山西大学学报(自然科学版)47(1):93－102,2024
Journal of Shanxi University(Nat. Sci. Ed.)

DOI:10.13451/j.sxu.ns.2023143

基于Stacking集成和偏探索贝叶斯优化的特征选择

孙林[1],郭嘉琪[1],朱雨晨[2],陈森[2*]
(1.天津科技大学 人工智能学院,天津 300457;
2.河南师范大学 计算机与信息工程学院,河南 新乡 453007)

摘 要:针对高维基因数据集的最优特征子集不易确定,以及传统的贝叶斯优化算法容易陷入局部最优,导致无法快速筛选出最优参数等问题,本文提出了一种基于Stacking集成和偏探索贝叶斯优化的基因选择方法。首先,使用卡方过滤法剔除原始特征空间中的冗余基因,获得相关性较高的基因,通过贝叶斯优化算法的采集函数进行改进,引入跳出系数,使得贝叶斯优化算法能够自适应地跳出局部最优,降低开销并加快寻优的效率;然后,使用偏探索贝叶斯优化寻找随机森林的最优参数,使用优化后随机森林模型筛选最优基因子集;最后,设计了一种Stacking集成模型框架来构建分类器,并对最优基因子集进行分类,进而构建了基于Stacking集成和偏探索贝叶斯优化的基因选择算法。在9个公开的基因表达谱数据集上进行仿真实验,结果表明所提算法可以快速筛选出最优的基因子集,且具有较高的分类精度。

关键词:基因选择;Stacking算法;贝叶斯优化算法;随机森林模型
中图分类号:TP391 **文献标志码**:A **文章编号**:0253-2395(2024)01-0093-10

Feature Selection Using Stacking Integration and Partial Exploration Bayesian Optimization

SUN Lin[1], GUO Jiaqi[1], ZHU Yuchen[2], CHEN Sen[2*]
(1. *College of Artificial Intelligence, Tianjin University of Science and Technology, Tianjin* 300457*, China;*
2. *College of Computer and Information Engineering, Henan Normal University, Xinxiang* 453007*, China*)

Abstract: To address the problems that the optimal feature subset of high-dimensional gene datasets is not easy to be determined and the traditional Bayesian optimization algorithm is prone to falling into local optimum, which cannot quickly select the optimal parameters, in this paper, we propose a gene selection method based on the Stacking integration and partial exploration Bayesian optimization. Firstly, the Chi-square filtering scheme is used to eliminate the redundant genes in the original feature space, so as to obtain the genes with high correlation. The acquisition function of the Bayesian optimization algorithm is improved, and the jump out coefficient is introduced, so that the Bayesian optimization algorithm can adaptively jump out of the local optimum. The cost can be reduced and the efficiency of optimization will be speeded up. Secondly, the partial exploration Bayesian optimization is used to find the optimal parameters of random forest. Then, the optimized random forest model is employed to screen the optimal feature subset. Finally, a framework of the Stacking integration model is designed to construct classifier and classify the optimal feature subset, and then a gene selection algorithm based on the Stacking integration and partial exploration Bayesian optimization is constructed. The experimental results on nine public gene expression profile datasets show that the proposed algorithm can quickly select the optimal gene subset with higher classification accuracy.

Key words: gene selection; stacking algorithm; bayesian optimization algorithm; random forest model

收稿日期:2023-06-23;**接受日期**:2023-09-18
基金项目:国家自然科学基金(61772176);河南省科技攻关项目(212102210136)
作者简介:孙林(1979-),男,河南南阳人,博士,教授,研究方向为粒计算、机器学习。E-mail: slinok@126.com
* **通信作者**:陈森(CHEN Sen),E-mail: chensen0420@126.com
引文格式:孙林,郭嘉琪,朱雨晨,等.基于Stacking集成和偏探索贝叶斯优化的特征选择[J].山西大学学报(自然科学版),2024,47(1):93-102. DOI:10.13451/j.sxu.ns.2023143

0 引言

近年来,从大量基因数据中选择有效的基因子集进行判别,已成为生物信息学和计算机科学的研究热点[1]。特征选择是从原始特征集合中选择具有区分能力的特征子集,在不改变特征含义的前提下减少特征的维度,提高分类的精度[2]。根据在选择过程中子集评价准则和后续的学习算法的关系,可将特征选择方法分为过滤、封装和嵌入[3]。

信噪比、卡方过滤、邻域互信息和皮尔逊相关系数等过滤法是根据数据的特点来进行特征选择,时间复杂度低但精选效果欠佳[4]。随机森林、遗传算法和粒子群优化算法等封装法由于有后续学习算法参与特征选择的评估和筛选,能够筛选出比过滤法更好的特征子集,分类效果较优但复杂度较高[5]。其中,随机森林算法适用进行特征选择和处理高维非线性数据的问题,而遗传算法和粒子群优化算法更适用解决全局搜索最优解的问题。由此,本文采用结合过滤式和封装式的特征选择方法,即两阶段的特征选择。其中,卡方过滤在进行特征过滤的同时,确保筛选出的特征是与标签相关的,提供了有效的决策信息;作为封装式的随机森林特征选择方法,通过训练得到各个特征的权重系数,依此进行特征选择,是一种快速高效的方法。贝叶斯优化算法多用来对模型的参数进行优化,使模型具有更高的表现和泛化性能,但目前的贝叶斯优化算法存在容易陷入局部最优的困境。为解决该问题,本文对贝叶斯优化算法的采集函数进行改进,提出了自适应跳出系数,使得在优化的过程中更加倾向于探索,从而跳出局部最优。

尽管目前大多数方法在特征选择方面取得了比较突出的研究成果,但是多采用单一的模型进行特征的筛选,模型的泛化性能不高。徐久成等[6]提出了一种基于信噪比和随机森林的肿瘤基因选择方法,快速筛选基因子集,但是存在随机森林泛化误差大的问题。唐容[7]通过贝叶斯优化得到决策树模型的最优参数,提出了基于贝叶斯优化和决策树的特征选择算法,但是贝叶斯优化存在陷入局部最优的风险。Deng等[8]提出了一种基于极端梯度提升算法和多目标优化遗传算法相结合的两阶段基因选择模型,可以快速的筛选分类精度高的基因子集,但需手动设置模型的参数。Lu等[9]提出了一种集成AdaBoost和遗传算法的混合模型,为每个基分类器设置权重,增加了基分类器池的随机性和多样性,但存在耗时长的问题。Mohammed等[10]提出了一种基于一维卷积神经网络和Stacking的堆叠集成深度模型,但是它不适用于小样本的数据集。基于此,本文提出一种Stacking框架,集成式完成特征选择的任务,提高模型表现能力的同时也确保了较高的泛化性能。

针对决策树模型容易出现过拟合、剪枝复杂和贝叶斯优化陷入局部最优,以及目前大多数方法使用单一分类器等问题,本文提出了基于偏探索贝叶斯优化随机森林模型,寻找最优特征子集,并使用Stacking集成模型构建集成分类器,提高模型的泛化性能和分类的准确度。实验结果也表明该算法筛选特征速度快,准确率高。本文的主要贡献为:

(1)设计一种偏探索贝叶斯优化模型,引入跳出系数,使贝叶斯优化尽早地跳出局部最优。

(2)使用贝叶斯优化算法改进随机森林特征选择模型,快速筛选特征的同时确保特征子集的精确性。

(3)针对高维基因数据,构建两阶段训练Stacking集成模型,结合Stacking集成和偏探索贝叶斯优化提出基因选择算法。

1 相关工作

1.1 贝叶斯优化

贝叶斯优化对于超参数的优化,可以看作是反应泛化误差的未知黑盒的优化[11]。采集函数评估采取点的有用性,以实现最大化黑盒函数的目标。

(1)概率最大化

利用后验估计模型值大于目前最大值的概率来确定下一采样点。概率最大化$P_1(x)$函数表示为

$$P_1(x)=\phi\left(\frac{\mu(x)-f(x^+)-\varepsilon}{\sigma(x)}\right), \tag{1}$$

其中$\mu(x)$和$\sigma(x)$是根据后验模型得到的期望和方差,$\varepsilon \geqslant 0$为可调参数,$\phi(\cdot)$为标准正态分

布的累计分布函数。P_{I}属于开发过程,方差$\sigma(x)$的作用很小,即使在不确定性很小的地方也可能计算得到一个较大的P_{I}值,会使整个优化过程陷入局部最优。

(2) 期望最大化

假设采集样本点的观测值为$f(x)$,若$f(x)>f(x^+)$,则定义$[\mu(x)-f(x^+)]^+$为采集点的提升,这种期望函数$E_{\mathrm{I}}(x)$表示为

$$E_{\mathrm{I}}(x)=\begin{cases}(\mu(x)-f(x^+))\phi(Z)+\sigma(x)\varphi(Z), & \sigma(x)>0,\\ 0, & \sigma(x)=0,\end{cases} \tag{2}$$

$$Z=\frac{\mu(x)-f(x^+)}{\sigma(x)}, \tag{3}$$

$$x_{\mathrm{new}}=\arg\max_{x}\{E_{\mathrm{I}}(x)\}, \tag{4}$$

其中,x^+为新的采集点,$\phi(\cdot)$和$\varphi(\cdot)$分别为累计分布函数和标准正态分布的概率密度函数。当$\mu(x)-f(x^+)$越大,则$E_{\mathrm{I}}(x)$越大。

(3) 上置信区间

为了比较置信区间的大小,则上置信区间函数$U_{\mathrm{CB}}(x)$定义为

$$U_{\mathrm{CB}}(x)=\mu(x)+\beta\sigma(x), \tag{5}$$

其中,$\mu(x)$为当前预测分布的高均值,β为调节参数,可理解为上置信边界。

1.2 过滤法

通过方差筛选特征,其方差越小,表示样本在特征上没有差异,应优先消除掉低方差的特征。方差过滤后,可继续筛选与分类关联且有意义的特征。卡方过滤通过计算每个非负特征和标签之间的卡方统计量,从高到低排名。排名越高表明特征与分类关联度越大。卡方统计量的计算公式为

$$\chi^2=\sum_{i=1}^{n}\frac{(x_i^{\mathrm{ob}}-x_i^{\mathrm{ep}})^2}{x_i^{\mathrm{ep}}}, \tag{6}$$

其中,x_i^{ob}表示第i个样本在该特征上的实际值,x_i^{ep}表示在第i个样本的期望值。卡方统计量值越大,表明该特征与分类的关联度越大。

1.3 随机森林特征选择

随机森林(Random Forest, RF)通过对弱分类器分类结果进行投票,获得最优的分类结果[6,12],即$T=\{\max(S(t_i),i=1,2,\cdots,n)\}$表示为RF模型的分类结果,$t_i$为每一棵决策树,$S_i$为每个决策树的投票结果,max表示取$n$个决策树的最优投票结果。

Bootstrap[12]是构成Bagging算法和RF模型的基础。在利用Bootstrap生成树的过程中,未使用的样本成为袋外数据(Out of Bag data, OOB data),特征重要性通过OOB data计算得出。于是,本文基于OOB data方法对特征进行筛选。假设有n棵决策树,使用OOB样本在训练好的决策树$t_i(i=1,2,\cdots,n)$上运行,其Bootstrap对应的数据集记为$D_i^{\mathrm{oob}}(i=1,2,\cdots,n)$,则特征$x_j$的重要性$F_j$的计算步骤为:

Step1:使用t_i对OOB数据D_i^{oob}进行分类,统计分类正确的数目,记为N_i^{oob};

Step2:对其中第j列特征加噪音,保持其他列不变,记扰动后的OOB数据集为D_{ij}^{oob},再次使用t_i进行分类,分类正确的数目记为N_{ij}^{oob};

Step3:对所有的决策树重复Step1和Step2;

Step4:得出特征x_j的重要性$F_j=\frac{1}{n}\sum_{i=1}^{n}(N_i^{\mathrm{oob}}-N_{ij}^{\mathrm{oob}})$。

2 特征选择方法

2.1 改进的贝叶斯优化

贝叶斯优化与是否找到全局最优及其效率和采集函数$A_{\mathrm{c}}(x)$有很大的关系。Snoek等[13]指出:在求解最优解的过程中,函数P_{I}的开发和探索是根据ε来平衡的,ε的取值需要针对不同的问题具体定义,而且对ε的值很敏感,过小则会陷入局部最优,过大则会大大降低效率,ε的大小不能动态地选择。期望函数E_{I}考虑$f(x)$比$f(x^+)$大多少,在贝叶斯寻优的过程中,会倾向于取x^+周围的点,原因是$\phi(Z)$在x^+周围的概率大,也会存在陷入局部最优的问题。

为了解决上述两种A_{c}函数存在的局部最优问题,结合P_{I}和E_{I}函数,自适应跳出系数定义为

$$\omega=\left|\lg|f(x)-f(x^+)|\right|+|f(x)-f(x^+)|, \tag{7}$$

$$Z=\frac{\mu(x)-f(x^+)-\varepsilon}{\sigma(x)}, \tag{8}$$

其中,ω由当前$f(x)$和当前最大$f(x^+)$得出,其大小随$|f(x)-f(x^+)|$改变,当两者越趋近,跳出系数越大,越容易跳出,是一种正则约束;当两者差别大时,跳出系数的范围为(0,1),继续进

行开发。基于P_{I}函数加入参数ε进行扰乱,扩大x的采集范围,是一种偏探索的策略。由此对采集函数A_{c}进行改进,则新的采集函数$P_{\mathrm{EI}\omega}(x)$定义为

$$P_{\mathrm{EI}\omega}(x)=\begin{cases}\dfrac{1}{\omega}\Big[\big(\mu(x)-f(x^{+})\big)\varphi(Z)+\sigma(x)\phi(Z)\Big], & \sigma(x)>0,\omega>1;\\ \omega\varphi\left(\dfrac{\mu(x)-f(x^{+})-\varepsilon}{\sigma(x)}\right), & \sigma(x)>0,0<\omega<1;\\ 0, & \sigma(x)=0。\end{cases}\tag{9}$$

由式(9)可知,在开始寻优的过程中,ω的值位于[0,1]之间,使用第二个函数对参数进行寻优,目的是偏向于寻找下一个值最大的概率;当ω的值大于1时,A_{c}函数选择式(9)中的第一个函数。因而,使用ω在寻优的过程中既能确保较大的期望,也能保持偏向探索的能力。当陷入局部最优时,通过ω减小其期望值,从而跳出局部最优。同时,$P_{\mathrm{EI}\omega}(x)$也可以避免模型多次使用重复值,从而提高模型效率。

综上可知,结合P_{I}和E_{I}这两种函数,构建新的A_{c}函数$P_{\mathrm{EI}\omega}(x)$,其中自适应跳出系数ω和集合杂乱参数ε可以扩大x的搜索范围,同时也帮助贝叶斯优化算法在寻优过程中跳出局部最优,加快搜索效率。

2.2 Stacking集成模型

Stacking模型使用不同的基分类器,得到了异质的集成模型,具有并行度高、泛化性能好的优点[14],于是使用Stacking模型构建肿瘤基因分类模型。该模型利用高级学习器和低级学习器获得更好的预测准确度,分为两层,第一层为基学习器,其具备比较高的精度,且第一层分类器差异较大。第二层为元学习器,是较为简单的分类器,解决过拟合的问题。由此,本文使用逻辑回归作为元分类器。在第一层不同的分类器上对同一数据集进行训练,得到多个不同的基学习器,其分类结果作为第二层元分类器的输入。元学习器的输入还包括真实样本的类别,元学习可以做到改正基分类器的分类错误,提升分类的准确度。Stacking算法的具体步骤描述如下:

算法1 Stacking算法

输入:数据集$D=\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$,基学习算法$\{A_1,A_2,\cdots,A_N\}$和元学习算法$A^{*}$;

输出:元学习器L_f和分类结果。

Step1:使用不同的基学习算法$\{A_1,A_2,\cdots,A_N\}$得到不同的基学习器$\{L_1,L_2,\cdots,L_N\}$;

Step2:用基学习器对原始数据集进行分类得到输出$\{o_1,o_2,\cdots,o_N\}$,并构建新的数据集

$D_{\mathrm{new}}=\{(o_1,y_1),(o_2,y_2),\cdots,(o_N,y_N)\}$;

Step3:使用元学习算法A^{*}训练得到元学习器L_f;

Step4:基于元学习器L_f进行分类,输出分类结果。

2.3 算法描述

本文基于偏探索的贝叶斯优化对RF模型的参数进行寻优,目标函数是使模型的交叉验证得分越高越好;对卡方过滤后的肿瘤基因数据集,设计基于偏探索的贝叶斯优化和RF模型的特征选择算法(Exploration Bayesian Optimization and Random Forest-Based Feature Selection, BRFS),寻找最优基因子集;对BRFS算法筛选出的最优特征子集进行Stacking集成模型研究(BRFS with Stacking Integrating Algorithm, BRFS-S);最后使用不同的分类器进行分类。其伪代码如算法2所示。

算法2 BRFS-S算法

输入:基因数据集、改进的贝叶斯优化模型、RF模型及其参数取值空间和Stacking集成模型

输出:最优基因子集

Step0:卡方过滤基因数据集,得到初步基因子集F_{start};

Step1:While不满足最大迭代次数或停止条件do;

Step2:归一化基因数据集,设定期望交叉验证得分y;

Step3:随机选取一组RF参数集合ξ_{old};

Step4:初始贝叶斯优化模型;

Step5:ξ_{old}为已知所有参数集合中交叉验证$tar(x)$取得最大值的参数;

Step6:if $tar(\xi_{\mathrm{old}})>y$ then 返回ξ_{old}和

$tar(\xi_{old})$;

Step7: else 随机初始化;

Step8: While $tar(\xi_{new})<y$;

Step9: 高斯回归求解未知参数集合,使用$P_{EI\omega}(x)$寻找贝叶斯优化猜测的最大值参数取值ξ_{new};

Step10: 返回ξ_{new}和$tar(\xi_{new})$,使用ξ_{new}构建RF分类器;

Step11: 对数据集进行卡方过滤并导入RF分类器;

Step12: 选择出重要性排名的前n个基因;

Step13: 挑选$tar(x)$最高的类别个数m;

Step14: End while;

Step15: 输出最优基因子集F_{end};

Step16: 构建Stacking集成模型进行分类。

在BRFS-S算法中,假设基因数据集包括m个样本和n个基因,则卡方过滤的计算复杂度为$O(m)$;第二阶段执行基于贝叶斯优化和RF的特征选择,假设树深度为q,则计算复杂度为$O(m\log q)$。由此分析可知,该算法的计算复杂度为$O(m+m\log q)$。

3 实验结果与分析

3.1 实验数据集和预处理

为评估本文算法的性能,选用支持向量机(Support Vector Machine, SVM)、K-近邻(K-Nearest Neighbor, KNN)和朴素贝叶斯(Naive Bayes, NB)这3种分类器,在http://portals.broadinstitute.org/cgi-bin/cancer/datasets.cgi上选择9个基因表达数据集(如表1所示),进行仿真实验。实验硬件配置为酷睿i5-1035G7、1.2 GHz和16 GB内存,在Pycharm 2020软件中编程实现。为保证对比算法的稳定性,选择验证数据集上分类正确率(Accuracy, Acc)最高作为特征选择的依据,并将10次迭代的平均分类正确率作为迭代所构建分类器的分类精度[15]。接下来,首先使用方差过滤对基因数据集进行过滤,将方差为0的过滤掉,然后对过滤后的基因计算卡方值并排序,由于卡方值大于0,需要对基因数据进行无量纲化,加快模型的求解速度。这里将基因数据集的值归一化,对每一个基因求其卡方值,其中卡方值越大,表明该基因对分类的作用越大,分类相关性越大。为选出较优的基因子集,使用学习曲线确定卡方值的阈值,筛选出若干个基因作为预选基因子集。于是,对表1中9个基因数据集进行学习曲线分析并确定卡方值的下限,以得到不同基因数据集的最优预选基因子集。在下面所有实验结果中,粗体表示最优值。

表1 9个基因数据集信息

Table 1 Information of nine gene datasets

序号	数据集	基因个数	样本个数	类别个数
D1	Conlon	2 000	62	2
D2	Breast	9 216	82	5
D3	SRBCT	2 308	83	4
D4	Leukemia	7 129	72	2
D5	MLL	12 582	72	3
D6	Lung	12 600	203	2
D7	Brain	10 509	102	2
D8	Prostate	12 600	136	2
D9	DLBCL	5 469	77	2

注:SRBCT: Small Round Blue Cell Tumors(小圆蓝细胞肿瘤);MLL:Mixed-lineage Leukemias(混合谱系白血病);DLBCL:Diffuse Large B-cell Lymphoma(弥漫性大B细胞淋巴瘤)

3.2 RF最优参数及Stacking集成模型设置

BRFS算法中使用贝叶斯优化对RF进行参数寻优,参数分别为树的个数n,树的深度d,最大特征F_{max},最小样本分割S_{min},最小样叶L_{min},将$P_{EI\omega}(x)$作为A_c函数对RF分类器的参数进行寻优,目标函数为十则交叉验证均值得分最高的结果。参数的取值空间分别为树的个数为[100, 800]、树的深度为[5, 40],最小样本分割为[5, 40],最小样叶为[5, 40],最大特征为[0.05, 1],贝叶斯优化迭代次数为300,寻找表1中的9种基因数据集进行RF分类器的最优参数集合。通过贝叶斯优化可得到多组最优解,考虑到树的个数对于内存的影响最大,树的个数越多,消耗的内存越多,因此选择多组最优解中树的个数最少的一组作为模型的优化参数。各基因数据集在取值空间内的最优参数如表2所示。

表2为使用$P_{EI\omega}(x)$得到的最优参数,在构建RF分类器时,由于RF模型的参数过多,本文只对上述5类重要的参数构造取值空间,当模型的参数过多时会造成过拟合现象。通过贝叶斯优化可以得到多组最优解,鉴于模型中树

表2 9个基因数据集上贝叶斯优化的RF模型的最优参数

Table 2 The optimal parameters of the RF model based on the Bayesian optimization on nine gene datasets

数据集	树的个数	树的深度	最大特征	最小分割	最小样叶
Conlon	172	8	0.37	18	5
Breast	736	5	0.6777	7	5
SRBCT	100	5	0.1	21	5
Leukemia	186	16	0.885 5	7	20
MLL	171	25	0.891 7	10	12
Lung	357	32	0.967 6	24	5
Brain	195	5	0.05	5	5
Prostate	165	15	0.999 9	39	6
DLBCL	180	16	0.695 0	23	19

的个数越多,需要的计算资源越多,以树的个数为主要评判依据,使用树的个数最少的一组参数来构建模型。在训练阶段,利用Stacking集成模型对9种基因数据集构建分类预测模型。将通过卡方过滤和BRFS算法筛选基因和原始的真实标签作为集成模型的输入。此外,在训练集成模型时,使用留一交叉验证(Leave-one-out Cross-validation,LOOCV)模型,对样本的利用率极高,尤其适合肿瘤基因小样本的情况。对于第一层分类模型,使用SVM、KNN和NB的分类模型进行分类预测。第二层模型考虑泛化性能,使用逻辑回归算法进行最后的分类。

3.3 卡方过滤实验结果

使用卡方过滤初选,对不同基因数据集筛选得分最高的基因,作为初步基因子集,在不降低Acc的前提下去除无关基因以及与分类关联度较低的基因。因此,用卡方过滤减少9个高维基因数据集的维度。对这9个数据集使用RF分类器进行分类,参数设置为本文改进的贝叶斯优化模型得出的最优参数(见表2),其分类结果如表3所示。针对基因数据集样本数少的情况,采用文献[11]中的准确度,将10次结果得分的平均值作为最终的Acc。通过卡方过滤在不同的基因数据集上初选的基因子集,筛选的依据为使用网格搜索在一定的范围内进行搜索,搜索的范围为[30,500],确保筛选出的初选基因子集维度固定,便于后续进行精选基因子集。使用RF分类器对各个数据集初选的基因子集进行分类。由表3可知,卡方过滤对基因数据集进行初择后,各数据集去除了无关基因,RF分类器的Acc在一定程度上均有所提升,这说明了初择保留了与分类关联度相关的基因。

表3 卡方过滤前后使用RF分类器的结果对比

Table 3 Comparison between results obtained using RF classifier before and after chi-square filtering

数据集	原始特征集合		初选特征子集	
	基因数	Acc	基因数	Acc
Conlon	2 000	0.832 7	**300**	**0.878 9**
Breast	9 216	0.859 7	**155**	**0.927 8**
SRBCT	2 308	0.952 3	**90**	**1.0**
Leukemia	7 129	0.973 2	**75**	**0.985 7**
MLL	12 582	0.9	**255**	**0.957 1**
Lung	12 600	0.97	**55**	**0.99**
Brain	10 509	0.880 1	**70**	**0.909 9**
Prostate	12 600	0.88	**90**	**0.91**
DLCBL	5 469	0.885 7	**50**	**0.92**

3.4 BRFS与RF的实验结果分析

为了验证BRFS算法的有效性,与使用偏探索贝叶斯优化后的RF模型在3种分类器下进行分类性能比较。表4为RF和BRFS这两种算法在SVM、KNN和NB这3种分类器下的分类效果,由于基因数据维度较高,而表4中数据集是经过卡方过滤后筛选出的新样本数据。由表4可知,在SVM分类器下的9个数据集上,BRFS算法的Acc均优于RF算法。具体来看,BRFS算法在SRBCT数据集上达到全局最优值,Acc达到99.60%,表现最优;在Lung数据集上,RF与BRFS算法的Acc差别不大,BRFS算法相较于RF算法的Acc高出0.17%;在Conlon数据集上,BRFS算法的Acc比RF算法高出3.34%。因此,在SVM分类器下,BRFS算法整体表现较好,提升效果为0.17%~3.34%。在KNN分类器下,在这9个基因数据集中,BRFS算法的Acc也均优于RF算法。具体来看,BRFS算法在Leukemia与Lung数据集上达到最优,Acc达到98.68%;在Lung数据集上,RF与BRFS算法的Acc表现相同;在Breast数据集上,BRFS算法的Acc比RF算法高出4.29%。因此,在KNN分类器下,BRFS算法整体上效果可观,提升了0%~4.29%。在NB分类器下,这9个基因数据集中,BRFS算法的Acc均优于RF算法。具体来看,BRFS算法在SRBCT数据集上达到全局最优值,Acc达到99.20%,在此数据

集上,RF与BRFS算法的Acc表现相同;在MLL数据集上,BRFS算法取得较好的提升,比RF算法的Acc高出4.55%。因此,在NB分类器下,BRFS算法整体表现较好,提升效果最高可达4.55%。结合在SVM、KNN和NB这3种分类器上Acc的实验对比,相较于RF算法,BRFS算法在Acc上表现良好。

为了更直观地呈现在不同分类器上的分类效果,图1—3给出3种分类器下10次实验结果的平均值。实验结果表明,在3种分类器上BRFS算法均优于RF算法,尤其是在MLL数据集上提升效果最为显著,提升效果为3%~5%。在分类器SVM上,分类预测的时间开销最短,这是因为SVM训练完成后只与支持向量有关,使得模型更加高效。在SRBCT数据集上,BRFS和RF这两种算法的Acc都很高,这表明本文使用卡方过滤来消除冗余基因的有效性。同时,在不同的基因数据集上,BRFS算法在各分类器下的Acc差距很小,表明经过卡方过滤后再进一步精选基因确保了后续的分类任务的稳定性。

表4 3种分类器上BRFS与RF在9个基因数据集上的正确率

Table 4 Comparison of accuracy between BRFS and RF on nine gene datasets under three classifiers

数据集	SVM		KNN		NB	
	RF	BRFS	RF	BRFS	RF	BRFS
Conlon	0.884 2	**0.917 6**	0.889 4	**0.915 7**	0.9	**0.910 5**
Breast	0.938 4	**0.95**	0.902 3	**0.946 1**	0.938 1	**0.953 8**
SRBCT	0.992 0	**0.996 0**	0.979 9	**0.984 0**	0.992 0	**0.992 0**
Leukemia	0.945 4	**0.977 2**	0.954 5	**0.986 8**	0.963 6	**0.968 1**
MLL	0.927 2	**0.954 5**	0.922 7	**0.963 6**	0.909 0	**0.954 5**
Lung	0.981 9	**0.983 6**	0.986 8	**0.986 8**	0.981 9	**0.983 6**
Brain	0.922 5	**0.945 1**	0.919 3	**0.929 0**	0.912 2	**0.935 4**
Prostate	0.912 1	**0.926 8**	0.904 3	**0.926 8**	0.902 4	**0.926 8**
DLCBL	0.904 1	**0.920 8**	0.912 5	**0.916 7**	0.9	**0.912 5**

为了进一步分析BRFS算法的有效性,这里比较了RF和BRFS在表现最好情况下筛选出的基因数量。两种算法中RF模型的参数均为本文改进贝叶斯优化算法得到的最优参数。这里以SVM、KNN和NB分类器上的最高得分作为其Acc。表5为RF算法与BRFS算法的Acc确度和筛选基因个数的对比,Acc取3种分类器中表现最优的。从表5可以看出,BRFS算法均展现了比RF的Acc高以及基因数少的优

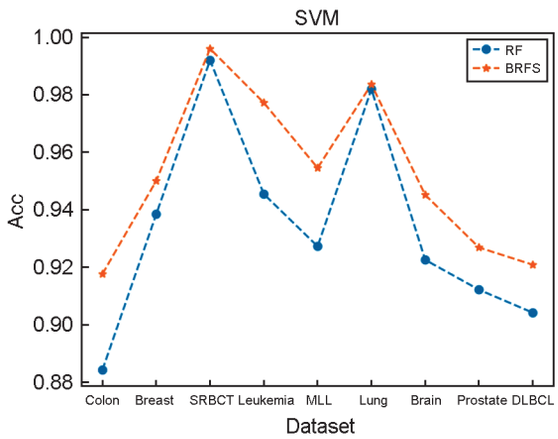

图1 SVM分类器上BRFS与RF在9个基因数据集上的正确率(Acc)

Fig. 1 Acc of BRFS and RF on nine gene datasets under the SVM classifier

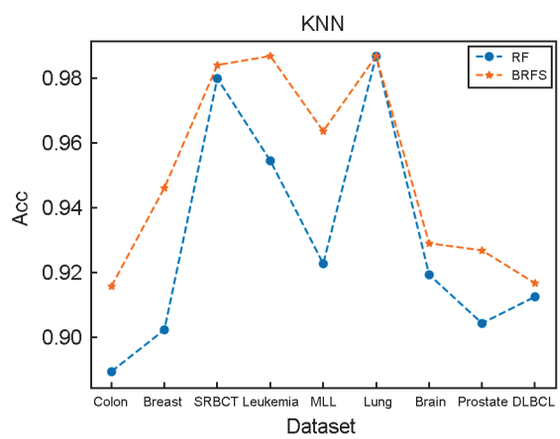

图2 KNN分类器上BRFS与RF在9个基因数据集上的正确率(Acc)

Fig. 2 Comparison of accuracy (Acc) of BRFS and RF on nine gene datasets under the KNN classifier

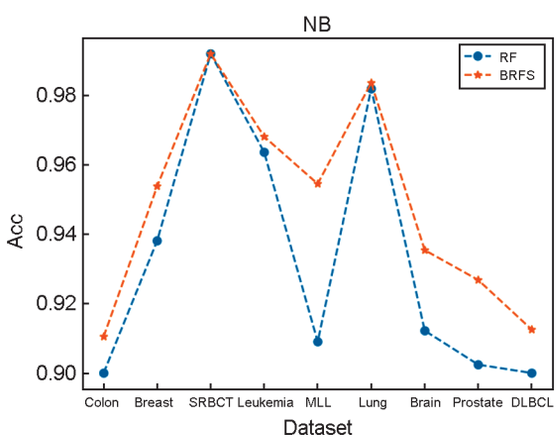

图3 NB分类器上BRFS与RF在9个基因数据集上的正确率(Acc)

Fig. 3 Comparison of accuracy (Acc) of BRFS and RF on nine gene datasets under the NB classifier

表5 RF与BRFS的Acc和筛选基因个数对比

Table 5 Comparison of accuracy and number of features selected by the RF and BRFS algorithms

数据集	RF		BRFS	
	Acc	选择基因数	Acc	选择基因数
Conlon	0.9	7	**0.917 6**	7
Breast	0.938 4	12	**0.953 8**	6
SRBCT	0.992 0	12	**0.996 0**	5
Leukemia	0.963 6	10	**0.986 8**	6
MLL	0.927 2	**10**	**0.963 6**	**10**
Lung	**0.986 8**	10	**0.986 8**	6
Brain	0.922 5	10	**0.945 1**	7
Prostate	0.912 1	8	**0.926 8**	6
DLBCL	0.912 5	9	**0.920 8**	5

点。具体来看,在SRBCT数据集上,BRFS算法在筛选出5个基因时的Acc达到99.60%,表现最优;在MLL数据集中,RF与BRFS算法所选择的基因数相同,均为10个,但BRFS算法的Acc比RF算法高3.64%,同样在Conlon数据集中,选择基因数均为7个的情况下,BRFS算法的Acc比RF算法高3.64%。因此,BRFS算法整体上效果可观,提升效果最高可达3.64%,可以说明贝叶斯优化寻优是有效的。需要说明的是,在Lung数据集上,这两种算法的Acc均为0.986 8。但是,BRFS算法选出的基因个数比RF特征选择算法少了4个,表明了BRFS算法提取了更加关键的基因子集。

3.5 BRFS与BRFS-S实验结果分析

为验证Stacking集成框架对于模型泛化性能的提升效果,在9个基因数据集上3种分类器下比较BRFS和BRFS-S这两种算法的Acc。表6给出这两种算法在9个基因数据集上的平均Acc。从表6结果可知,BRFS-S算法在SRBCT和Leukemia这两个数据集上的Acc均为100%,而BRFS算法在这两个数据集上的Acc分别为0.996 0和0.986 8,表明单分类器存在一定的分类误差,通过引入Stacking集成模型可以很好地解决这个问题;在Lung数据集上,BRFS-S算法虽未取得最优值,但两种算法的Acc差距较小,仅为0.23%;在DLCBL数据集上,BRFS-S算法的Acc比BRFS算法高出6.25%,提升效果最为明显。总体来看,在大部分数据集上,BRFS-S算法能够带来较好的分类效果,提升效果为0.4%~6.25%。

为了更形象地展示BRFS和BRFS-S这两种算法的最优结果,表7给出了9个数据集上在召回率(recall)[6]、精确率(precision)[6]以及调和精确率和召回率的均值($F1$)[15]的实验结果。由表7可知,从recall指标来看,BRFS-S算法在Brain、MLL与Lung这3个数据集上的Acc可以达到100%;在Leukemia与SRBCT这两个数据集上,BRFS算法与BRFS-S算法的Acc均为100%;在Conlon数据集上,BRFS-S算法的Acc比BRFS算法高出5.55%;在DLCBL数据集上,虽然BRFS-S算法的Acc略逊于BRFS算法,但差距仅为1.67%。除在DLBCL数据集上,BRFS-S算法均优于BRFS算法。从precision指标来看,BRFS-S算法在9个基因数据集上均最优。在Conlon数据集上,BRFS-S算法的Acc比BRFS算法高出5.87%,提升最为明显;在Leukemia与SRBCT这两个数据集上,BRFS算法与BRFS-S算法的Acc均为1.0;BRFS-S算法在MLL与Lung这两个数据集上的Acc达到100%,而BRFS的Acc分别为99.28%和99.24%。从$F1$指标来看,BRFS-S算法在9种数据集上同样均优于BRFS算法。在Breast数据集上,BRFS-S算法的Acc比BRFS算法高出4.08%,提升较为明显;BRFS-S算法在Leukemia、MLL和Lung这3个数据集上的Acc可以达到100%。总体上看,在3种指标上,BRFS-S算法在绝大部分基因数据集上均能够带来较好的分类效果。

3.6 与其他算法的实验对比

为进一步验证本文所提算法的有效性,与现有8种其他算法进行比较,主要包括:文献[8]中结合XGBoost和多目标遗传算法(Genetic Algo-

表6 9个基因数据集上BRFS和BRFS-S的平均Acc

Table 6 Average Acc of BRFS and BRFS-S on nine gene datasets

算法	Colon	Breast	SRBCT	Leukemia	MLL	Lung	Brain	Prostate	DLCBL
BRFS	0.917 6	0.953 8	0.996 0	0.986 8	0.963 6	**0.986 8**	0.945 1	0.926 8	0.920 8
BRFS-S	**0.947 3**	**0.961 5**	**1.0**	**1.0**	**0.986 3**	0.983 6	**0.974 2**	**0.975 6**	**0.983 3**

表7 9个基因数据集上BRFS与BRFS-S关于召回率(recall)、精确率(precision)和均值(*F*1)的实验结果

Table 7 Experimental results of BRFS and BRFS-S in terms of recall, precision and *F*1

指标	算法	Brain	DLBCL	Colon	Leukemia	MLL	Prostate	SRBCT	Lung	Breast
recall	BRFS	0.967 7	**1.0**	0.905 2	**1.0**	0.990 9	0.914 6	**1.0**	0.991 8	0.915 3
	BRFS-S	**1.0**	0.983 3	**0.960 7**	**1.0**	**1.0**	**0.931 7**	**1.0**	**1.0**	**0.950 0**
precision	BRFS	0.970 0	0.968 8	0.916 3	**1.0**	0.992 8	0.927 1	**1.0**	0.992 4	0.937 2
	BRFS-S	**0.977 2**	**0.984 9**	**0.975**	**1.0**	**1.0**	**0.937 8**	**1.0**	**1.0**	**0.963 9**
*F*1	BRFS	0.967 7	0.974 5	0.905 3	0.994 1	0.991 1	0.915 1	**1.0**	0.991 8	0.910 0
	BRFS-S	**0.975 1**	**0.983 4**	**0.916 0**	**1.0**	**1.0**	**0.931 7**	**1.0**	**1.0**	**0.950 8**

表8 9种算法在5个代表性基因数据集上的正确率(Acc)对比结果

Table 8 Comparison results of nine algorithms on five representative gene datasets in terms of accuracy (Acc)

Algorithms	Colon	Leukemia	Breast	Lung	Brain
XGBoost-MOGA	0.902 4	0.985 7	0.823 3	0.988 9	—
AdaBoost-RF	**0.966 7**	0.984 6	0.908 8	0.999 2	0.963 4
AdaBoost-SVM	0.903 5	0.963 1	0.883 4	0.896 9	0.837 0
AdaBoost-BPNN	0.938 1	0.989 7	0.896 2	0.975 8	0.911 1
SLFDD	0.903 0	0.986 0	—	**1.0**	—
IG-SGA	0.855 0	0.971 0	—	**1.0**	—
TSLR	0.938 0	0.937 0	—	0.964 0	—
BRFS	0.917 6	0.986 8	0.953 8	0.986 8	0.945 1
BRFS-S	0.947 3	**1.0**	**0.961 5**	0.986 6	**0.974 2**

rithm, GA)优化的XGBoost-MOGA算法,将RF[12]、SVM和反向传播(Backpropagation, BP)神经网络[10]分别与文献[9]的AdaBoost算法结合设计3种算法:AdaBoost-RF、AdaBoost-SVM和AdaBoost-BPNN,以及半监督学习故障检测与诊断算法(Semi-supervised Learning Fault Detection and Diagnosis, SLFDD)[16]、信息增益和标准遗传算法(Information Gain and Standard Genetic Algorithm, IG-SGA)[17]和两阶段稀疏逻辑回归T算法(Two-stage Sparse Logistic Regression, TSLR)[18]。参照文献[16]的实验安排,从表1中选择5个代表性的基因数据集,对这9种算法开展Acc的对比实验,其结果如表8所示。

从表8可知,XGBoost-MOGA算法虽然在Lung和Leukemia这两种数据集上表现较优,但在Breast数据集上的Acc较低,波动较为明显。IG-SGA算法也存在类似问题,在Lung数据集上的Acc可达100%,但在Conlon数据集上Acc较低,为85.50%。基于AdaBoost的3种算法在5种基因数据集上的表现效果较为稳定,其中AdaBoost-RF算法在Colon数据集上取得了最好的结果。综合来看,与其他8种算法相比,BRFS-S算法在Leukemia、Breast和Brain这3个基因数据集上取得了最优结果,且均超过其他算法的平均水平,表现稳定。具体来讲,在Leukemia数据集上表现最优,BRFS-S算法的Acc达到100%,相较于其他算法高出0.93%~1.94%;在Breast数据集上,BRFS-S算法的Acc相较于其余算法高出0.77%~13.82%;在Brain数据集上,相较于其余算法的Acc高出1.08%~13.72%。除了Colon数据集之外,BRFS算法的实验结果也均高出平均水平。这也充分验证了本文提出的BRFS和BRFS-S这两种算法的高效性和稳定性。

4 结束语

为了有效识别和分类肿瘤基因,本文提出一种基于Stacking集成和偏探索贝叶斯优化的基因选择方法。首先,为了设计偏探索的贝叶斯优化算法,将两种采集函数结合并引入跳出系数设计了新的采集函数,进而约束贝叶斯优化算法,提高寻优速度,避免局部最优;然后,使用偏探索的贝叶斯优化改进RF模型,并将其应用到特征选择上,快速精选特征子集,引入Stacking集成模型来提高分类效果并降低泛化误差;最后,一系列实验证明了本文所研究的算法具有更好的分类性能和稳定性。但是,本文算法在计算开销、贝叶斯搜索过程的并行优化以及Stacking模型的层数设置上有待进一步改进。因此,在下一步研究工作中专注解决上述问题。

参考文献:

[1] SUN L, WANG L Y, DING W P, *et al*. Feature Selection Using Fuzzy Neighborhood Entropy-based Uncertainty Measures for Fuzzy Neighborhood Multigranulation Rough Sets[J]. *IEEE Trans Fuzzy Syst*, 2021, **29**(1): 19-33. DOI: 10.1109/TFUZZ.2020.2989098.

[2] 孙林, 徐枫, 李硕, 等. 基于ReliefF 和最大相关最小冗余的多标记特征选择[J]. 河南师范大学学报(自然科学版), 2023, **51**(6): 21−29+175. DOI: 10.16366/j. cnki. 1000-2367.2023.06.003.
SUN L, XU F, LI S, *et al*. Multilabel Feature Selection Algorithm Using ReliefF and Mrmr[J]. *J Henan Norm Univ Nat Sci Ed*, 023, **51**(6): 21−29+175. DOI: 10.16366/j. cnki.1000-2367.2023.06.003.

[3] 刘艳, 程璐, 孙林. 基于K-S检验和邻域粗糙集的特征选择方法[J]. 河南师范大学学报(自然科学版), 2019, **47**(2): 21−28. DOI: 10.16366/j.cnki.1000-2367.2019.02.004.
LIU Y, CHENG L, SUN L. Feature Selection Method Based on K-S Test and Neighborhood Rough Set[J]. J *Henan Norm Univ Nat Sci Ed*, 2019, **47**(2): 21−28. DOI: 10.16366/j.cnki.1000-2367.2019.02.004.

[4] 钱宇华, 王川杭, 王婕婷. 消除随机一致性的互信息及决策树算法[J]. 山西大学学报(自然科学版), 2022, **45**(5): 1206−1215. DOI: 10.13451/j.sxu.ns.2021016.
QIAN Y H, WANG C H, WANG J T. Mutual Information and Decision Tree Algorithm with Eliminating Random Consistency[J]. *J Shanxi Univ Nat Sci Ed*, 2022, **45**(5): 1206−1215. DOI: 10.13451/j.sxu.ns.2021016.

[5] 李冰晓, 万睿之, 朱永杰, 等. 基于种群分区的多策略综合粒子群优化算法[J]. 河南师范大学学报(自然科学版), 2022, **50**(3): 85−94. DOI: 10.16366/j. cnki. 1000-2367.2022.03.011.
LI B X, WAN R Z, ZHU Y J, *et al*. Multi-strategy Comprehensive Article Swarm Optimization Algorithm Based on Population Partition[J]. *J Henan Norm Univ Nat Sci Ed*, 2022, **50**(3): 85−94. DOI: 10.16366/j. cnki. 1000-2367.2022.03.011.

[6] 徐久成, 冯森, 穆辉宇. 基于信噪比与随机森林的肿瘤特征基因选择[J]. 河南师范大学学报(自然科学版), 2017, **45**(2): 87−92. DOI: 10.16366/j.cnki.1000-2367.2017.02.017.
XU J C, FENG S, MU H Y. Tumor Feature Gene Selection Based on SNR and Random Forest[J]. *J Henan Norm Univ Nat Sci Ed*, 2017, **45**(2): 87−92. DOI: 10.16366/j.cnki.1000-2367.2017.02.017.

[7] 唐容. 基于特征选择的CART算法研究[D]. 成都: 电子科技大学, 2020. DOI: 10.27005/d.cnki.gdzku.2020.000752.
TANG R. Research on CART algorithm based on feature selection[D]. Chengdu: University of Electronic Science and Technology of China, 2020. DOI: 10.27005/d. cnki. gdzku.2020.000752.

[8] DENG X S, LI M, DENG S B, *et al*. Hybrid Gene Selection Approach Using XGBoost and Multi-objective Genetic Algorithm for Cancer Classification[J]. *Med Biol Eng Comput*, 2022, **60**(3): 663−681. DOI: 10.1007/s11517-021-02476-x.

[9] LU H J, GAO H Y, YE M C, *et al*. A Hybrid Ensemble Algorithm Combining AdaBoost and Genetic Algorithm for Cancer Classification with Gene Expression Data[J]. *IEEE/ACM Trans Comput Biol Bioinform*, 2021, **18**(3): 863−870. DOI: 10.1109/TCBB.2019.2952102.

[10] MOHAMMED M, MWAMBI H, MBOYA I B, *et al*. A Stacking Ensemble Deep Learning Approach to Cancer Type Classification Based on TCGA Data[J]. *Sci Rep*, 2021, **11**: 15626. DOI: 10.1038/s41598-021-95128-x.

[11] 曹建芳, 田晓东, 贾一鸣, 等. 基于改进PSPNet网络的古代壁画分割方法[J]. 河南师范大学学报(自然科学版), 2022, **50**(4): 65−75. DOI: 10.16366/j. cnki. 1000-2367.2022.04.010.
CAO J F, TIAN X D, JIA Y M, *et al*. Segmentation Method of Ancient Murals Based on Improved PSPNet [J]. *J Henan Norm Univ Nat Sci Ed,* 2022, **50**(4):65−75. DOI: 10.16366/j.cnki.1000-2367.2022.04.010.

[12] AMINI N, MAHDAVI M, CHOUBDAR H, *et al*. Automated Prediction of COVID-19 Mortality Outcome Using Clinical and Laboratory Data Based on Hierarchical Feature Selection and Random Forest Classifier[J]. *Comput Methods Biomech Biomed Eng*, 2023, **26**(2): 160−173. DOI: 10.1080/10255842.2022.2050906.

[13] SNOEK J, LAROCHELLE H, ADAMS R P. Practical Bayesian Optimization of Machine Learning Algorithms [C]//Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 2. New York: ACM, 2012: 2951−2959. DOI: 10.5555/2999325.2999464.

[14] PAVLYSHENKO B. Using Stacking Approaches for Machine Learning Models[C]//2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP). IEEE, 2018: 255−258. DOI: 10.1109/dsmp.2018.8478522.

[15] SUN L, ZHANG X Y, QIAN Y H, *et al*. Feature Selection Using Neighborhood Entropy-based Uncertainty Measures for Gene Expression Data Classification[J]. *Inf Sci*, 2019, **502**: 18−41. DOI: 10.1016/j.ins.2019.05.072.

[16] YAN K, ZHONG C W, JI Z W, *et al*. Semi-supervised Learning for Early Detection and Diagnosis of Various Air Handling Unit Faults[J]. *Energy Build*, 2018, **181**: 75−83. DOI: 10.1016/j.enbuild.2018.10.016.

[17] SALEM H, ATTIYA G, EL-FISHAWY N. Classification of Human Cancer Diseases by Gene Expression Profiles [J]. *Appl Soft Comput*, 2017, **50**: 124−134. DOI: 10.1016/j.asoc.2016.11.026.

[18] ALGAMAL Z Y, LEE M H. A Two-stage Sparse Logistic Regression for Optimal Gene Selection in High-dimensional Microarray Data Classification[J]. *Adv Data Anal Classif*, 2019, **13**(3): 753−771. DOI: 10.1007/s11634-018-0334-1.