谢文涌,柴琴琴,林 旋,等. 基于 Stacking 集成学习的马兜铃酸及其类似物鉴别[J].江苏农业学报,2021,37(2):503-508.
doi:10.3969/j.issn.1000-4440.2021.02.028

基于 Stacking 集成学习的马兜铃酸及其类似物鉴别

谢文涌[1,2], 柴琴琴[1,2], 林 旋[3], 李祥辉[3], 王 武[1,2]
(1.福州大学电气工程与自动化学院,福建 福州 350108; 2.福建省医疗器械和医药技术重点实验室,福建 福州 350108; 3.福建医科大学医学技术与工程学院,福建 福州 350004)

摘要: 以中草药中所含成分马兜铃酸及其类似物为研究对象,针对传统中药鉴定存在的主观性强、操作复杂等不足以及单一机器学习模型鉴别精度不高的问题,提出多模型融合的 Stacking 集成学习分类模型,用来实现马兜铃酸及其类似物的鉴别。采集马兜铃酸、1,10-菲咯啉-4,7-二甲酸、菲醌、β-谷甾醇 4 种样品的近红外光谱数据,对其进行数据预处理与主成分分析降维,基于降维后的数据特征,通过遍历搜索策略构建了以随机森林、支持向量机、朴素贝叶斯为基分类器,随机森林为元分类器的 Stacking 集成学习分类模型。结果表明,Stacking 集成学习分类模型具有最佳表现性能,鉴别正确率最高达到 99.38%,比 K 最近邻、决策树、随机森林、支持向量机、朴素贝叶斯分类模型的平均鉴别正确率高 8.23 个百分点,并且在精确率、召回率、综合评价指标(F_1 值)方面有优异表现。综上可见,本研究提出的 Stacking 集成学习分类模型能够快速有效地鉴别马兜铃酸及其类似物。

关键词: 马兜铃酸; 近红外光谱; 主成分分析; Stacking 集成学习
中图分类号: TP391 **文献标识码:** A **文章编号:** 1000-4440(2021)02-0503-06

Discrimination of aristolochic acid and its analogues based on stacking ensemble learning

XIE Wen-yong[1,2], CHAI Qin-qin[1,2], LIN Ni[3], LI Xiang-hui[3], WANG Wu[1,2]
(1.*College of Electrical Engineering and Automation, Fuzhou University, Fuzhou* 350108, *China*; 2.*Fujian Key Laboratory of Medical Instrument and Pharmaceutical Technology, Fuzhou* 350108, *China*; 3.*School of Medical Technology and Engineering, Fujian Medical University, Fuzhou* 350004, *China*)

Abstract: Aristolochic acid and its analogues contained in Chinese herbal medicine were taken as the research objects. Classification model based on Stacking ensemble-learning with multi-model fusion was proposed to identify aristolochic acid and its analogues, aiming at the shortcomings in traditional Chinese medicine identification such as strong subjectivity, complex operations and low accuracy of single classifier model. The near-infrared spectroscopy data of aristolochic acid, 1,10-phenanthroline-4,7-dicarboxylic acid, phenanthraquinone and β-sitosterol samples were collected. The data were preprocessed and principal component analysis was used to reduce dimensionality. Stacking ensemble-learning model was constructed through traversal search strategies based on the data features after dimensionality reduction, with random forest (RF), support vector machine (SVM), naive bayes (NB) as base classifiers and RF as meta classifier. The results showed that classification model based on Stacking ensemble-learning showed the best performance, with a discrimination accuracy rate of 99.38%, which was 8.23 percentage point higher than the average discrimination accuracy rate of classification models like K nearest, decision tree, RF, SVM and NB. Moreover, the proposed method showed excellent performance in precision, recall ratio and comprehensive evaluation index (F_1 score). Therefore, the method proposed in this study can quickly and effectively identify aristolochic acid and its analogues.

收稿日期: 2020-08-04
基金项目: 国家自然科学基金项目(61773124);晋江市福大科教园区发展中心科研项目(2019-JJFDKY-48)
作者简介: 谢文涌(1994-),男,福建三明人,硕士研究生,主要从事机器学习研究。(E-mail)1024396820@qq.com
通讯作者: 柴琴琴,(E-mail)qq.chai@fzu.edu.cn

Key words: aristolochic acid; near infrared spectroscopy; principal component analysis; Stacking ensemble-learning

关木通、青木香、马蹄香、细辛等马兜铃科的中草药具有调血压、抗肿瘤、抗菌、镇痛、消炎等作用,被广泛应用于临床医学中[1-2]。然而,这些中草药中含有马兜铃酸及其衍生物,长期服用有明显的肾毒性和致癌性[3]。近年来,国内外发生的多起马兜铃酸中毒事件引起了公众极大的关注。国际癌症研究中心已经将马兜铃酸列为第 1 类致癌物,多个国家也明确提出要禁止这类中草药的流通和使用[4]。中国中草药种类繁多,含有马兜铃酸的中成药就有 101 种[5],马兜铃酸及其类似物在结构上具有很大的相似性,如将马兜铃酸误识别成其类似物用于医学治疗将会造成难以估量的损失。因此,如何准确有效地区分马兜铃酸及其类似物,对于维护消费者权益和传承发展中医药具有重大意义。

常用的马兜铃酸检测方法主要有高效液相色谱法、薄层扫描法、荧光分析法、气相色谱分析法等,这些方法存在主观性强、灵敏度低、操作过程繁琐等缺陷[6-7],而近红外光谱(Near infrared spectroscopy, NIRS)检测技术[8]具有快速、无损且精度高等特点,因此将该技术与机器学习方法相结合,能有效对中草药进行定性和定量分析。然而,近红外光谱数据含有较多冗余信息,需要在建模前对其进行降维处理,以去除无关信息,常用的方法有主成分分析(Principal component analysis, PCA)法[9]、线性判别分析法[10]等。在中草药鉴别方面,传统的机器学习方法如 K 最近邻(K nearest neighbor, KNN)法[11]、支持向量机(Support vector machine, SVM)法[12]、决策树(Decision tree, DT)法[13]、极限学习机法[14]、朴素贝叶斯分类器(Naive bayes, NB)法[15]等均要求被测样本的数据集与训练样本的数据集分布一致,其假设的特征分布函数与实际情况最符合。然而,由于中草药产地、品种多样,在众多假设空间中找到一个与实际相符的函数作为分类函数十分困难。由此可见,以上单一分类器往往由于随机性而导致泛化性能不佳。

为了解决由样本不确定性带来的分布函数难以估计的问题,本研究提出了基于分层堆叠式的 Stacking 集成学习算法[16]。目前,集成学习已经成为机器学习的热门研究方向之一,集成学习包括并行化集成的 Bagging[17]、序列化集成的 Boosting[18]和堆叠式集成的 Stacking。其中 Stacking 集成学习由基分类器和元分类器组成,基分类器对原始数据进行训练预测,元分类器综合多个基分类器的输出特征作出最后决策。因此,Stacking 集成学习具有更高的模型准确性、鲁棒性和整体归纳能力。Stacking 集成学习在教育、医学、社会科学等方面的应用广泛[19-21]。然而,基分类器和元分类器的组合方式是构造 Stacking 集成学习的难重点,现有方法大多是对基分类器的多样性和分类正确率等指标进行权衡来获得“好而不同”的基分类器[22-23],但是多样性度量方法多样,难以找到最佳分类器组合。因此,本研究在基分类器的选择中设计了遍历搜索策略,以分类正确率为评价指标选择 Stacking 模型的最佳组合方式。

综上,鉴于目前鲜有关于近红外光谱技术结合集成学习方法对中草药进行分类鉴别的研究,本研究使用 PCA 法进行光谱数据降维,提出基于遍历搜索策略构建的两阶段 Stacking 集成学习模型,以期有效提高马兜铃酸及其类似物鉴别的精度,解决传统机器学习模型鉴别效果不佳的问题。

1 材料与方法

1.1 样品制备

本研究所用样本为关木通(马兜铃酸Ⅰ含量约为 0.05%)及其 3 种马兜铃酸类似物(分别为 1,10-菲咯啉-4,7-二甲酸、菲醌、β-谷甾醇)。其中,关木通样品采购于福建省福州市某药房,1,10-菲咯啉-4,7-二甲酸采购于上海毕得医药科技有限公司,菲醌采购于上海麦克林生化科技有限公司,β-谷甾醇采购于北京索莱宝科技有限公司,药材关木通经福建医科大学教授专业认证。将采购的关木通置于中药粉碎机中粉碎,过 60 目筛网,得到含马兜铃酸的粉末。将上述 4 种化合物分别与淀粉混合制备成 8 种质量浓度(1.3×10^{-3} mg/ml、1.2×10^{-3} mg/ml、1.1×10^{-3} mg/ml、1.0×10^{-3} mg/ml、0.9×10^{-3} mg/ml、0.8×10^{-3} mg/ml、0.7×10^{-3} mg/ml、0.6×10^{-3} mg/ml)的中药制剂样品,每种质量浓度的中药制剂制备 4 个样本。

1.2 数据采集与划分

本试验采用配有高灵敏度 InGaAs 检测器、积分球采样系统及内置自动金箔背景采集方式的 ANTARIS 型傅里叶变换近红外光谱分析仪(Thermo,德国)采集得到样品的近红外光谱集。光谱分辨率为

8 cm^{-1},光谱波长扫描范围为4 000~10 000 cm^{-1},平均扫描次数为 32 次。以采集的空白数据作为测量的背景数据来设置仪器的流程参数,在室温为 25 ℃、空气相对湿度为 60%的条件下测定样品的近红外光谱,每个样品采集 5 条光谱数据。最终,每类样品得到 160 组近红外光谱数据,总共有 640 组数据。采用随机划分样本数据集的方式,将原始数据集按照3∶1 的比例划分为训练集和预测集,具体划分情况如表 1 所示。

表 1　样本数据集的划分情况

Table 1　Division of sample data sets

样本名称	类别标签	样本总数(个)	训练集数量(个)	预测集数量(个)
马兜铃酸	AA	160	120	40
1,10-菲咯啉-4,7-二甲酸	A1	160	120	40
菲醌	A2	160	120	40
β-谷甾醇	A3	160	120	40
合计		640	480	160

1.3　数据分析方法

1.3.1　主成分分析(PCA)　PCA 是一种经典的无监督聚类算法,也是常用的数据降维与特征提取方法。该算法通过正交变换将高维线性相关变量投影至低维空间,由此获取线性不相关的新变量,即主成分。主成分能够反映原始数据的主要方差信息,并且去除了大量冗余特征,减少了计算的复杂度,有效避免了由维数灾难造成的模型过拟合现象。

1.3.2　Stacking 集成学习　Stacking 集成学习框架由 2 级分类器构成,第 1 层分类器称为基学习器,第 2 层分类器称为元学习器,其基本结构见图 1。

具体的训练过程如下:将原始数据集按照一定比例划分为训练集和预测集,训练集用于第 1 层分类模型训练,并将第 1 层中各个基分类器的输出特征作为第 2 层分类器的输入特征,预测集用于元分类器的预测,由元分类器输出最终预测结果。假定原始的数据集为 $L=\{(y_i,x_i),i=1,2,\cdots,N\}$,其中 y_i 为第 i 个样本的类别,x_i 为第 i 个样本的特征向量,N 为样本总数,p 为特征向量的数量,即 x_i 中包含 $x_1,x_2,\cdots,x_p$。按照 K 折交叉验证方法,将原始数据集 L 划分为 K 个大小相等的子集 $D_1,D_2,\cdots,D_K$,$\overline{D_K}=L-D_K$,其中 $\overline{D_K}$ 为交叉验证中的第 K 折训练集,D_K 为第 K 折预测集。设基分类器的数量为 n,每个分类器对于交叉验证中的第 K 折训练集进行训练和测试,对于预测集中的样本 x_i,基分类器的预测结果为 z_{ni},将每个基分类器的 K 次测试结果合并,与原始数据标签(y_i)一起构成元分类器的输入向量,即 $L_{new}=\{(y_i,z_{1i},z_{2i},\cdots,z_{ni}),i=1,2,\cdots,N\}$。元分类器通过学习新构成的数据特征,输出最终判别属性,以此来增强模型的泛化能力。

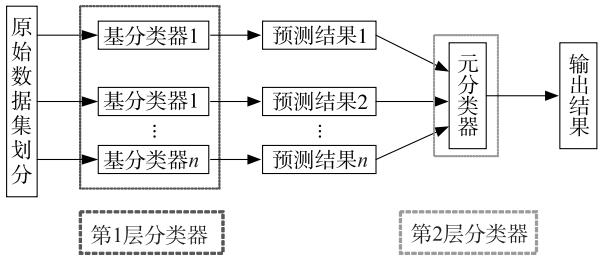


图 1　Stacking 算法结构

Fig.1　Structure diagram of the Stacking algorithm

1.4　模型评价指标

为更加直观准确地评价本研究构建的 Stacking 集成学习模型的性能,本研究提出结合鉴别正确率(A)、精确率(P)、召回率(R)和综合评价指标 F_1 值作为评价指标,计算公式如下:

$$A=\frac{\text{鉴别正确样本数}}{\text{总样本数}}\times 100\% \tag{1}$$

$$P=\frac{TP}{TP+FP} \tag{2}$$

$$R=\frac{TP}{TP+FN} \tag{3}$$

$$F_1=\frac{2PR}{P+R} \tag{4}$$

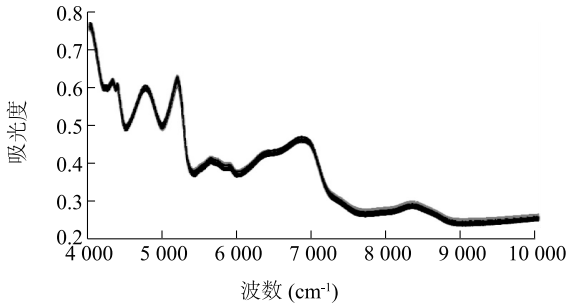
式中,TP 表示实际为正类的样本预测为正类的数量,FP 表示实际为负类的样本预测为正类的数量,FN 表示实际为正类的样本预测为负类的数量。

2　结果与分析

2.1　光谱分析

用 Pycharm 集成开发环境对所得样品的近红外光谱数据集进行分析,得到马兜铃酸及其类似物样品在4 000~10 000 cm^{-1}的原始近红外光谱(图 2)。可以看出,马兜铃酸、1,10-菲咯啉-4,7-二甲酸、菲醌、β-谷甾醇这 4 种化合物样品的近红外光谱相似度很高,

在多个波段出现交叉重叠情况,并在4 800 cm^{-1}、5 100 cm^{-1}、6 900 cm^{-1}附近有明显的吸收峰。其中,4 800 cm^{-1}附近的特征吸收峰为 C-H 的二级倍频与组合频,5 100 cm^{-1}附近的特征吸收峰为 C-H 的倍频,6 900 cm^{-1}附近的特征吸收峰为 O-H 或 N-H 的二级倍频。总体看出,这 4 种化合物全光谱吸收波段的形状与位置都十分相似,无法通过肉眼对其进行区分,需要进一步结合机器学习方法实现有效鉴别。



— 马兜铃酸; — 1,10-菲咯啉-4,7-二甲酸; ···· 菲醌; ····β-谷甾醇

图 2 马兜铃酸及其类似物的近红外光谱

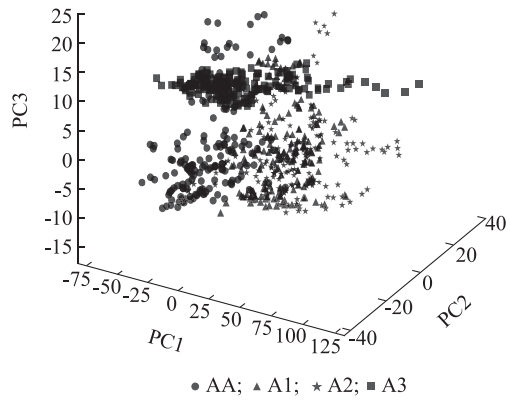
Fig.2 Near-infrared spectrum of aristolochic acid and its analogues

2.2 PCA 法降维处理

本试验采集的近红外光谱数据有1 557个特征值,且含有大量噪声,如果直接将其作为分类模型的输入向量,会增加模型的复杂度,降低模型的鉴别精度。因此,本研究在建立分类模型之前,先对原始光谱数据进行标准化处理以去除噪声干扰,再使用 PCA 法降低数据维度。经 PCA 法分析得出,前 3 个主成分的方差贡献率分别为 83.49%、10.97%、3.77%,累计贡献率达到 98.23%,足以解释原始数据信息。图 3 显示了前 3 个主成分(PC1、PC2、PC3)的得分分布,可以看出,这 4 种化合物相互混合,无法通过 PCA 聚类方法直接得到区分。因此,在 PCA 法降维的基础上,本试验进一步通过建立定性分析模型来实现马兜铃酸及其类似物的鉴别。

2.3 Stacking 集成学习分类模型分析

由于 Stacking 集成学习的元分类器训练集是由基分类器的输出产生的,如果直接对原始数据进行多折交叉验证,会导致元分类器与基分类器使用相同的数据集,从而造成严重的过拟合。因此,需要在原始数据集划分为训练集和预测集的基础上,再对训练集进行五折交叉验证。对于每个单一基学习器,依次使



AA、A1、A2、A3 设置见表 1。

图 3 马兜铃酸及其类似物的主成分得分

Fig.3 Principal component score of aristolochic acid and its analogues

用其中 4 个数据块作为训练子集,将对应的 1 个数据块作为验证子集。经过 5 次训练测试,将 5 次验证的子集合并,得到与原始训练集大小相同的新数据集,结合原始分类标签,一起作为元分类器的训练集;将 5 次预测集的结果取平均值,得到与原始预测集大小相同的新数据集,作为元分类器预测集。

此外,对于 Stacking 集成学习而言,基分类器和元分类器的组合是重点。本研究选择常见的 5 个异质分类器 KNN、DT、SVM、NB 和随机森林(Random forest,RF)[24]作为待选基分类器,以 RF 作为元分类器。为了提升模型的分类效果,本研究在待选基分类器的基础上,设计了遍历搜索策略,以鉴别正确率(A)作为评价指标,选择与 RF 结合且使正确率达到最大值的基分类器。

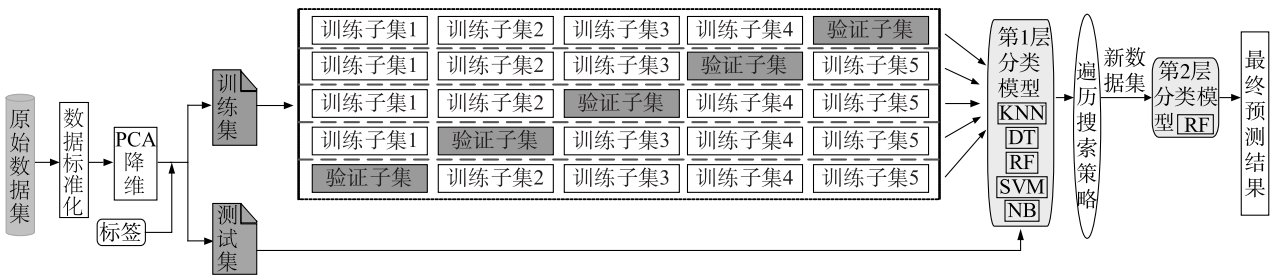
具体训练过程如图 4 所示,实现了原始数据输入特征至输出特征的变换,元分类器的训练集和预测集均未参与基分类器的训练过程,大大减少了过拟合风险,并且通过选择不同分类器组合,获取了更好的分类精度。不同分类器的具体组合结果见表 2。

表 2 不同基分类器的组合情况

Table 2 Combination of different base classifiers

组合方式	组合数量(个)	最高正确率(%)
5 选 2	10	98.13
5 选 3	10	99.38
5 选 4	5	98.75
5 选 5	1	96.25

5 选 2、5 选 3、5 选 4、5 选 5 指从 K 最近邻分类器、决策树分类器、支持向量机分类器、朴素贝叶斯分类器、随机森林分类器这 5 种分类器中随机选择 2 个、3 个、4 个、5 个进行组合。



PCA:主成分分析;KNN:K 最近邻;DT:决策树;RF:随机森林;SVM:支持向量机;NB:朴素贝叶斯。

图 4 Stacking 集成学习模型的马兜铃酸及其类似物鉴别

Fig.4 Aristolochic acid and its analogues identifification using Stacking ensemble-learning model

由表 2 可以看出,Stacking 集成学习通过不同基分类器与元分类器(RF)的组合,最高正确率均达 96.25%及以上。当基分类器组合方式为 5 选 3 时,与 RF 一起构成 Stacking 集成学习分类模型,此时的鉴别正确率达到最高值 99.38%,对应的基分类器为 RF、SVM、NB。为了充分验证所构建的 Stacking 集成学习分类模型的优越性,对于每个类别,使用精确率(P)、召回率(R)和综合评价指标 F_1 值进行度量评价。由表 3 可知,Stacking 集成学习分类模型具有良好的表现性能,在马兜铃酸、1,10-菲咯啉-4,7-二甲酸的鉴别中,精确率、召回率、综合评价指标 F_1 值均达到 97%及以上;在菲醌、β-谷甾醇的鉴别中,精确率、召回率、综合评价指标 F_1 值 3 个指标均达 100%。

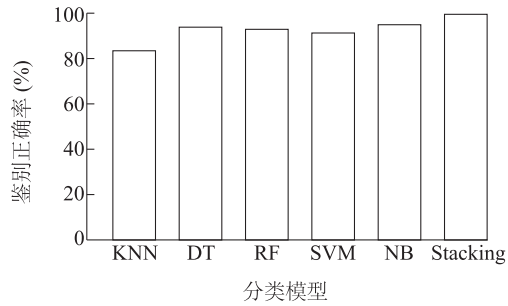
表 3 Stacking 集成学习分类模型评价结果

Table 3 Performance evaluation results of the Stacking ensemble-learning classification model

样品名称	精确率(%)	召回率(%)	综合评价指标 F_1 值(%)
马兜铃酸	100	97	99
1,10-菲咯啉-4,7-二甲酸	98	100	99
菲醌	100	100	100
β-谷甾醇	100	100	100

2.4 不同模型分类结果的对比

为了进一步验证本研究提出的 Stacking 集成学习分类模型的有效性,将其与使用相同训练集和预测集的单一分类模型进行对比。从图 5 可以看出,KNN、DT、RF、SVM、NB、Stacking 各分类模型的鉴别正确率分别为 83.13%、93.62%、92.88%、91.25%、94.87%、99.38%。与其他单一分类模型相比,Stacking 集成学习分类模型拥有最高的鉴别正确率。相比于单一分类模型 KNN,Stacking 集成学习分类模型的鉴别正确率提高了 16.25 个百分点,比各单一分类模型平均高 8.23 个百分点。



KNN:K 最近邻分类模型;DT:决策树分类模型;RF:随机森林分类模型;SVM:支持向量机分类模型;NB:朴素贝叶斯分类模型;Stacking:Stacking 集成学习分类模型。

图 5 各分类模型鉴别正确率的对比

Fig.5 Comparison of accuary of different classification models

以上结果表明,Stacking 集成学习分类模型综合了基分类器的优势,拥有比单一分类模型更好的表现性能,在一定程度上提升了鉴别正确率,能够更加有效地鉴别马兜铃酸及其类似物。这是因为单一分类模型在训练过程中可能陷入局部最优点,局部最优往往导致模型泛化性能不佳,而本研究提出的 Stacking 集成学习分类模型通过遍历搜索找到基分类器和元分类器的最佳组合方式,从而有效减少陷入局部最优点的风险。

3 结 论

本研究在近红外光谱技术的基础上,提出采用多模型融合的 Stacking 集成学习分类模型对马兜铃酸及其 3 种类似物进行鉴别。首先,对获取的光谱

数据进行数据标准化,使用 PCA 降维方法去除数据中含有的大量冗余信息,以此降低模型的复杂度。其次,为了充分学习数据特征,通过遍历搜索策略构建了以随机森林(RF)、支持向量机(SVM)、朴素贝叶斯(NB)为基分类器的 Stacking 集成学习分类模型。试验结果表明,Stacking 集成学习分类模型的鉴别正确率最高达 99.38%,优于单一分类模型 RF、SVM、NB、K 最近邻(KNN)和决策树(DT)。此外,Stacking 集成学习分类模型在精确率、召回率、综合评价指标 F_1 值方面均达 97%及以上,表现出优越性能。在今后的研究中,可以进一步研究改进 Stacking 集成学习分类模型,使用更多数据集验证其性能。

参考文献:

[1] HOLZBACH J C, NASCIMENTO I R, LOPES L M X. Phenylethylpyranone and aristolochic acid derivatives from *Aristolochia urupaensis*[J]. Journal of the Brazilian Chemical Society, 2017, 28(11): 2275-2279.

[2] JIN K, SU K K, LI T, et al. Hepatic premalignant alterations triggered by human nephrotoxin aristolochic acid Ⅰ in canines[J]. Cancer Prevention Research, 2016, 9(4): 324-334.

[3] 薛寿征,曾广先. 马兜铃酸肾病:研究及启示[J]. 科学(上海), 2018, 70(4): 27-31.

[4] 柏兆方,王春宇,王伽伯,等. 马兜铃酸与肝癌相关性的研究及思考[J]. 世界科学技术:中医药现代化, 2019,21(7): 1275-1279.

[5] 宋亚刚,苗艳艳,苗明三. 含马兜铃酸中药毒性分析[J]. 中华中医药杂志, 2018, 33(5): 1950-1954.

[6] 章 莹,肖 榕,黄 杰,等. 不同产地马兜铃蜜炙前后 HPLC 指纹图谱分析[J]. 中国药学杂志, 2017, 52(16): 1397-1402.

[7] 刘欣欣,王 莉,肖红斌. 不同产地马兜铃药材中马兜铃总酸的含量[J]. 时珍国医国药, 2017,28(1):74-76.

[8] LIN W Q, CHAI Q Q, WANG W, et al. A novel method for geographical origin identification of *Tetrastigma hemsleyanum* (Sanyeqing) by near-infrared spectroscopy[J]. Analytical Methods, 2018, 10(25): 2980-2988.

[9] MORAIS C L M, LIMA K M G. Principal component analysis with linear and quadratic discriminant analysis for identification of cancer samples based on mass spectrometry[J]. Journal of the Brazilian Chemical Society, 2018, 29: 472-481.

[10] LI C N, SHAO Y H, YIN W T, et al. Robust and sparse linear discriminant analysis via an alternating direction method of multipliers[J]. IEEE Transactions on Neural Networks and Learning Systems, 2019, 31(3): 915-926.

[11] CHEN Y W, HU X L, FAN W T, et al. Fast density peak clustering for large scale data based on kNN[J]. Knowledge-Based Systems, 2020, 2020(187): 104824.

[12] 马 娜,李艳文,徐 苗. 基于改进 SVM 算法的植物叶片分类研究[J]. 山西农业大学学报(自然科学版), 2018, 38(11): 33-38.

[13] 张晓忆,李卫国,景元书,等. 多种光谱指标构建决策树的水稻种植面积提取[J]. 江苏农业学报, 2016, 32(5): 1066-1072.

[14] 唐云峰,柴琴琴,林双杰,等. 可见/近红外光谱的葡萄籽油掺伪检测系统[J]. 光谱学与光谱分析, 2020, 40(1): 202-208.

[15] 陈 曦,张 坤. 一种基于树增强朴素贝叶斯的分类器学习方法[J]. 电子与信息学报, 2019, 41(8): 2001-2008.

[16] 袁培森,杨承林,宋玉红,等. 基于 Stacking 集成学习的水稻表型组学实体分类研究[J]. 农业机械学报, 2019, 50(11):144-152.

[17] ANDIOJAYA A, DEMIRHAN H. A bagging algorithm for the imputation of missing values in time series[J]. Expert Systems with Application, 2019, 129(9): 10-26.

[18] WANG B Y, PINEAU J. Online bagging and boosting for imbalanced data streams[J]. IEEE Transactions on Knowledge and Data Engineering, 2016, 28(12): 3353-3366.

[19] ELAYIDOM S, IDIKKULA S M, ALEXANDER J. A hybrid stacking ensemble framwork for employment predicyion problems[J]. Advances in Computational Research, 2011, 3(1): 25-30.

[20] DINAKAR K, WEINSTEIN E, LIEBERMAN H, et al. Stacked generalization learning to analyze teenage distress[C]//Association for the Advancement of Artificial Intelligence. Eighth International AAAI Conference on Weblogs and Social Media. Ann Arbor, Michigan, USA:Association for the Advancement of Artificial Intelligence,2014.

[21] HADDAD B M, YANG S, KARAM L J, et al. Multifeature, sparse-based approach for defects detection and classification in semiconductor units[J]. IEEE Transactions on Automation Science and Engineering, 2016, 15(1): 145-159.

[22] 孙 博,王建东,陈海燕,等. 集成学习中的多样性度量[J]. 控制与决策, 2014, 29(3): 385-394.

[23] 章 宁,陈 钦. 基于 AUC 及 Q 统计值的集成学习训练方法[J]. 计算机应用, 2019, 39(4):935-939.

[24] GUI L, XIA Y, LI H, et al. Prediction of NOX emission from coal-fired boiler based on RF-GBDT[C]//KIM YH. 2017 6th International Conference on Energy and Environmental Protection. Zhuhai, China:KIM YH, 2017.

(责任编辑:徐 艳)