第42卷，第7期 光 谱 学 与 光 谱 分 析 Vol. 42, No. 7, pp2148-2155
2022年7月 Spectroscopy and Spectral Analysis July, 2022

随机森林结合 CatBoost 的近红外光谱药品鉴别

蒋 萍[1]，路皓翔[2]，刘振丙[2*]

1. 广西警察学院信息技术学院，广西 南宁 530028
2. 桂林电子科技大学计算机与信息安全学院，广西 桂林 541004

摘 要 药品质量关乎人民健康和国家命脉，随着社会经济的飞速发展对药品质量的快速、有效鉴别具有极其重要的作用。光谱分析技术具有较高的准确性、较快的分析速度且对样品不存在污染等突出优点，广泛应用在化工、石油以及医药等重要的领域。为了解决传统药品鉴别模型存在的鉴别精度低、鉴别速度不能满足实际需求且鉴别模型稳定性差的问题，采用光谱仪采集药品的近红外光谱数据达到对药品无污染鉴别的目的。结合随机森林和 CatBoost 对药品进行分类鉴别，以实现快速且准确的鉴别。首先采用随机森林(RF)对光谱仪采集的光谱数据进行有效特征波长的筛选，从而将药品光谱数据中的无关波长去除、筛选出最能表征样品属性的特征波长，然后以极限学习机(ELM)作为 CatBoost 的弱分类器分析筛选的特征波长对药品的属性鉴别。由于 ELM 仅只含有一个隐含层且无需多次迭代寻优保证了鉴别模型运行速度更快，CatBoost 通过集成弱分类器以改善模型鉴别准确性。为对所提出的药品鉴别模型性能进行有效评估，采用随机抽取训练集的方式构造不同规模药品光谱数据并分别上进行独立实验且以 10 次运行结果的均值作为其最终结果，并通过与 CatBoost、持向量机(SVM)、反向传播网络(BP)、ELM、波形叠加极限学习机(SWELM)和 Boosting 进行对比，进一步对模型的性能进行评估。从不同规模训练集的分类结果可看出，随着训练集样本的增加分类精度最高为 100%且预测标准偏差趋于 0。实验结果表明，所建立 RF-CatBoost 鉴别模型在不同规模的药品数据集上较对比方法具有更高的分类准确率、更快的速度且其鲁棒性更强，能够广泛应用于药品类别的准确鉴别，从而实现药品质量的有效监督。

关键词 近红外光谱；随机森林；极限学习机；CatBoost
中图分类号：O657.33 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2022)07-2148-08

引 言

近红外光谱分析技术具有对检测样品零污染、检测速度快等突出优点，其结合化学计量学和机器学习广泛应用在石油化工[1]、疾病诊断[2-3]、农副产品质量检测[4]等领域。药品质量的有效监督对于维系国计民生至关重要，引起了全球各国政府的关注，我国也专门成立了国家食品药品监督管理总局对药品质量进行监督[5]。然而传统的药品鉴别模型较为复杂、精度较低且运行时间较长不能满足实际的需要，因此构建快捷、准确的药品鉴别模型是一项极为重要的工作。

近红外光谱分析技术依据构成样品的不同成分对于近红外光谱的吸收性不同实现样品属性及质量的检测，机器学习可对高维数据进行处理挖掘出最能表征样本属性的特征[6-8]，国内外学者尝试将机器学习和近红外光谱分析技术相结合起来应用于药品质量检测[9,11]。有研究采用小波变换对药品光谱数据进行处理，通过稀疏降噪自编码提取药品光谱数据深层特征并由持向量机(support vector machine，SVM)进行药品类比鉴别。周颖[12]等通过构建女贞子的近红外光谱快速检测模型，实现了真假女贞子及其产地的准确鉴别。Rodionova[13]等建立了一种数据收集、模型构建和模型校验的研究过程，能够有效区分假冒药品和真实药品。Sampaio[14]等利用偏最小二乘判别分析(partial least squares discriminant analysis，PLS-DA)和 SVM 对米粉的光谱数据进行分析，有效解决了米粉制造商的准确区分。然而，由于样品的近红外光谱维度较高且存在严重的谱区重叠问题，无疑会对模型的鉴别性能产生较大的影响[15-16]，因此筛选出有效的、最能表征样品特征属性的波长或波长范围对于构建有效且可靠的近

收稿日期：2022-01-12，**修订日期**：2022-03-28
基金项目：国家自然科学基金项目(61866009)，广西重点研发计划项目(桂科 AB22035034)，广西警察学院校级科研课题(2021KYA01)资助
作者简介：蒋 萍，女，1981 年生，广西警察学院信息技术学院副教授 e-mail：j_pingzi@163.com
* 通讯作者 e-mail：zbliu@quet.edu.cn

红外光谱分析模型具有重要意义。陈文丽[17]等采用最小角回归算法筛选柑橘叶片的近红外光谱有效波长，并利用极限学习机(extreme learning machine，ELM)对筛选的有效波长进行分析实现柑橘叶片是否带有黄龙病的检测。沈东旭[18]等通过在神经网络的约束损失函数进行光谱数据中有效数据的筛选，提高了血液鉴别模型的性能。Chen[19]等研究了基于卷积神经网络(convolutional neural networks，CNN)的特征波长选择方法，研究发现 CNN 的参数对其性能产生较大的影响。虽然利用筛选出来的样品特征波长用于构建近红外光谱分析模型可以有效改善其性能，但是在药品鉴别领域的报道仍较少。

本研究将无损、检测快速的近红外光谱分析技术与机器学习方法相结合用于药品的准确鉴别。为减少光谱数据谱区重叠及无关变量对药品鉴别模型性能的影响，结合随机森林(random forest，RF)和 CatBoost 提出了一种新的近红外光谱药品鉴别方法。首先采用随机森林筛选出药品近红外光谱数据中最能表征样品特征的波长，再利用 CatBoost 对筛选出来的样品特征波长进行分析实现不同厂商药品的分类鉴别；以药品的近红外光谱数据为实例评价该方法的有效性，并与同类方法进行实验对比。本研究主要特点：

(1)将随机森林算法用于筛选最能表征样品属性的特征波长点，可有效剔除样品近红外光谱中无关变量对模型性能的影响；

(2)为确保模型具有较高的预测精度，采用决策树作为 CatBoost 中的弱分类器保证模型的预测精度更高、鲁棒性更强。

1　RF-CatBoost 模型

RF 是一种结合决策树和特征选择的集成学习方法，解决了传统决策树分类规则复杂易陷入局部最优解的问题，常用于特征变量选择、分类及异常点检测等[20]。CatBoost[21-22]以对称决策树为弱分类器，将样本特征组合在一起便于充分利用样本特征间的信息且丰富了样本的特征；此外，为了降低样品数据中噪声对模型性能的影响，采用排序提升的方法对数据进行处理，能够解决模型过拟合的问题，提升其准确性及泛化能力。结合随机森林较优的特征选择能力和 CatBoost 较强的分类能力提出了一种新的药品鉴别模型——RF-CatBoost，其模型结构如图 1 所示。

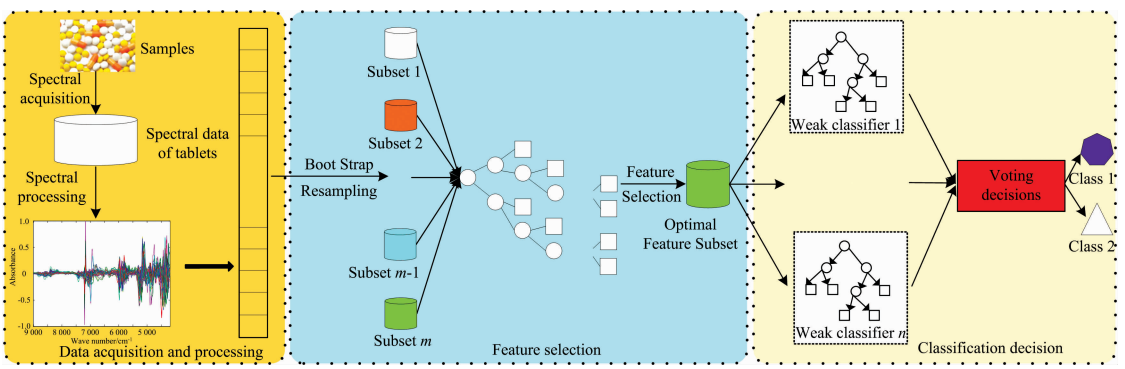

图 1　RF-CatBoost 的结构

Fig. 1　The structure of RF-CatBoost

鉴别模型主要分为两个部分：RF 特征波长选择和 CatBoost 分类决策，即首先采用 RF 对经过预处理后药品近红外光谱数据的特征波长进行筛选，然后将筛选出来的样品波长送入 CatBoost 中对样品类别进行决策。若样品的原始集合为 D，其中 N 为样品总数，X_i 表示第 i 个样品的特征波长集合，y_i 表示第 i 个样品的类别属性，则 RF-CatBoost 实现类别确定的详细过程如下：

Stage Ⅰ：波长选择

袋外误差是袋外数据真实值与预测值之差，袋外误差总和是所有袋外总据误差总和。

首先，从样品总数 N 中有放回 Bootstrap 采样 n 次构成子集 B_1，B_2，…，B_j，见式(1)

$$B_j = \text{Bootstrap}(D\{(X_i, Y_i), i = 1, \cdots, N\}) \tag{1}$$

采用 B_j 对随机森林中的决策树进行训练并计算 B_j 对应的袋外数据 OOB_j 预测结果的误差，

$$\text{err}OOB_j = \text{err}\{\text{Test}[OOB_j, \text{model}(B_j)]\} \tag{2}$$

则 n 个子集的袋外误差总和 $\text{err}OOB_{jT_1}$ 为

$$\text{err}OOB_{jT_1} = \sum_1^n \text{err}OOB_j \tag{3}$$

在袋外数据 OOB_j 的特征变量 S_m 上加入噪声记为 OOB'_j 重新依据式(1)—式(3)计算其袋外误差总和 $\text{err}OOB_{jT_2}$，见式(4)

$$\text{err}OOB_{jT_2} = \sum_1^n \text{err}OOB'_j \tag{4}$$

并计算两个袋外误差的均值 err_M[见式(5)]，差值越大说明特征变量 S_m 越重要。

$$\text{err}_M = \frac{\text{err}OOB_{jT_2} - \text{err}OOB_{jT_1}}{n} \tag{5}$$

重复以上操作计算出所有特征变量的 err_M，并按照 err_M 从小到达的顺序排列将最重要的前 n 个特征变量作为第 i 个样本 X_i 的特征集合 $X_i = \{s_1, s_2, \cdots, s_n\}$，输入 CatBoost 进行分类决策。

Stage Ⅱ：类别决策

首先需要将筛选的样品特征集合 $X_i = \{s_1, s_2, \cdots, s_n\}$ 中的

特征波长进行随机排列，构成新的样品特征集合 S'_i，见式(6)

$$X'_i = \text{Random}(X_i) = \text{Random}(s_1, s_2, \cdots, s_n) \quad (6)$$

利用 S'_i 构建新的决策树 $f(X_i)$拟合 CatBoost 的梯度，最终得到 CatBoost 分类模型，见式(7)

$$P(y \mid X_i) = \frac{1}{1 + e^{-f(X_i)}} \quad (7)$$

为了对训练集样本的类别进行确认，采用 RF 对其特征波长进行选择，由式(7)对依据其选择的特征波长进行类别的确定。

2 实验部分

为保证 RF-CatBoost 在进行药品鉴别时具有较高的训练精度，需对模型中的参数进行确定。首先对实验室数据、数据预处理及数据集划分进行简要概述，然后对 RF-CatBoost 中决策树数目做确定，最后给出 RF-CatBoost 模型建立的过程。

2.1 实验数据

实验以湖南方盛制药、江苏正大、山东鲁抗和山东罗欣生产的铝塑和非铝塑两种包装方式的头孢克肟片光谱数据为例。该光谱数据由中国食品药品检定所提供，采用 Bruker Matrix 光谱仪采集，光谱仪的采样间隔设定为 4 cm^{-1}，采样范围为 4 196～9 002 cm^{-1}，每个样本的吸光点为2 074个。实验中头孢克肟片近红外光谱数据如表 1 所示。

表 1 头孢克肟片近红外光谱数据信息

Table 1 Near infrared spectral data of cefixime tablets

厂商	非铝塑包装	铝塑包装	合计
湖南方盛制药股份有限公司	54	54	108
江苏正大清江制药有限公司	63	56	119
山东鲁抗医药股份有限公司	51	40	91
山东罗欣药业股份有限公司	48	48	96
共计	216	198	414

2.2 数据预处理

四个厂商生产的铝塑和非铝塑包装方式头孢克肟片的光谱共 414 条，这些光谱间存在明显的重叠且包含噪声，影响了样品光谱信息的解析。为了消除样品光谱间的重叠、提高样品光谱间的辨识度，采用对样品的光谱数据依次进行平滑化、归一化处理消除光谱数据中的背景干扰，消除光程差异带来的光谱变化。经过预处理后的头孢克肟片光谱信息如图 2 所示。多阶段预处理增加了样品光谱数据间的辨识，提高药品鉴别模型的准确度。

数据白化是指将数据的协方差矩阵进行单位化处理，保证数据的方差一致且特征间相互独立。其详细过程如下：

首先，构建预处理后光谱数据 x 的协方差矩阵，见式(8)

$$\boldsymbol{\Sigma} = \boldsymbol{E}(xx^{\mathrm{T}}) \quad (8)$$

若光谱数据 x 的变量相关，则其 $\boldsymbol{\Sigma}$ 为非对角矩阵。

将协方差矩阵 $\boldsymbol{\Sigma}$ 对角化，见式(9)

$$\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\Phi} = \boldsymbol{\Lambda} \quad (9)$$

式(9)中，$\boldsymbol{\Lambda}$ 为对角矩阵，其对角元素由协方差矩阵 $\boldsymbol{\Sigma}$ 的特征值组成。$\boldsymbol{\Phi}$ 为特征值对应的特征向量。对 x 进行解相关，见式(10)

$$y = \boldsymbol{\Phi}^{\mathrm{T}}x \quad (10)$$

y 为解除相关后的数据，其协方差矩阵 $\boldsymbol{E}(yy^{\mathrm{T}})$为对角矩阵。

最后，将光谱数据与对角矩阵相乘即可得到白化后的数据 w，见式(11)

$$w = \boldsymbol{\Lambda}^{1/2}y = \boldsymbol{\Lambda}^{1/2}\boldsymbol{\Phi}^{\mathrm{T}}x \quad (11)$$

采用单位化处理后的协方差矩阵构建模型有利于提高模型鉴别能力，故而对经过预处理后的药品光谱数据进行白化处理，并将药品光谱数据协方差矩阵的对角元素按照“从大到小”原则排列，其值越小包含的有效信息越少，颜色越接近深蓝色。白化处理前后药品光谱数据的协方差矩阵如图 3 所

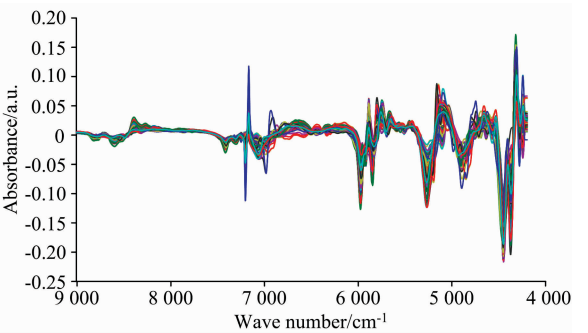

图 2 预处理后头孢克肟片的近红外光谱

Fig. 2 NIR spectra of pretreated cefixime tablets

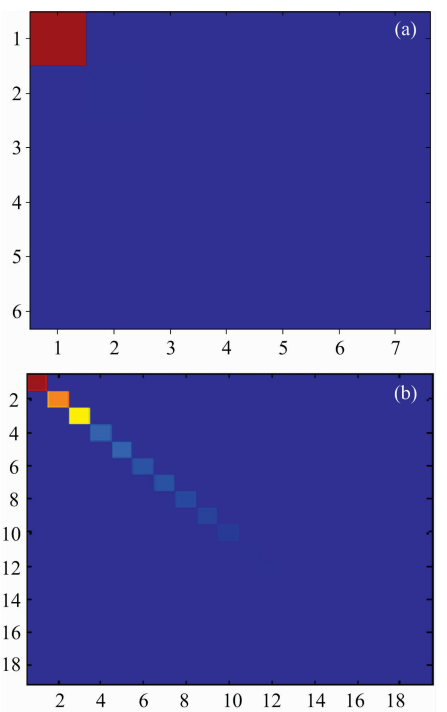

图 3 白化处理前(a)和白化处理后(b)药品近红外光谱数据的协方差矩阵

Fig. 3 Covariance matrix of drug NIR data before (a) and after (b) pretreatment

示。从图 3(a,b)可看出，预处理前药品光谱协方差矩阵的前三个分量大于其他分量，说明样品等光谱信息主要集中在前三个分量；预处理后药品光谱协方差矩阵对角线所占面积缩小，说明更多的药品光谱信息显示出来。

2.3　数据集划分

本实验中将非铝塑包装方式的头孢克肟片光谱数据记为 A 组、铝塑包装方式的头孢克肟片光谱数据记为 B 组。其中 A 组和 B 组中均将江苏正大生产的头孢克肟片的光谱数据作为正类样本，其他厂商生产的头孢克肟片的光谱数据作为负类样本，按照表 2 构建出不同规模的训练集集进行实验，验证各模型在不同规模训练集中的性能。

表 2　A 和 B 组中不同数量训练集配置表

Table 3　Configuration table of different number of training sets in group A and B

数据集	样本总数	正样本数	负样本数
A	40	15	25
	60	20	40
	80	25	55
	100	30	70
	120	35	85
	140	40	100
	160	45	115
	180	50	130
B	30	10	20
	50	15	35
	70	20	50
	90	25	65
	110	30	80
	130	35	95
	150	40	110
	170	45	125

2.4　CatBoost 中决策树个数的确定

CatBoost 中决策树数目较多则会增加其运行时间，决策树数目较少则会降低其鉴别精度。因此在建立 RF-CatBoost 鉴别模型时需确定 CatBoost 中决策树数目。图 4(a,b)分别为 CatBoost 模型在不同规模训练集、不同决策树数目下两种包装形式的头孢克肟片光谱数据的分类精度。从图中可看出，CatBoost 中决策树的数目在 200～300 间时，其在两种包装形式的不同规模头孢克肟片光谱训练集的分类精度较高。当决策树的数目超过 300 时，随着决策树数目的增加 CatBoost 模型的分类精度反而降低。据此分类，本次在构建 RF-CatBoost 药品鉴别模型时将 CatBoost 中的决策树数目设定为 250。

2.5　模型实现

基于 RF-CatBoost 的药品鉴别模型编程采用 MATLAB 2014A 实现，其中 RF 的源代码使用的是 Abhishek Jaiantilal 开源的工具箱（https://code.google.com/p/randomforest-matlab/）。RF-CatBoost 模型的性能评估实验运行在 Intel(R) Core(TM) i5-2450M CPU 的计算机上，系统版本是 Windows 10 专业版，其详细过程如下：

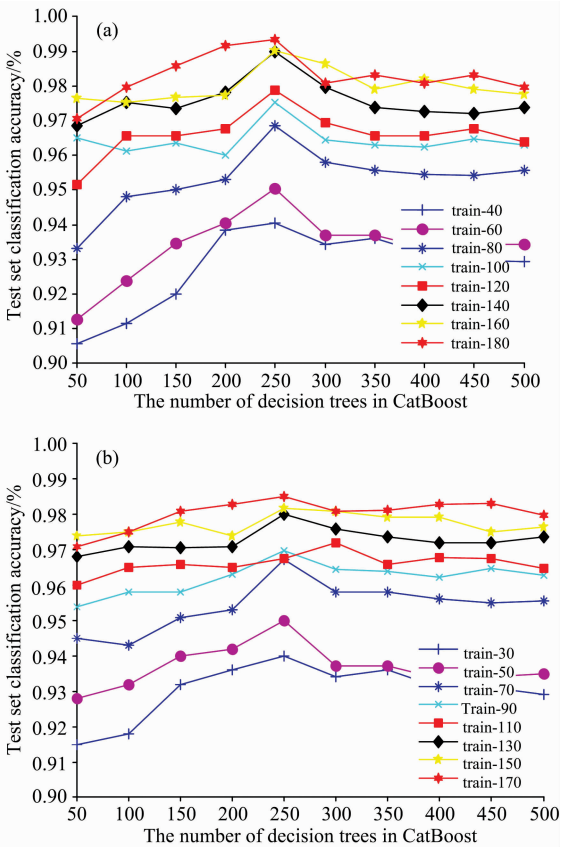

图 4　CatBoost 中不同决策树数目在 A 组(a)和 B 组(b)不同规模数据集上的分类精度

Fig. 4　Classification accuracy of different decision tree numbers in Catboost on datasets of different sizes in group A (a) and group B (b)

(1)数据预处理

由于药品的光谱数据中存在重叠且包含噪声，故采用多阶段预处理的方式对药品的光谱数据进行处理。为了提高模型性能，对经过多阶段预处理后的样品光谱数据协方差矩阵进行白化处理。

(2)特征筛选

采用随机森林筛选出预处理后药品光谱数据中最能表征其属性的特征波长，用于训练 CatBoost 分类决策模型。

(3)分类决策

按照表 2 划分出不同规模的训练集，并将筛选出的药品特征波长输入 CatBoost 中进行模型的训练。将测试数据输入训练好的 CatBoost 模型中进行药品类别的确定。

(4)对比分析

以 CatBoost、ELM、SVM、反向传播网络(back propagation，BP)、波形叠加极限学习机(summation wavelet extreme learning machine，SWELM)、Boosting 作为对比方法验证该方法在运行时间、分类精度以及稳定性方面的表现。其中 CatBoost 中决策树的数目选择为 250；SWELM 和 ELM 网络结构的构成均为 2074-Train_num * 0.4-400-2，均选用 Sigmoid 作为网络的激活函数，迭代次数设定为 50 次，设定

两层的学习率均为 0.03；SVM 的核函数选择为线性函数，且设定 $C=1$，gamma＝0.3；BP 的网络结构设置为 2074-800-400-2，选用 Sigmoid 作为网络的激活函数，迭代次数设定为 50 次，设定网络学习率为 3%。

3 结果与讨论

为评估 RF-CatBoost 在不同规模训练集中的表现，每个规模的训练集按照表 2 中正样本和负样本的数目随机抽取 10 次进行实验并与 CatBoost、ELM、SVM、BP、SWELM 和 Boosting 模型对比，将各模型在每个规模训练集上 10 次运行时间、分类精度及预测标准偏差的均值作为各模型的最终性能指标值。

(1)分类精度

分类精度是对 RF-CatBoost、CatBoost、ELM、SVM、BP、SWELM 和 Boosting 模型药品鉴别结果可靠性的衡量，分类精度越高说明药品鉴别模型的可靠性越高。各药品鉴别模型在不同规模训练集 A 和 B 上的分类精度如表 3 所示。从表 3 可看出随着，在 A 和 B 两组数据集中各模型的性能随着训练集样本的增加均逐渐增加，当 A 组中训练集增强到 120 后 RF-CatBoost 药品鉴别模型的分类精度均达到 100%；当 B 组训练集增加到 130 后，RF-CatBoost 药品鉴别模型的分类精度均达到 100%。其中在各组数据集中，与 CatBoost、ELM、SVM、BP、SWELM 和 Boosting 相比，无论训练集规模大小 RF-CatBoost 的分类精度均最高，CatBoost 和 Boosting 次之。分析认为集成学习能够将弱分类器集成在一起从而提高各弱分类器模型的非线性分析能力；与 CatBoost 相比，RF-CatBoost 分类精度较高，主要因 RF 能够将样品光谱数据中无效特征波长剔除从而筛选出最具样品属性特征的波长。此外 CatBoost 较 Boosting 分类精度更高，主要由于 CatBoost 利用对称树将类别特征组合在一起，丰富了各类别的特征维度。BP 分类精度最差，说明其非线性建模能力较差。ELM 和 SWELM 表现出了相当的分类精度且比 SVM 低，说明核函数几乎对 ELM 模型鉴别能力没有影响但其非线性建模比 SVM 差。

(2)运行时间

运行时间是对药品鉴别模型工作效率的重要衡量指标，运行时间越短说明药品鉴别模型的效率越高。表 4 给出了 RF-CatBoost、CatBoost、ELM、SVM、BP、SWELM 和 Boosting 在不同规模的 A、B 两组数据集上的运行时间。

表 3 各模型在不同规模的 A 和 B 两组数据集上的分类精度(%)

Table 3 Classification accuracy of each model on different sizes data sets in group A and B (%)

组别	训练/测试集	ELM	SWELM	SVM	BP	Boosting	CatBoost	CatBoost RF-CatBoost
A	40/176	92.36	92.33	93.68	89.36	94.44	94.99	96.79
	60/156	93.65	93.88	94.09	90.31	94.95	95.03	97.89
	80/136	94.99	95.11	95.35	91.01	96.22	96.85	98.82
	100/116	96.03	96.29	96.88	91.85	97.05	97.52	99.05
	120/96	97.88	97.95	97.23	92.99	97.99	97.89	99.95
	140/76	97.99	97.64	98.88	93.35	98.85	98.98	100
	160/56	98.05	98.01	99.05	94.99	99.01	99.02	100
	180/36	98.88	98.91	99.01	95.89	99.18	99.35	100
B	30/168	91.28	90.99	92.34	88.75	92.86	93.95	95.97
	50/148	92.69	91.67	93.38	90.31	94.01	94.98	96.59
	70/128	93.19	93.11	94.20	91.11	95.19	96.82	98.79
	90/108	94.25	94.26	95.38	91.85	96.21	96.98	99.92
	110/88	94.95	95.89	96.21	92.99	96.89	97.99	100
	130/68	95.92	97.22	98.09	93.35	97.96	98.09	100
	150/48	97.95	98.09	98.89	94.99	98.39	98.19	100
	170/28	98.88	98.85	99.00	95.89	99.08	98.51	100

表 4 各模型在不同规模的 A、B 两组数据集上的运行时间(s)

Table 4 Runningtime of each model on different sizes data sets in group A and group B (s)

组别	训练/测试集	ELM	SWELM	SVM	BP	Boosting	CatBoost	CatBoost RF-CatBoost
A	40/176	0.009 4	0.003 1	0.017 0	38.098 8	15.928 8	8.717 0	6.088 3
	60/156	0.013 0	0.005 0	0.032 1	38.339 7	17.985 4	15.903 0	7.408 3
	80/136	0.013 8	0.015 8	0.063 3	38.449 8	20.111 6	23.147 4	9.237 4
	100/116	0.021 1	0.017 1	0.113 8	38.864 7	22.273 0	30.441 0	9.365 9
	120/96	0.031 1	0.030 2	0.151 8	39.800 8	24.884 6	38.090 2	10.297 3
	140/76	0.044 2	0.040 1	0.219 8	40.885 0	26.690 6	45.062 6	10.988 8
	160/56	0.056 7	0.059 7	0.287 8	41.297 8	28.565 2	52.382 0	12.083 4
	180/36	0.078 7	0.074 0	0.359 5	42.380 7	30.661 2	59.526 4	12.530 8

续表 4

B	30/168	0.017 5	0.009 4	0.007 3	2.171 0	2.510 6	1.496 2	0.477 1
	50/148	0.053 5	0.023 7	0.024 3	4.159 9	3.322 7	2.395 7	1.164 1
	70/128	0.127 5	0.053 3	0.052 3	6.208 0	4.081 7	3.325 8	2.031 3
	90/108	0.215 5	0.098 3	0.102 7	8.350 6	5.336 2	4.422 0	2.907 4
	110/88	0.339 1	0.167 0	0.177 8	10.416 0	6.174 3	5.406 6	3.831 2
	130/68	0.495 8	0.260 2	0.295 3	12.404 0	6.887 9	6.411 8	4.785 0
	150/48	0.706 1	0.406 3	0.439 3	14.357 2	7.694 5	7.363 1	5.779 7
	170/28	0.912 0	0.616 8	0.643 0	16.322 8	8.565 5	8.346 4	6.829 1

由表 4 中可看出，RF-CatBoost、CatBoost、ELM、SVM、BP、SWELM 和 Boosting 随着训练样本数目的增加运行时间均逐步增加，且不论训练集样本的大小，BP 的运行时间最长，RF-CatBoost、CatBoost 和 Boosting 的运行时间次之，ELM、SVM 和 SWELM 的运行时间最短。分析认为由于 BP 神经网络采用多次循环迭代求解网络的最优参数实现网络的训练，因此延长了其运行时间；由于集成学习需要训练多个弱分类器实现最终网络的训练，所以造成 RF-CatBoost、CatBoost 和 Boosting 的运行时间比 ELM、SVM 和 SWELM 的运行时间长。此外，由于 ELM 和 SWELM 为只含有一个隐含层的网络且无需多次迭代寻优，故而缩短了网络的运行时间。

(3)模型稳定性

为了保证药品鉴别模型具有较强的应用稳定性，采用预测标准偏差(standard deviation，STD)对 RF-CatBoost、CatBoost、ELM、SVM、BP、SWELM 和 Boosting 的稳定性进行评估。各模型在 A、B 两组不同规模训练集上的 STD 如图 5 所示。

由图 5(a,b)中可看出，与 ELM、SVM、BP、SWELM 相比，在 A、B 两组不同规模训练集上无论训练集样本数目如何，RF-CatBoost、CatBoost 和 Boosting 均表现出了较低的 STD 且 RF-CatBoost 最低、CatBoost 次之、Boosting 最差。结果表明集成学习方法有利于提高决策树的稳定性，且在 RF-CatBoost、CatBoost 和 Boosting 这 3 个集成学习算法中 RF-CatBoost 的稳定性最强、Boosting 的稳定性最差。BP 比 ELM、SVM 和 SWELM 的 STD 较强，说明 BP 的稳定性较对比方法较差；与 ELM 相比，SWELM 在不同规模训练集上均表现出了较低的 STD，说明核函数对于 ELM 的稳定性会产生影响。

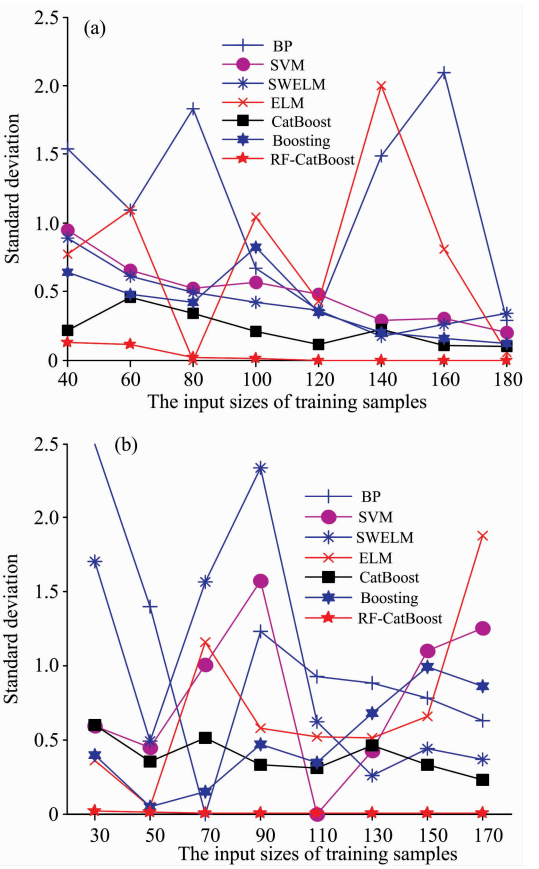

图 5　各模型在 A 组(a)和 B 组(b)不同规模训练集上的预测标准偏差

Fig. 5　Standard deviations of each model on different sizes data sets in group A (a) and group B (b)

4　结　论

采用近红外光谱分析技术实现了药品光谱信息的无损采集；采用多阶段预处理和白化处理消除了药品光谱数据中存在噪声和基线漂移等；采用随机森林能够准确地筛选出最能表征样品属性的特征波长并采用筛选的特征送入 CatBoost 实现了药品生产厂商的准确鉴别。以不同厂商生产的铝塑和非铝塑包装形式药品的光谱数据为例，构建了不同规模的训练集对 RF-CatBoost 的性能进行评估，并与 CatBoost、ELM、SVM、BP、SWELM 和 Boosting 模型进行对比，其中 RF-CatBoost 模型的分类精度最高达 100%且预测标准偏差趋于 0。结果表明 RF-CatBoost 在不同规模训练集上均表现出了最优的鉴别性能，可用于药品生产厂商的鉴别。

References

[1] CHU Xiao-li, CHEN Pu, LI Jing-yan, et al(褚小立, 陈 瀑, 李敬岩, 等). J. Instr. Anal.(分析测试学报), 2020, 39(10): 1181.
[2] Pavlek L R, Mueller C, Jebbia M R, et al. Front. Pediatr., 2021, 8: 624113.
[3] WANG Li-qun, LI Yu-yu, JIN Rong-jiang, et al(王丽群, 李雨谿, 金荣疆, 等). J. Tissue. Eng.(中国组织工程研究), 2021, 25(11): 1799.
[4] FU Dan-dan, WANG Qiao-hua, GAO Sheng, et al(付丹丹, 王巧华, 高 升, 等). Chin. J. Anal. Chem.(分析化学), 2020, 48(2): 289.
[5] CHU Xiao-li, YUAN Hong-fu, LU Wan-zhen(褚小立, 袁洪福, 陆婉珍). Chin. J. Anal. Chem.(分析化学), 2002, 30(1): 114.
[6] Siddiqui M R, Alothman Z A, Rahman N. Arab. J. Chen., 2017, 44(1): 1409.
[7] Huang Y, Meng S, Zhao P, et al. Appl. Optics, 2019, 58(18): 5122.
[8] Nguyen K, Duong D Q, Almeida F T, et al. J. Dent. Res., 2020, 99(1): 1054.
[9] Morellos A, Pantazi X E, Moshou D, et al. Biosyst. Eng., 2016, 152: 104.
[10] Clua P G, Jo E, Nikolic S, et al. J. Pharmaceut. Biomed., 2020, 183(8): 113163.
[11] Zheng A, Yang H, Pan X, et al. Sensors, 2021, 21(4): 1088.
[12] ZHOU Ying, LIU Jia-ming, LI Xiu-yun(周 颖, 刘佳明, 李秀芸). China Pharm.(中国药师), 2020, 23(1): 172.
[13] Rodionova Y, Titova A V, Balyklo K S. Talanta, 2019, 205: 120150.
[14] Sampaio P S, Castanho A, Almeida A S, et al. Eur. Food Res. Technol., 2020, 246(3): 527.
[15] Kim S Y, Hong S J, Kim E, et al. Appl. Eng. Agric., 2021, 37(4): 653.
[16] Nasir R, Saleem M R, Nisar A, et al. Optik, 2021, 225(11): 165714.
[17] CHEN Wen-li, WANG Qi-bin, LU Hao-xiang, et al(陈文丽, 王其滨, 路皓翔, 等). J. Instr. Anal.(分析测试学报), 2020, 39(10): 1267.
[18] SHEN Dong-xu, HONG Ming-jian, DONG Jia-lin(沈东旭, 洪明坚, 董家林). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2020, 40(11): 3457.
[19] Chen B, Wang Z B. Chemometr. Intell. Lab., 2019, 191: 103.
[20] Breiman L. Mach. Learn., 2001, 45(1): 5.
[21] Tang J, Fan B, Xiao L, et al. SPE Journal, 2020, 26(1): 482.
[22] Pinto P A, Dias A A, Fraga I, et al. Bioresour. Technol., 2012, 111: 261.

Drugs Identification Using Near-Infrared Spectroscopy Based on Random Forest and CatBoost

JIANG Ping[1], LU Hao-xiang[2], LIU Zhen-bing[2*]

1. School of Computer and Information Technology, Guangxi Police College, Nanning 530028, China
2. College of Computer and Information Security, Guilin University of Electronic Technology, Guilin 541004, China

Abstract Drug quality is related to people's health and national lifeblood. The rapid development of the economy and society plays an extremely important role in the rapid and effective identification of drug quality. Spectral analysis technology has high accuracy, fast analysis speed and no pollution to samples, and is widely used in the chemical industry, petroleum, medicine and other important areas of people's livelihood. In order to solve the problems of low accuracy, low identification speed and poor stability of the traditional drug identification model, the spectrometer was used to collect near-infrared spectroscopy data of drugs to achieve the purpose of pollution-free drugs. Then, random forest and CatBoost were combined to classify and identify drugs quickly and accurately. The proposed method firstly uses Random Forest (RF) to screen the effective characteristic wavelength of the spectrometer's spectral data to eliminate the irrelevant wavelength in the drug spectral data and screen out the characteristic wavelength that can best characterize the sample properties. Then Extreme Learning Machine (ELM) was used as CatBoost weak classifier to analyze the feature wavelengths of the screening for drug attribute identification. Since ELM only contains one hidden layer and no iterative optimization is required to ensure the faster running of the identification model, CatBoost can improve the model's identification accuracy by integrating a weak classifier. In order to effectively evaluate the performance of the drug identification model proposed in this paper, the spectral data of drugs of different sizes were constructed

by randomly selected training sets, and experiments were carried out independently. The mean value of 10 running results was taken as the final result. In addition, Back Propagation with CatBoost, Support Vector Machine (SVM), BP, ELM, Summation Wavelet Extreme Learning Machine (SWELM) and Boosting were compared to evaluate the performance of the proposed model further. As can be seen from the classification results of training sets of different sizes, with the increase of training sets, the highest classification accuracy is 100%, and the prediction standard deviation tends to be 0. The experimental results show that the RF-CATBoost identification model proposed in this paper has higher classification accuracy, faster speed and stronger robustness than the comparison method on drug data sets of different sizes and can be widely used in the accurate identification of drug categories, to achieve effective supervision of drug quality.

Keywords　Near-infrared spectroscopy; Random Forest; Extreme learning machine; CatBoost

(Received Jan. 12, 2022; accepted Mar. 28, 2022)

* Corresponding author

关于《光谱学与光谱分析》调整审稿费收费标准的通知

尊敬的《光谱学与光谱分析》广大作者、读者：我刊自 2018 年 7 月 1 日以后登记的稿件向投稿作者收取审稿费 200 元/篇，在您投稿之前，为免受经济损失，请您必须考虑：

1. 没有创新的一般性稿件，请您不要投稿。

2. 没有国家级基金资助的稿件，请您不要投稿。

3. 不是光谱专业的稿件，请您不要投稿。

4. 与其他文章重合率超过 10%的稿件，请您不要投稿。

所投稿件经初审通过后，作者会收到缴纳审稿费的通知。请作者及时从我刊网站(http://www.gpxygpfx.com)查询稿件是否处于交审稿费状态，在收到通知后，请及时缴纳审稿费；如在 10 天之内没有收到您的审稿费，被视为自动放弃，我刊不再受理。交费后我刊开据增值税电子普通发票，并传至作者提供的电子邮箱，作者可自行打印。

联系电话：010-62181070，62182998

电子邮箱：chngpxygpfx@vip.sina.com

感谢您多年来对《光谱学与光谱分析》的支持和厚爱！

《光谱学与光谱分析》期刊社

2018 年 6 月 30 日