基于二分搜索结合修剪随机森林的特征选择算法在近红外光谱分类中的应用

刘 明[1]，李忠任[2]，张海涛[2]，于春霞[2]，唐兴宏[2]，丁香乾[1]

[1]中国海洋大学信息科学与工程学院，山东 青岛 266100；

[2]云南中烟工业有限责任公司技术中心，云南 昆明 650024

摘要 针对随机森林(RF)在高维空间特征选择过程中计算繁琐和内存开销大、分类准确率低等问题，提出了基于二分搜索(BS)结合修剪随机森林(RFP)的特征选择算法(BSRFP)；该算法首先根据纯度基尼指数获取特征重要性评分，删除重要性评分较低的特征，然后利用BS算法结合基分类器差异性的修剪技术得到最优特征子集和最高分类准确率的分类器；为了验证算法的有效性，构建卷烟质量识别模型并与其他方法进行比较。结果表明：BS算法简化了特征搜索过程，RFP算法缩减了RF算法的规模；RFP算法的分类准确率可达96.47%；BSRFP算法选择出的特征相关性更强，对卷烟质量识别具有更高的准确度。

关键词 光谱学；特征选择；修剪随机森林；分类；纯度基尼指数；近红外光谱

中图分类号 O433.4 **文献标识码** A

doi：10.3788/LOP54.103001

Feature Selection Algorithm Application in Near-Infrared Spectroscopy Classification Based on Binary Search Combined with Random Forest Pruning

Liu Ming[1], Li Zhongren[2], Zhang Haitao[2], Yu Chunxia[2], Tang Xinghong[2], Ding Xiangqian[1]

[1] *College of Information Science and Engineering, Ocean University of China, Qingdao, Shandong* 266100, *China*;

[2] *Technical Research Center, China Tobacco Yunnan Industrial Co., Ltd., Kunming, Yunnan* 650024, *China*

Abstract In view of the problems of the random forest in the feature selection process in high-dimensional spaces, such as calculation complexity, large model memory overhead, and low classification accuracy, a feature selection algorithm named binary search random forest pruning (BSRFP) is proposed. This algorithm firstly obtains the feature importance scores according to the purity Gini index, and deletes features with low importance scores. The optimal feature subset and the classifier with the highest classification accuracy are then obtained with utilization of the pruning technique combining binary search with the diversity among base classifiers. To verify the effectiveness of this algorithm, a cigarette quality recognition model is established and compared with other methods. The results show that the binary search algorithm simplifies the feature search process, and the RFP algorithm reduces the size of random forest algorithm. The classification accuracy of the random forest pruning algorithm is 96.47%. The features selected by using BSRFP algorithm are more correlated, and the algorithm provides higher accuracy of cigarette quality recognition.

Key words spectroscopy; feature selection; random forest pruning; classification; purity Gini index; near-infrared spectrum

OCIS codes 300.6340; 070.4790; 070.5010

收稿日期：2017-04-26；**收到修改稿日期**：2017-05-31

基金项目：国家科技支撑计划(2015BAF12B01)、云南中烟工业有限责任公司项目(JSZX2014YL01，2053000102015200086)

作者简介：刘 明(1992—)，男，硕士研究生，主要从事近红外光谱学、数据挖掘、机器学习方面的研究。E-mail：lium.prc@gmail.com

导师简介：丁香乾(1962—)，男，博士，教授，博士生导师，主要从事机器学习、智能信息系统、工业大数据方面的研究。E-mail：dingxq1995@vip.sina.com(通信联系人)

1 引 言

近红外光谱分析技术是一种简单、高效、低成本的物理测量技术,已广泛应用于化工、制药、农业、烟草等领域[1-4]。近红外光谱包含了大量反映样品特征的信息,可用于物质的定量和定性分析[5]。但近红外光谱数据的高维、高噪特点使得全光谱建模分析计算过程异常繁琐,且样本的分类准确率较低[6]。因此,如何从近红外光谱中筛选与分类相关的特征尤为重要。

随机森林(RF)算法是一种基于分类回归树(CART)的集成学习算法,它可综合各决策树的投票情况对样本类别进行判定,同时还可对属性的重要性进行度量,广泛应用于高维特征选择和分类研究中[7-8]。杨珺雯等[9]基于 RF 后向特征消除(RF-RFE)法计算了特征的重要性,并对特征进行排序,通过与支持向量机(SVM)对比,证明了 RF 分类器是一种较好的高维光谱数据分类器。姚登举等[10]提出了一种基于 RF 的封装式特征选择(RFFS)算法,该算法在特征数量不是非常大的情况下具有较高的分类准确率。许勇刚等[11]基于分类间隔加权对 RF 进行修剪,减小了 RF 的规模,避免了因存储过多的基分类器而开销大量内存的问题,提高了分类准确率。虽然上述方法均取得了一定效果,但在消除低重要性分值的特征时,均采用了逐一删除评分最低的特征的方法,并且在高维空间中未将特征选择与 RF 修剪过程进行有效结合,从而未从根本上对算法的计算成本进行有效控制。

针对上述问题,本文提出了一种结合二分搜索(BS)算法和修剪随机森林(RFP)算法的特征选择算法——BSRFP 算法。该算法首先采用纯度基尼指数对特征变量的重要性进行评分,删除评分较低的特征,然后结合 BS 算法和 RFP 算法搜索最优特征子集,以获得最优特征子集及其对应的 RFP 模型,解决了高维空间中特征选择计算成本高、模型分类准确率低等问题。

2 算 法

2.1 RF 算法

RF 算法是由 Breiman 提出的基于多个 CART 的集成学习算法[12-13],该算法利用 Bootstrap 重采样法生成训练集,根据基尼指数最小原则对属性进行度量,逐步建立 CART,然后综合各决策树的投票情况判别样本的类别归属。同时,将未在训练集中出现的样本作为“袋外数据”来预测算法的准确度。RF 算法可以对属性的重要性进行度量,并根据投票机制确定样本的类别。

基尼指数是一种基于杂度的属性分裂方法。杂度与可获取的有用信息呈负相关。杂度越小,变量的离散程度越差,得到的信息量越大[14-15]。杂度基尼指数 G 的计算公式[14]为

$$G(a)=1-\sum_{i=1}^{c}p_i^2, \tag{1}$$

式中 c 为样本类别数,p_i 为特征 f 中属性 a 对应的样本属于类别 C_i(C_i 表示第 i 类,$i=1,2,\cdots,c$)的概率。

2.2 纯度基尼指数

由于杂度基尼指数与可用信息呈负相关,因此,为了更直观地体现特征项对分类效果的影响,本研究采用纯度基尼指数将纯度与可获得的有用信息转换为正相关,即

$$G_{\text{purity}}(a)=\sum_{i=1}^{c}p_i^2。 \tag{2}$$

特征 f 的纯度基尼指数为

$$G(f)=\sum_{l=1}^{K}\frac{n_l}{N}G_{\text{purity}}(a_l), \tag{3}$$

式中 N 为样本数,K 为某属性 a 的类别个数,a_l 为某特定类别的属性,n_l 为某特定类别属性对应的样本数。特征的纯度越大,表明该特征对样本的识别能力越强。特征的重要性评分为

$$S(v)=\frac{1}{m'}\sum_{u=1}^{m'}G(f_{uv}), \tag{4}$$

式中 m' 为 RF 中训练集的个数;$G(f_{uv})$ 为第 u 个训练集中第 v 维特征的向量纯度,$v=1,2,\cdots,k$,k 为样本的维度。

将特征重要性集合中的元素降序排列，设置阈值 α，删除特征重要性评分低于 α 的特征，获得与样本识别具有一定相关性的特征集合。

2.3 RFP 算法

RF 算法在执行过程中，基分类器的存储和投票耗费大量的计算成本，尤其是内存开销[16]。由此本研究提出了基分类器间相似度差异性的修剪方法，删除相似的基分类器，构造具有较强差异性的基分类器集合。RFP 算法和步骤如下：

1）采用 Bagging 策略生成 RF 训练模型。根据 RF 算法对样本的分类结果，构造 $n\times m$ 型基分类器相似度差异性矩阵 $\boldsymbol{S}_{nm}=\begin{bmatrix} s_{11} & \cdots & s_{1m} \\ \vdots & & \vdots \\ s_{n1} & \cdots & s_{nm} \end{bmatrix}$，$s_{nm}$ 为在第 m 个决策树中与第 n 个样本被分到同一个叶节点的样本数。矩阵 S_{nm} 的每一列代表某个决策树对样本的分类效果。

2）将基分类器相似度差异性矩阵改为列向量的形式 $\boldsymbol{S}=(\boldsymbol{S}_1,\boldsymbol{S}_2,\cdots,\boldsymbol{S}_m)$。此时，基分类器相似度差异性度量转化为列向量之间差异性度量。采用欧氏距离法计算各列向量之间的距离 $D_{xy}=\sqrt{\sum_{d=1}^{n}(s_{dx}-s_{dy})^2}$，其中 $x=1,2,\cdots,m-1$；$y=x+1$。得到列向量之间的距离矩阵 $\boldsymbol{D}=\begin{bmatrix} 0 & D_{12} & \cdots & D_{1m} \\ 0 & 0 & \cdots & D_{2m} \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}$，矩阵 $\boldsymbol{D}$ 是一个主对角线及以下元素均为 0 的矩阵。

3）考虑到矩阵 $\boldsymbol{D}$ 中与基分类器 T_x 和 T_y 有关的变量有 $\{D_{x,x+1},D_{x,x+2},\cdots,D_{xm}\}$ 和 $\{D_{y,y+1},D_{y,y+2},\cdots,D_{ym}\}$，则删除基分类器 T_x 和 T_y 后，列向量之间距离的改变量 $\Delta d=\sum_{p=x+1}^{m}D_{xp}+\sum_{q=y+1}^{m}D_{yq}$，对 Δd 降序排列，删除 Δd 的最大值所对应的 D_{xy}，并将其对应的基分类器 T_x 和 T_y 加入最优基分类器集合 T 中，计算此时集合 T 中基分类器组成的 RF 对样本的分类准确率。

4）重复步骤 3），直到 $D=$null，选取分类准确率最高的基分类器集合 T 作为最优基分类器集合。

2.4 BSRFP 算法

BSRFP 算法流程如图 1 所示。BSRFP 算法主要为删除无关变量的过程以及利用 BS 算法查找最优特征子集并对节点所包含的特征构造 RFP 的过程。BSRFP 算法的具体方法和步骤如下：

1）利用重采样方法构造 t 个训练集，记为 $\boldsymbol{F}=\{\boldsymbol{F}_1,\boldsymbol{F}_2,\cdots,\boldsymbol{F}_t\}$，每个训练集含 n' 个样本和 k' 个特征，$\boldsymbol{F}_w=\begin{bmatrix} f_{11} & \cdots & f_{1k'} \\ \vdots & & \vdots \\ f_{n'1} & \cdots & f_{n'k'} \end{bmatrix}=(\boldsymbol{f}_1,\boldsymbol{f}_2,\cdots,\boldsymbol{f}_{k'})$，$w=1,2,\cdots,t$。计算 $G(\boldsymbol{F}_w)=\{G(\boldsymbol{f}_1),G(\boldsymbol{f}_2),\cdots,G(\boldsymbol{f}_{k'})\}$，然后根据(4)式获得特征重要性评分集合，将集合中的评分结果进行降序排列。根据设置的阈值，删除评分小于该阈值的特征变量，得到长度为 l_h 特征集合 $\boldsymbol{F}'$。

2）将特征集合 $\boldsymbol{F}'$作为 BSRFP 算法的输入，根据排列好的特征构造二叉树。

BS 算法的步骤如下：

1）二叉树的根节点 R 包含了集合 $\boldsymbol{F}'$的 r 个特征，令 $N_{\min}$指向 $\boldsymbol{f}_1$，初始值为 1；$N_{\max}$指向 $\boldsymbol{f}_r$，初始值为 r；N_{index}表示中间变量，初始值为 1。根据 2.3 节所述方法，由根节点所包含的特征集合构造 RFP 模型，获取最优基分类器集合，并记录修剪后的 RF 分类准确率 a_1。

2）N_{mid}表示中间特征，令 $N_{\text{mid}}=N_{\text{mid}}+(N_{\max}-N_{\text{mid}})/2$。根节点 R 左子树的根节点为 R_1，右子树的根节点为 R_2，R_1 的左节点为 R_{11}，右节点为 R_{12}；同理，R_2 的左节点为 R_{21}，右节点为 R_{22}；以此类推。R_1 的特征集合为$\{\boldsymbol{f}_{\min},\cdots,\boldsymbol{f}_{\text{mid}}\}$，$R_2$ 的特征集合为$\{\boldsymbol{f}_{\text{mid}+1},\cdots,\boldsymbol{f}_{\max}\}$。可以看出，左节点的特征重要性明显大于右节点的特征重要性。计算由节点 R_1 的特征集合建立的 RFP 分类准确率 a_2。

3）若 $a_2>a_1$，则更新当前的精度，令 $a_1=a_2$；若 $N_{\max}>N_{\text{mid}}+1$，令 $N_{\max}=N_{\text{mid}}$。若 $a_2\leqslant a_1$ 且

$N_{max}>N_{mid}+1$，则令 $N_{index}=N_{mid}$，$N_{mid}=N_{mid}+(N_{max}-N_{mid})/2$，$R_{21}$ 节点的特征集合为 $\{\boldsymbol{f}_{min},\cdots,\boldsymbol{f}_{mid}\}$，$R_{22}$ 节点的特征集合为 $\{\boldsymbol{f}_{mid+1},\cdots,\boldsymbol{f}_{max}\}$。

4) 在二叉树上循环步骤 2)和步骤 3)，直到 $N_{max}=N_{mid}+1$ 时，算法停止。

由此可以得到具有最高分类准确率的最优特征子集及其对应的最优基分类器组合。

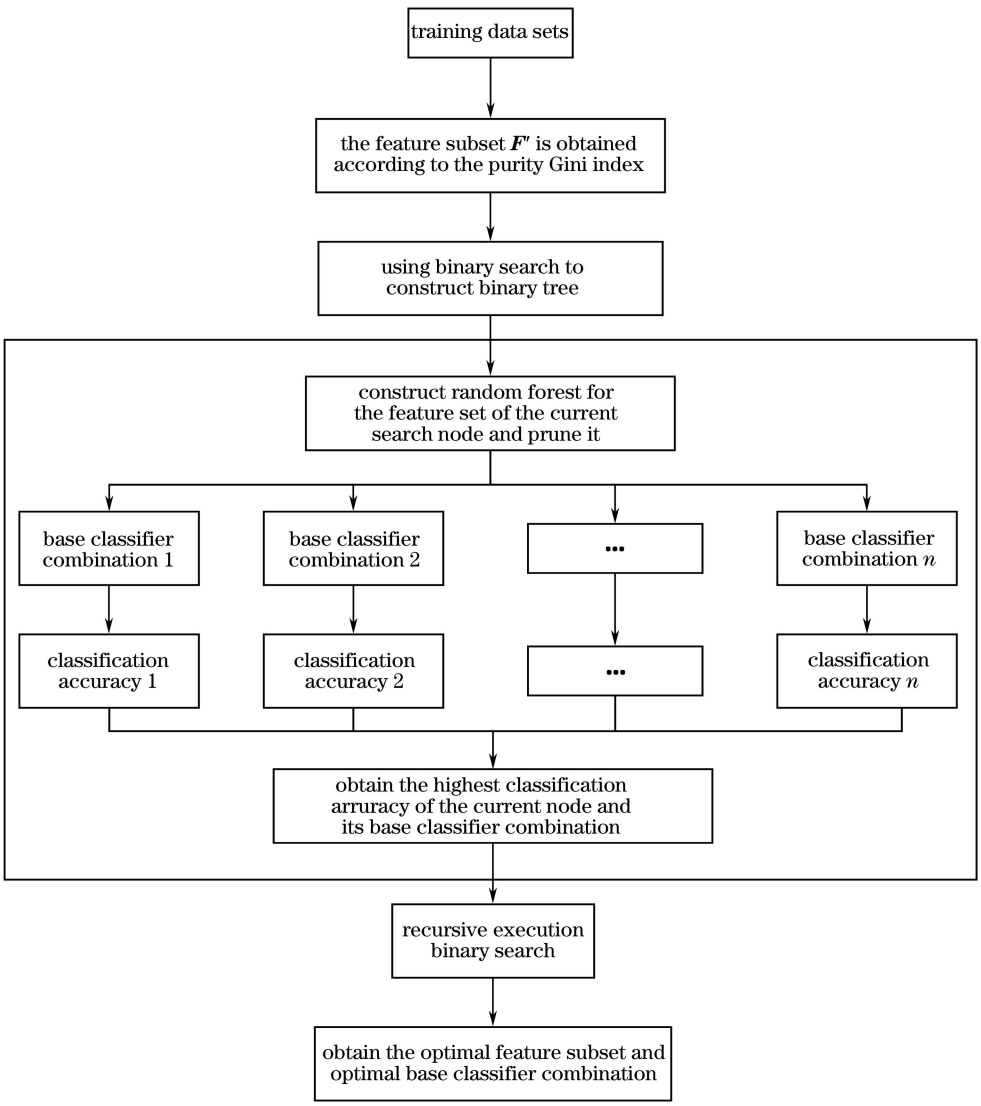

图 1 BSRFP 算法流程图

Fig. 1 Flow chart of BSRFP algorithm

3 实 验

3.1 数据来源及样品制备

实验材料购自市售各品牌卷烟，按照价格进行等级划分。每个样品购买 2 盒进行制样，共制作 5 档不同等级的卷烟样品，样品共 300 个，每档卷烟样品 60 个，随机选取其中的 40 个作为训练样本，剩余的 20 个作为测试样本。将每支烟剥开取出烟丝并置于烘箱中，在 40 ℃下干燥 2 h，磨碎后过 40 目(1 目=354 μm)筛，在(20±2) ℃下密封避光储存。为了取得比较精确的实验结果，取 20 次实验结果的平均值。

3.2 近红外光谱采集

采用 Antaris Ⅱ型傅里叶近红外光谱仪采集烟叶样本的漫反射光谱，扫描范围为 4000～10000 cm^{-1}，分辨率为 8 cm^{-1}，共 1557 个光谱数据点。将 15 g 样品置于洁净的样品杯中，放置在光谱仪上进行扫描。每个样品重复装样、扫描 3 次，然后计算其均值，将均值作为样本光谱数据。所有样本的原始近红外光谱图如图 2 所示。

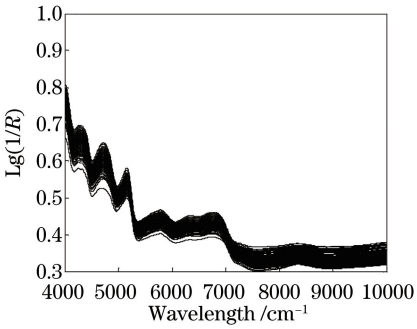

图 2 烟叶样本的原始近红外光谱图

Fig. 2 Original near-infrared spectra of cigarette samples

3.3 数据预处理

为了消除仪器误差、环境因素、基线漂移等干扰信息对光谱数据的影响，采用标准正态变量变换(SNV)并求二阶导数的方法作为预处理方法，使用 Unscrambler 9.7 软件对原始光谱进行预处理，实验建模使用 MATLAB R2014a 软件。

4 结果与分析

4.1 特征选择

预处理后的全光谱中含有 1539 个特征。利用 Bootstrap 重采样法构建含有 2000 个决策树的 RF。由(4)式计算特征重要性评分，重复实验 10 次，求平均值，得到所有特征的综合评分。特征重要性度量结果如图 3 所示。由图 3 可知，不同特征对卷烟质量等级分类的重要性不同，大多数特征的重要性为 0～0.05，特征重要性较强的特征评分高于 0.05。

根据特征重要性评分删除对卷烟质量等级不重要的特征，简化模型。将所有元素按特征重要性评分进行降序排列。调节特征重要性评分的阈值，得到对应的特征子集以及分类模型的准确度，并进行综合比较，将阈值设置为 $\alpha=0.02$，删除特征重要性评分小于 α 的特征，得到长度为 342 的特征子集 $\boldsymbol{F}'$。经过 BSRFP 算法的第 1)步，删除全光谱中的无关特征，得到与质量识别有一定相关性的特征集合；而特征集合 $\boldsymbol{F}'$中仍含有相关性较弱的特征，为了提高分类准确率，需要获得最优特征子集。图 4 所示为 BSRFP 算法所选的最优特征子集。由图 4 可知，BSRFP 算法所选特征个数为 69，明显少于原特征个数 1539 和纯度基尼指数重要性度量方法所选的特征个数 342。

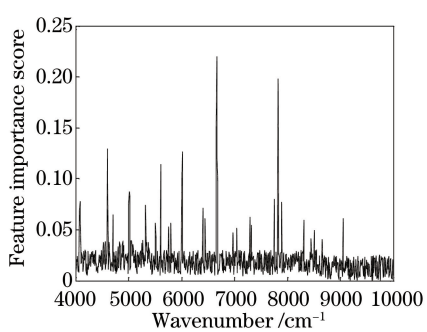

图 3 特征重要性的度量结果

Fig. 3 Measurement result of feature importance

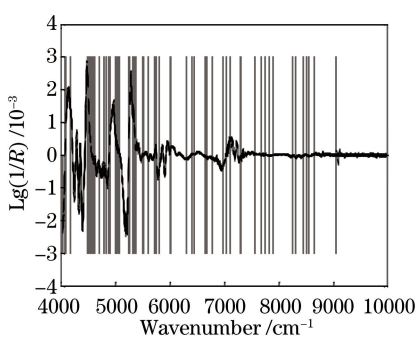

图 4 BSRFP 算法的选择结果

Fig. 4 The selection result of BSRFP algorithm

4.2 RFP 与 RF 算法的性能对比

图 5 所示为基于 BSRFP 算法所选的最优特征子集，采用 RFP 和 RF 算法的决策树数量(基分类器数量)与测试集分类准确率的对比。由图 5 可知，当分类准确率达到最高时，RFP 和 RF 基分类器数量分别为 87 和 131，分类准确率分别为 96.47%和 92.35%。与未修剪的 RF 相比，基于基分类器相似度差异性的修剪

技术不仅使基分类器的规模缩减了 44 个，而且使分类准确率提高了 4.12%。此外，随着基分类器数量增加，RF 的分类准确率逐步升高，越来越多差异性较大的基分类器被组合成 RF，分类能力逐步提高；然而当基分类器数量达到 87 之后，分类准确率开始下降，原因是随着 RF 规模增加，基分类器之间开始出现冗余，削弱了 RF 的分类性能。

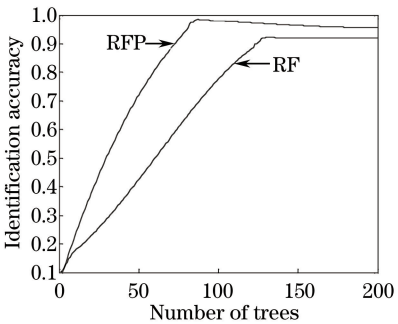

图 5 RFP 与 RF 算法的性能对比

Fig. 5 Performance comparison of RFP and RF algorithms

4.3 分类性能对比

为了进一步验证 BSRFP 算法的性能，分别采用预处理后的全光谱特征、纯度基尼筛选的特征、遗传算法(GA)选择的特征、无信息变量消除(UVE)算法选择的特征和 BSRFP 所选特征作为 RF 分类器的输入来建立分类模型，各算法选择的特征数量和分类准确率如表 1 所示。由表 1 可知，经过特征选择后建立的各分类模型的分类准确率均比基于全光谱特征建立的分类模型高，这表明在近红外光谱分类中特征选择是有必要的。全光谱特征中含有大量的噪声、无关信息和冗余信息，这些都会降低模型的分类准确率。与 GA、UVE 算法相比，基于 BSRFP 算法选出的特征所建立的分类模型对卷烟质量等级的分类准确率较高，并且所选特征数量最少，这说明 BSRFP 算法选出的特征与卷烟质量等级的相关性很高。

表 1 不同特征选择算法的特征数量与分类准确率

Table 1 Feature number and classification accuracy of different feature selection algorithms

Feature selection algorithm	Feature number	Classification accuracy
All features	1539	0.7242
Purity Gini	342	0.8601
GA	107	0.8967
UVE	176	0.8833
BSRFP	69	0.9235

为了验证 RFP 算法的有效性，将 RFP 分类算法与传统 RF 算法、K 最近邻(KNN)算法、最小二乘支持向量机(LS-SVM)算法、反向传输(BP)神经网络算法进行对比，以 BSRFP 算法选择的 69 个特征变量作为各分类算法的输入，各分类算法模型的分类准确率如表 2 所示。由表 2 可知，RFP 分类算法具有最高的分类准确率，并且分类准确率明显高于 RF、KNN、BP 神经网络、LS-SVM 算法的分类准确率，证明了本文所提 RFP 算法的有效性。传统的 RF 算法和本文所提 RFP 算法的分类准确率高于 KNN、LS-SVM、BP 神经网络算法的分类准确率，这表明 RF 算法对近红外光谱分类具有一定的优势。与传统 RF 分类算法相比，本文所提 RFP 分类算法不仅有较高的分类准确率，而且模型更简单，占用存储空间更少。

表 2 不同分类算法的分类准确率

Table 2 Classification accuracy of different classification algorithms

Classification algorithm	KNN	BP	LS-SVM	RF	RFP
Classification accuracy	0.8460	0.8870	0.9023	0.9235	0.9647

总体看来，BSRFP 算法解决了逐一删除评分最低特征来获取最优特征子集方法计算异常繁琐的问题，同时也解决了因基分类器数量过多导致的算法消耗大量内存的问题。上述结果表明，本文所提方法建立的模型不仅具有较高的分类准确率，而且模型更简单。

5 结 论

针对高维空间中特征选择计算成本高和分类准确率低的问题,提出了 BSRFP 算法。在删除无关特征的过程中,采用纯度基尼指数更直观地对特征重要性进行了度量。利用 BS 算法搜索效率高和 RFP 算法内存开销小的优势,提高了 BSRFP 算法的执行效率。仿真实验结果表明,所提算法具有更低的计算成本和更高的分类准确率,有望成为卷烟质量评估领域的一种有效解决方案。如何结合更多领域的实际应用,进一步改进分类算法并提高分类性能是下一步研究的方向。

参 考 文 献

[1] Sun Tong, Wu Yiqing, Li Xiaozhen, *et al*. Discrimination of camellia oil adulteration by NIR spectra and subwindow permutation analysis[J]. Acta Optica Sinica, 2015, 35(6): 0630005.
孙通, 吴宜青, 李晓珍, 等. 基于近红外光谱和子窗口重排分析的山茶油掺假检测[J]. 光学学报, 2015, 35(6): 0630005.

[2] Liu Wei, Chang Qingrui, Guo Man, *et al*. Detection of leaf nitrogen content of summer corn using visible/near infrared spectra[J]. Journal of Infrared and Millimeter Waves, 2011, 30(1): 48-54.
刘炜, 常庆瑞, 郭曼, 等. 夏玉米可见/近红外光小波主成分提取与氮素含量神经网络检测[J]. 红外与毫米波学报, 2011, 30(1): 48-54.

[3] Chen Xiaofeng, Long Changjiang, Niu Zhiyou, *et al*. Classification research of Chinese medicine based on latent semantic analysis and NIR[J]. Acta Optica Sinica, 2014, 34(9): 0930001.
陈晓峰, 龙长江, 牛智有, 等. 基于潜在语义分析与 NIR 的中药材分类研究[J]. 光学学报, 2014, 34(9): 0930001.

[4] Li Huimei, Liu Gang, Ou Quanhong, *et al*. Analysis of eight bean species by two-dimensional correlation infrared spectroscopy[J]. Laser & Optoelectronics Progress, 2016, 53(3): 033003.
李会梅, 刘刚, 欧全宏, 等. 8 种豆的二维相关红外光谱的分析[J]. 激光与光电子学进展, 2016, 53(3): 033003.

[5] Chu Xiaoli, Lu Wanzhen. Research and application progress of near infrared spectroscopy analytical technology in China in the past five years[J]. Spectroscopy Spectral Analysis, 2014, 34(10): 2595-2605.
褚小立, 陆婉珍. 近五年我国近红外光谱分析技术研究与应用进展[J]. 光谱学与光谱分析, 2014, 34(10): 2595-2605.

[6] Guo Zhiming, Huang Wenqian, Peng Yankun, *et al*. Adaptive ant colony optimization approach to characteristic wavelength selection of NIR spectroscopy[J]. Chinese Journal of Analytical Chemistry, 2014, 42(4): 513-518.
郭志明, 黄文倩, 彭彦昆, 等. 自适应蚁群优化算法的近红外光谱特征波长选择方法[J]. 分析化学, 2014, 42(4): 513-518.

[7] Fang Kuangnan, Wu Jianbin, Zhu Jianping, *et al*. A review of technologies on random forests[J]. Statistics & Information Forum, 2011, 26(3): 32-38.
方匡南, 吴见彬, 朱建平, 等. 随机森林方法研究综述[J]. 统计与信息论坛, 2011, 26(3): 32-38.

[8] Zhang Aiwu, Xiao Tao, Duan Yihao, *et al*. A method of adaptive feature selection for airborne LiDAR point cloud classification[J]. Laser & Optoelectronics Progress, 2016, 53(8): 082802.
张爱武, 肖涛, 段乙好, 等. 一种机载 LiDAR 点云分类的自适应特征选择方法[J]. 激光与光电子学进展, 2016, 53(8): 082802.

[9] Yang Junwen, Zhang Jinshui, Zhu Xiufang, *et al*. Random forest applied for dimension reduction and classification in hyperspectral data[J]. Journal of Beijing Normal University (Natural Science Edition), 2015, 51(s1): 82-88.
杨珺雯, 张锦水, 朱秀芳, 等. 随机森林在高光谱遥感数据中降维与分类的应用[J]. 北京师范大学学报(自然科学版), 2015, 51(s1): 82-88.

[10] Yao Dengju, Yang Jing, Zhan Xiaojuan. Feature selection algorithm based on random forest[J]. Journal of Jilin University (Engineering and Technology Edition), 2014, 44(1): 137-141.
姚登举, 杨静, 詹晓娟. 基于随机森林的特征选择算法[J]. 吉林大学学报(工学版), 2014, 44(1): 137-141.

[11] Xu Yonggang, Zhang Jianye, Gong Xiaogang, *et al*. A method of real-time traffic classification in secure access of the power enterprise based on improved random forest algorithm[J]. Power System Protection and Control, 2016, 44(24): 82-89.

许勇刚, 张建业, 龚小刚, 等. 基于改进随机森林算法的电力业务实时流量分类方法[J]. 电力系统保护与控制, 2016, 44(24): 82-89.

[12] Breiman L. Random forests[J]. Machine Learning, 2001, 45(1): 5-32.

[13] Zhang Shihui, He Huan, Kong Lingfu, *et al*. Fusing multi-feature for video occlusion region detection based on graph cut[J]. Acta Optica Sinica, 2015, 35(4): 0415001.
张世辉, 何欢, 孔令富, 等. 融合多特征基于图割实现视频遮挡区域检测[J]. 光学学报, 2015, 35(4): 0415001.

[14] Tang Wei, Liu Fengnian, Chen Chongbang, *et al*. Application of improved Gini index in the text classification[J]. Journal of Changsha University, 2013, 27(5): 55-63.
唐伟, 刘丰年, 陈崇帮, 等. 改进的基尼指数在文本分类中的应用研究[J]. 长沙大学学报, 2013, 27(5): 55-63.

[15] Fan W L, Hu P, Liu Z G. Multi-attribute node importance evaluation method based on Gini-coefficient in complex power grids[J]. IET Generation, Transmission & Distribution, 2016, 10(9): 2027-2034.

[16] Yang F, Lu W H, Luo L K, *et al*. Margin optimization based pruning for random forest[J]. Neurocomputing, 2012, 94: 54-63.