艾叶与伪品艾叶不同比例混合样品。

Note: 1 – 51 – genuine Artemisiae argyi folium samples; 52 – 74 – counterfeit Artemisiae argyi folium samples; 75 – 96 – randomly mixed samples of counterfeit Artemisiae argyi folium samples; t1 – t14 – mixed samples of genuine and counterfeit Artemisiae argyi folium in different proportions.

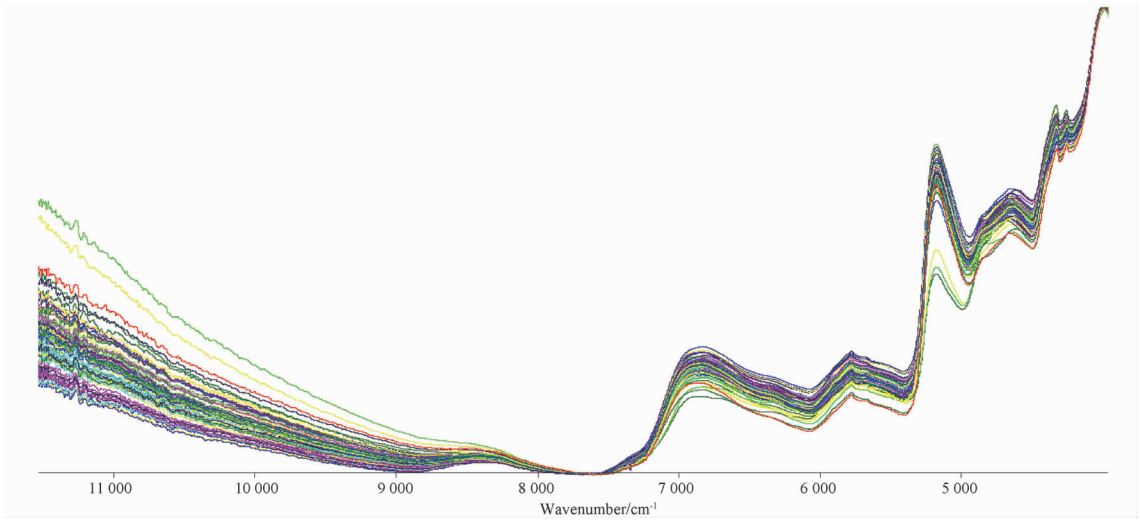

图 1 正品与伪品艾叶样品近红外光谱图

Fig. 1 Near-infrared spectrograms of genuine and counterfeit Artemisiae Argyi Folium samples

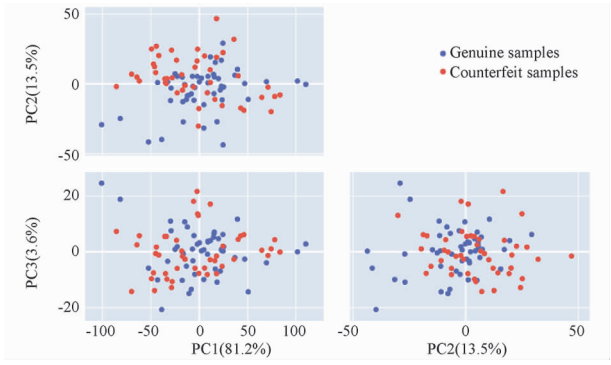

图 2 正品与伪品艾叶样品近红外光谱数据主成分分析得分分布图

Fig. 2 Scatter plots of genuine and counterfeit Artemisiae Argyi Folium sample near-infrared spectral data by principal component analysis

判别结果仅有 1 批艾叶伪品错判为正品，其余 28 批判别完全正确，准确率（accuracy）达 97%。测试集对模型的测试各项结果见表 2。

将编号为 t1 ~ t14 号的正品与伪品不同比例混合样品的近红外数据按照“2.3.2”项下操作步骤进行数据的特征筛选、衍生及以训练集样品数据的标准化参数进行预处理，然后使用上述训练得到的逻辑回归模型进行判别，结果 14 批样品中有 4 批样品判为正品，分别为 t5、t6、t10、t11。根据逻辑回归模型输出结果，这 4 批样品判为伪品概率分别为 7%、23%、11%、30%。

3 讨 论

3.1 艾叶正品与伪品近红外光谱的数据特点分析

由于艾叶正品与伪品的生物学种属关系较近，近红外光谱表现出较高一致性，这点直观表现在图 1 样品的近红外光谱图中。在“2.2”部分中，使用主成分分析法对样品的光谱数据进行降维分布展示（图 2），在 95% 以上累计方差情况下，正品与伪品分布重叠性极高，进一步表明本研究所使用的样品近红外光谱数据一致性极高，依靠简单而常规的化学计量学分析方法可能较难达到准确判别的目的，这一点在此后的聚类分析、相关性分析以及相似度分析结果中也得到验证（图 3 ~ 5）。图 3 表明，正品与伪品近红外光谱数据即使进行了标准化处理，聚类并不正确；图 4 表明，大多数样品的近红外光谱彼此之间都具有很强的相关性，相关系数与是否为正品和伪品无关；图 5 则表明，各批伪品与正品的平均近红外光谱数据的相似度多分布在 0.998 ~ 1.00 之间，最低相似度也接近 0.992，可见伪品近红外光谱与正品具有极高的相似度。

以上对数据进行探索的结果不仅仅与艾叶正品和伪品近红外光谱数据本身强一致性有关，可能也与本研究在 2 分类判别模式中，将不同植物种的艾叶伪品统一归为 1 类有一定的关系。

上述一系列化学计量学分析结果表明，艾叶的正品与伪品近红外光谱组间一致性较高，差异并不显著，仅使用简单直接的算法可能无法准确判别正品与伪品。这一点在本研究前期使用光谱仪器商业软件中距离匹配法（distance match）算法，结合 Savizky-Golay 卷积平滑去噪、多元散射校正、标准正则变换、一阶或二阶导数等多种数据前处理方法，分析判别准确率最高仅为 82% 的结果也可以得到佐证。

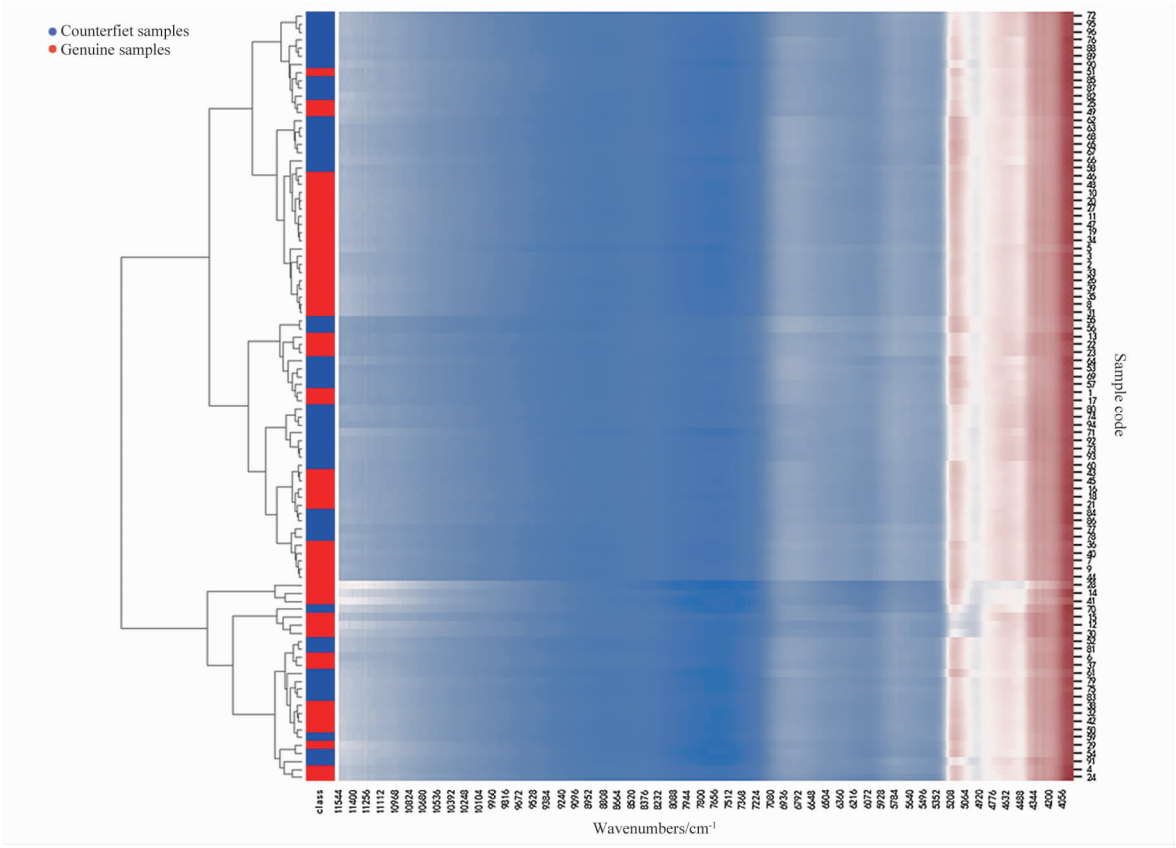

图 3 正品与伪品艾叶样品近红外光谱聚类分析热图

Fig. 3 Cluster analysis heat map of genuine and counterfeit Artemisiae Argyi Folium sample near-infrared spectral data

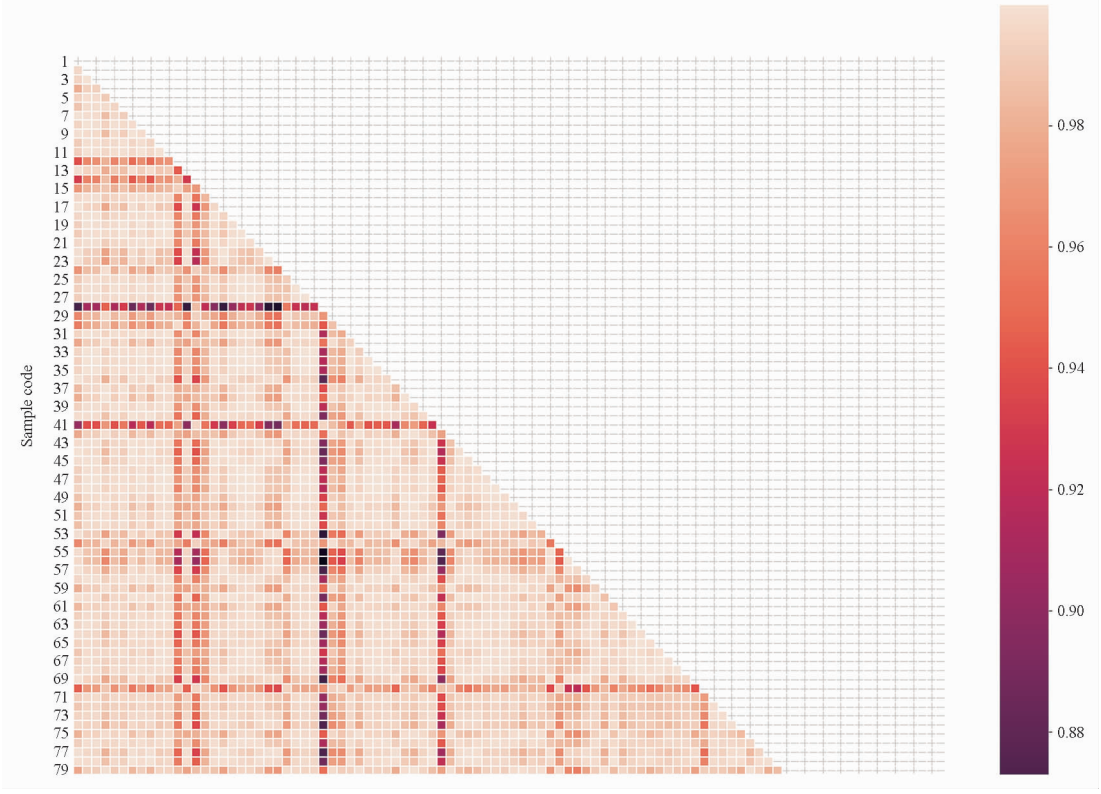

图 4 正品与伪品艾叶样品近红外数据 Pearson 相关性热图

Fig. 4 Pearson correlation heat map of genuine and counterfeit Artemisiae Argyi Folium sample near-infrared spectral data

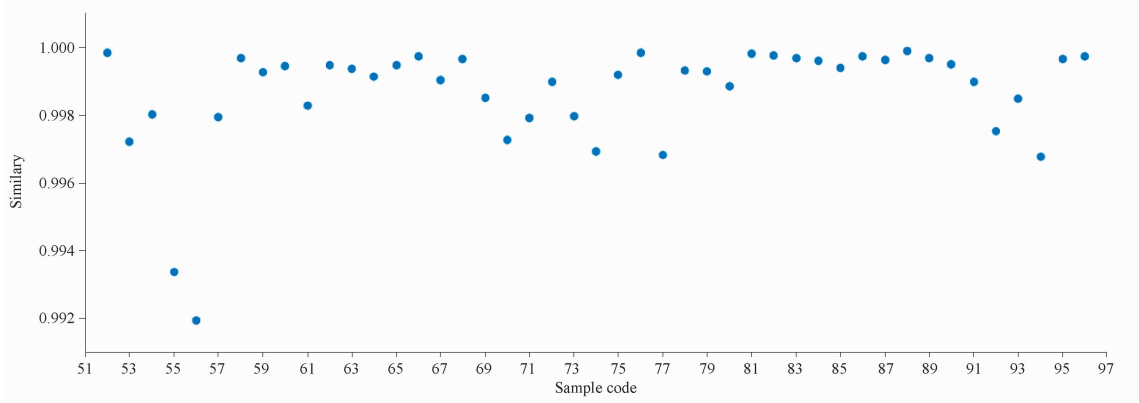

图 5 伪品艾叶样品近红外光谱数据相似度散点图

Fig. 5 Scatter plot of similarity of counterfeit Artemisiae Argyi Folium sample near-infrared spectral data

表 2 逻辑回归模型对测试集判别结果

Tab. 2 Discrimination results of test set by logistic regression model

Class	Number of batches	Precision /%	Recall /%	F1-score /%
Authentic drugs	17	94	100	97
Counterfeit drugs	12	100	92	96

3.2 机器学习判别模型算法的选择

逻辑回归模型属于广义的线性模型，是通过在线性模型中加入 Sigmoid 函数作为联系函数的反函数构建而得。其数学概念及原理在机器学习相关专著[7-10]及相关研究文献[11-13]中有较详细的描述，在此不再赘述。逻辑回归模型是机器学习领域内的经典的分类判别模型，由于其模型算法含有线性回归部分，故对于线性相关的数据具有天然的优势，在金融、医学、农林资源等领域内有着广泛的应用[11-16]。

基于药品监管实际情况，优选将样品数据按照正品与伪品划分为 2 个类别，而传统的逻辑回归模型正是属于 2 分类判别模型，且光谱行为也常为线性相关，因此本研究选择以传统 2 分类逻辑回归模型作为判别艾叶真伪的算法模型。

3.3 近红外光谱数据的特征工程

在机器学习领域内，对获得的原数据进行特征工程往往能够极大地提升机器学习模型的判别能力。本研究将样品按照“2.3.1”项下随机划分训练集和测试集，然后使用训练集样品的近红外光谱原数据训练逻辑回归模型，再使用测试集样品对该逻辑回归模型进行评估，结果模型对于测试集判别准确率分别仅为 52%，表明此时模型几乎没有判别能力。

为解决这一问题，本研究从特征工程传统角度出发，分别采取了数据预处理、特征筛选以及特征衍生的技术对样品的近红外光谱数据进行了处理。

在数据预处理过程中，根据光谱数据特点，选取相同波长点数值标准化作为预处理方法。样品数据经过标准化预处理后，训练得到的模型对测试集的判别准确率已达到 97% 左右，即 29 批测试集样品仅有 1 批判错。然而，训练集以及交叉验证的准确率低于测试集，结合逻辑回归模型特性，判断模型可能存在欠拟合情况，需要通过增加特征复杂程度解决。原数据中每批样品在测定的近红外波长范围内共有 1 899 个特征数据，若直接对这些特征进行衍生处理，会极大增加数据量，对算力造成不必要的负担，因此考虑特征筛选。

在特征筛选时，根据逻辑回归可表征特征与类别标签关系这一性质，使用了基于逻辑回归的嵌入法（embedded）。图 6 表明，在部分阈值范围内，经过特征选择后训练得到的模型可以比使用全部特征训练得到的模型具有更高的判别准确率。同时可知，当阈值为 0.052 1 ~ 0.091 2 时，特征选择效果最佳。阈值越高，选择得到的特征数越少，基于尽量剔除无用特征的考虑，本研究选择阈值为 0.091 2，对应的特征数为 391 个。经过数据预处理和特征筛选后，逻辑回归模型的判别准确率有了极大的提升，对训练集样品的判别准确率由 57% 提升至 92%，对测试集样品的判别准确率由 52% 提升至 96%。然而样品的训练集和测试集的判别准确率的差异，提示模型可能欠拟合，因此尝试对数据进行特征衍生处理，具体衍生手段为多项式法。多项式法是特征自身、特征之间的组合衍生，经过处理后，上述筛选得

到的 391 个特征可衍生至 77 028 个，通过特征衍生后模型对训练集和测试集样品的判别准确率分别为 100% 和 97%。仅有 1 批测试集中的样品判别错误。

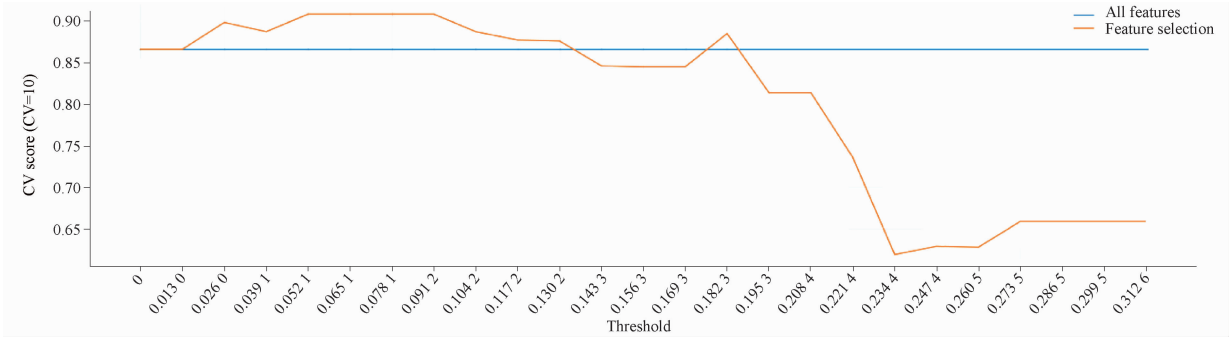

图 6 正品与伪品艾叶样品近红外光谱数据对基于逻辑回归模型的特征选择阈值学习曲线

Fig. 6 Learning curve of feature selection threshold based on logistic regression model for genuine and counterfeit Artemisiae Argyi Folium sample data

图 7 为经过特征工程处理后，分别以训练集、10 折交叉验证以及测试集判别准确率对模型的评估结果比较。一般来说，测试集对模型的效果评估最有意义，从这方面来看，后 3 种特征工程处理方式和流程效果相当，结合训练集及 10 折交叉验证的准确率结果，最终选择特征筛选、特征衍生和数据标准化的组合。

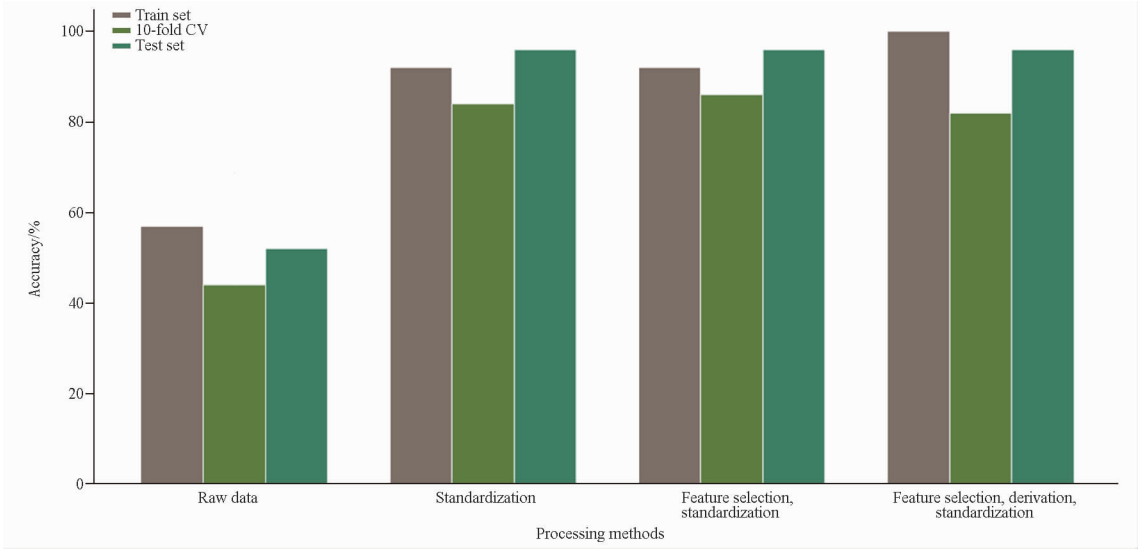

图 7 正品与伪品艾叶样品近红外光谱数据经不同方式处理后准确率比较

Fig. 7 Comparison of accuracy of different processing methods for genuine and counterfeit Artemisiae Argyi Folium samples

3.4 对于正品与伪品混合样品的判别

在药品监管工作的场景中，不仅会面临样品完全是伪品的情况，还有可能出现需要对正品和伪品混合的样品进行判别的情况。因此，正确而客观地评估模型的判别效果以及其在药品监管工作中的实际应用价值，应该考虑不同比例正品和伪品混合的情况。鉴于此，本研究制作了不同比例正品和伪品的混合样品（编号 t1 ~ t14）作为额外的测试集，对模型进行评估。结果除了 4 批样品（编号 t5、t6、t10 和 t11）判别为正品外，其余样品均判别为伪品。这 4 批样品的正品与伪品混合的比例分别为 8∶2、9∶1、9∶1 和 1∶1。这一结果表明，该模型具备一定的艾叶的正品和伪品混合的情况进行判别的能力，即使严格按照 2020 年版《中国药典》中“药材和饮片检定通则”的规定“药屑及杂质通常不得过 3%”，将所有正品和伪品混合的样品作为不合格样品，那么对这一额外的测试集的判别正确率为 71%。

逻辑回归模型经过 Sigmoid 函数处理后，输出

结果被限定在 0 ~ 1 之间。在通常情况下,以数值 0.5 为阈值应用于二分类的判别,而输出结果可视为在逻辑回归模型中,待判别样品被判别为 1 代表类别的概率 P,由此可推出该样品为 0 代表类别的概率为 1-P。按照此理论,本研究中的逻辑回归模型以 1 代表伪品类别,因此得出了 t5、t6、t10、t11 为伪品的概率值分别为 7%、23%、11%、30%。概率最高的 t11 样品也是这 4 批样品中实际伪品混合比例最高的。

4 结论与展望

本研究是基于机器学习相关技术与模型,采用近红外光谱技术对艾叶进行的真伪判别研究。基于药品监管工作中实际可能面临的情况,将不同品种的伪品相混合组成部分伪品数据,且在额外测试集使用了正品艾叶与不同伪品进行混合的样品。机器学习模型在药品监管工作中的使用,数据的获得条件可能是多样化的,因此采用近红外光谱测定时选择了夏、冬两个温湿度差异相对大的季节进行,最终结果也令人满意。由于条件所限,本研究难以完全将药品监管工作可能面临的情况一一考察,然而所幸的是,机器学习可以通过后期的在线学习与增量学习继续不断实现模型的迭代与优化,使其更加贴近和适配实际应用。

REFERENCES

[1] *Ch. P*(2020) Vol Ⅰ(中国药典 2020 年版. 一部)[S]. 2020: 91.

[2] LAN X Y, ZHANG Y, ZHU L B, *et al*. Research progress on chemical constituents from Artemisiae Argyi Folium and their pharmacological activities and quality control[J]. *China J Chin Mater Med*(中国中药杂志), 2020, 45(17):4017-4030.

[3] YUAN X L, WU H Y, QIU C L. A brief analysis of pharmaceutical active ingredients, pharmacological action and clinical application of Artemisia argyi[J]. *Contemp Med Symp*(当代医药论丛), 2020, 18(2):171-173.

[4] ZHANG X L, CHEN X W, WU Y M. Research progress on chemical constituents and pharmacological activities of volatile oil of Aiye (Artemisia Argyi)[J]. *Chin Arch Tradit Chin Med*(中华中医药学刊), 2021, 39(5):111-119.

[5] ZHANG X X, LIU R X, WANG Y X, *et al*. Processing evolution and modern research progress of *Artemisia argyi*[J]. *J China Pharm*(中国药房), 2023, 34(6):758-762.

[6] Editorial Committee of Flora of China, Chinese Academy of Sciences. *Flora of China*(中国植物志)[M]. Vol 76. Beijing: Science Press, 1991:2.

[7] LI H. *Statistical Learning Methods*(统计学习方法)[M]. Beijing: Tsinghua University Press, 2019:91-110.

[8] HASTIE T, TIBSHIRANI R, FRIEDMAN J. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*[M]. Berlin: Springer, 2009:119-128.

[9] GÉRON A. *Hands-on Machine Learning with Sciket-Learn, Keras & Tensorflow: Concepts, Tools, and Techniques to Build Intelligent Systems*[M]. 2nd, ed. Sebastopol, CA: O′Reilly Press, 2019: 144-154.

[10] MURPHY P K. *Machine Learning: A Probabilistic Perspective*[M]. Massachusetts: The MIT Press, 2012: 21-22.

[11] LIANG H L, WANG W H, GUO F T, *et al*. Comparing the application of logistic and geographically weighted logistic regression models for Fujian forest fire forecasting[J]. *Acta Ecol Sin*(生态学报), 2017, 37(12): 4128-4141.

[12] SHI Y, LIU W, WEI F, *et al*. UPLC-QDA and machine learning for distinguishing different commodity specifications of Fritillariae Cirrhosae Bulbus and application of data augmentation technology[J]. *China J Chin Mater Med*(中国中药杂志), 2023, 48(16):4370-4380.

[13] LI B Y, GUO D H, ZHU Y, *et al*. Automatic monitoring of thrombocytopenia caused by bevacizumab in 4864 inpatients and related influencing factors[J]. *Chin J Pharmacovigil*(中国药物警戒), 2022, 19(12):1362-1367.

[14] WU P Q, YANG Y L, ZHOU Y L, *et al*. Construction of MRI radiomic prediction models for the differentiation of benign and malignant lesions of breast[J]. *J Mol Imaging*(分子影像学杂志), 2021, 44(5):764-770.

[15] WU Q S, LU Z Q, LIU Y, *et al*. Machine learning for early warning of cardiac arrest: a systematic review[J]. *Chin J Evid-Based Med*(中国循证医学杂志), 2021, 21(8):942-952.

[16] WANG H, ZHANG X P, GONG H W, *et al*. Evaluation on the effects of different machine learning algorithms on the postoperative hypoproteinemia risk prediction model for elderly orthopedic patients[J]. *Mod Tradit Chin Med Mater Med World Sci Technol*(世界科学技术-中医药现代化), 2020, 22(10):3615-3621.

(收稿日期:2024-01-03)