计算机科学
COMPUTER SCIENCE

基于投影相关和随机森林融合模型的疾病诊断

韩怡梅, 李东喜

引用本文

韩怡梅, 李东喜. 基于投影相关和随机森林融合模型的疾病诊断[J]. 计算机科学, 2023, 50(11A): 230200172-6.

HAN Yimei, LI Dongxi. Disease Diagnosis Based on Projection Correlation and Random Forest Fusion Model [J]. Computer Science, 2023, 50(11A): 230200172-6.

相似文章推荐（请使用火狐或 IE 浏览器查看文章）

Similar articles recommended (Please use Firefox or IE to view the article)

基于模型融合思想的程序化交易投资者识别研究

Study on Programmatic Trading Investors Recognition Based on Model Fusion

计算机科学, 2023, 50(11A): 230300131-6. https://doi.org/10.11896/jsjkx.230300131

基于spike-and-slab先验的贝叶斯时间序列模型

Bayesian Time-series Model Based on spike-and-slab Prior

计算机科学, 2023, 50(11A): 221200131-6. https://doi.org/10.11896/jsjkx.221200131

基于图嵌入的正交局部保持投影无监督特征选择

Orthogonal Locality Preserving Projection Unsupervised Feature Selection Based on Graph Embedding

计算机科学, 2023, 50(11A): 220900003-9. https://doi.org/10.11896/jsjkx.220900003

基于子图特征的节点排序算法

Node Ranking Algorithm Based on Subgraph Features

计算机科学, 2023, 50(11A): 230100122-7. https://doi.org/10.11896/jsjkx.230100122

一种约束验证神经网络的方法

Constraint-based Verification Method for Neural Networks

计算机科学, 2023, 50(11A): 221000045-5. https://doi.org/10.11896/jsjkx.221000045

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230200172

基于投影相关和随机森林融合模型的疾病诊断

韩怡梅[1]　李东喜[2]
1 太原理工大学数学学院　太原 030024
2 太原理工大学大数据学院　太原 030024
(hanyimei1998@163.com)

摘　要　针对高维数据的处理方法已成为当前研究大数据的热点问题之一。提出一种基于投影相关系数的两阶段随机森林模型(Projection Correlation-Random Forest,PC-RF),它将度量随机变量相关性的投影相关系数与随机森林算法相融合,在预测性能上表现出更优的结果。使用3种基因微阵列数据进行实证分析,在Leukemia和Colon数据集实验中,所提模型比现有算法准确率提升了2.4%～6.5%;在Breast数据集实验中,所提模型比传统随机森林模型准确率提升了3.55%～9.26%,同时在不同规模高维数据中的多种评价指标上表现稳定且优良。所提模型应用在基于微阵列数据的疾病诊断领域,将为疾病预防和诊断治疗提供更加科学有效的决策支持。

关键词:投影相关系数;随机森林;高维数据;特征选择;机器学习
中图法分类号　TP391

Disease Diagnosis Based on Projection Correlation and Random Forest Fusion Model

HAN Yimei[1] and LI Dongxi[2]
1 College of Mathematics,Taiyuan University of Technology,Taiyuan 030024,China
2 College of Big Data,Taiyuan University of Technology,Taiyuan 030024,China

Abstract　The processing method for high-dimensional data has become one of the hot issues in the study of big data. In this paper,a two-stage random forest algorithm based on projection correlation is proposed,which integrates the projection correlation to measure the correlation of random variables with the random forest algorithm,and shows better results in prediction performance. Three kinds of gene data are used for experimental analysis. In the experiments on Leukemia and Colon datasets,the accuracy of the proposed model improves by 2.4%～6.5% compared with the existing algorithms. In the experiment on Breast data set,the accuracy rate of the proposed algorithm increases by 3.55%～9.26% compared with the traditional random forest model,and it also performs stably and well in various evaluation indexes of high-dimensional data of different scales. The application of the model in the field of disease diagnosis based on microarray data will provide more scientific and effective decision support for disease prevention,diagnosis and treatment.

Keywords　Projection correlation,Random forest,High dimensional data,Feature selection,Machine learning

1　引言

近二十年来,科学技术的飞速发展促进了新技术的变革,改变了人类生活的方方面面,也推动了社会生产力的进步与发展。人类正从IT(Information Technology)时代走向DT(Data Technology)时代。计算能力和技术突飞猛进的发展促使数据收集到达了前所未有的规模,各科研领域在数据收集成本不断降低的同时,数据收集能力也大大提高,高维数据频繁出现在金融数据、大脑图像、微阵列分析、蛋白质组学、环境数据、肿瘤分类等领域,且出现了数据量大、超高维、价值密度低等新的特征。

在当前信息爆炸式增长的时代,虽然各行各业的高维数据中暗藏了很多冗余,但其中也更大程度地包含了信息背后的"真相"。探求各种有效分析和利用高维数据的方法,不仅成为学术界热烈讨论的话题,而且从长远来看能够为社会的发展带来巨大财富。因此,高维数据的降维处理方法自提出以来,就成为了数据分析领域关注的重点。

Fan和Lv[1]在2008年首先提出了SIS(Sure Independence Screening),根据响应变量和各协变量之间的皮尔逊相关系数(Pearson Correlation Coefficient,PCC)来度量它们之间的相关性,从而能够将协变量的维数降到低于样本量的大小,并证明了在一定的正则条件下SIS具有筛选一致性(Sure Screening Property),即当样本量趋于无穷时,所有显著变量被选入模型的概率趋近于1。SIS方法自提出,就为高维数据的特征选择方法奠定了理论基础和思路,至今还在世界各地学者不断的讨论和研究下扩展到新的领域,衍生出了多种

基金项目:国家自然科学基金项目(11571009);山西省回国留学人员科研资助项目(2022-074)
This work was supported by the National Natural Science Foundation of China(11571009) and Research Project Supported by Shanxi Scholarship Council of China(2022-074).
通信作者:李东喜(dxli0426@126.com)

特征选择算法。例如，Li[2]基于 Kendall 相关系数度量协变量和响应变量之间的相关性，提出了稳健秩相关筛选(Robust Rank Correlation Screening，RRCS)；Fan 等[3]将 SIS 扩展到超高维稀疏模型，提出了非参数独立筛选(Nonparametric Independence Screening，NIS)等。

SIS 方法及其在不同模型假定或无模型假定下的改进方法有一大共性，它们都是通过某种统计指标来度量响应变量与协变量之间的相关关系，如皮尔逊相关系数、Kendall 相关系数、距离相关系数等，并据此采用边际筛选方法选出重要协变量。边际筛选方法是使得高维特征筛选得以快速实施的关键。这一关键思路在 Niu[4]关于特征筛选方法的综述以及 He[5]的基于最大边际效用的特征筛选方法中有过呈现。

本文基于投影相关系数(Projection Correlation，PC)来度量两个随机变量间的相关关系，将其应用于特征与类别之间的相关性计算，再结合随机森林算法(Random Forest，RF)对降维后的数据进行分类预测，建立基于投影相关系数的两阶段随机森林模型——PC-RF 模型(Projection Correlation-Random Forest)。降维与分类分两步走的两阶段算法使得本文模型有较强的泛化能力，能够更好地模拟真实情况，适应来自现实生活的高维数据。本文模型的创新之处在于将投影相关系数与随机森林算法相结合，拓宽了投影相关系数作为特征选择指标时的适用范围，同时也解决了使用随机森林算法对高维数据集进行预测时计算量大、准确性不足的缺点。

2 模型构建

2.1 投影相关系数

本文使用投影相关系数(Projection Correlation)[6]来度量两个随机变量间的相关关系。

定义任意维数的向量 X 和 Y，由单位向量 α 和 β 可构成向量 $U=\alpha^{\mathrm{T}}X$ 和 $V=\beta^{\mathrm{T}}Y$，并且用 $F_U(u)$ 和 $F_V(v)$ 分别表示其边际分布函数，用 $F_{U,V}(u,v)$ 表示其联合分布函数，那么向量 X 和 Y 相互独立等价于：

$\forall\alpha,\beta$，向量 $U=\alpha^{\mathrm{T}}X$ 和 $V=\beta^{\mathrm{T}}Y$ 相互独立。

那么对于任意给定单位向量的 α 和 β，U 和 V 相互独立等价于：

$$F_{U,V}(u,v)-F_U(u)F_V(v)=\mathrm{cov}\{I(\alpha^{\mathrm{T}}X\leqslant u),I(\beta^{\mathrm{T}}Y\leqslant v)\}=0 \tag{1}$$

因此，检验向量 X 和 Y 相互独立，等价于检验下述公式是否成立：

$$\iint \mathrm{cov}^2\{I(\alpha^{\mathrm{T}}X\leqslant u),I(\beta^{\mathrm{T}}Y\leqslant v)\}\mathrm{d}F_{U,V}(u,v)\mathrm{d}\alpha\,\mathrm{d}\beta=0 \tag{2}$$

其中，$I(\cdot)$是示性函数。根据 Fubini 定理和 Escanciano[7]中的引理可对上式进行展开与推导，具体运算过程和推导细节见文献[6]。

从而定义随机向量 X 和 Y 投影协方差(Projection Covariance)的平方为：

$$\begin{aligned}\{P\ \mathrm{cov}(X,Y)\}^2 &=S_1+S_2-2S_3\\ &=E\left[\arccos\left\{\frac{(X_1-X_3)^{\mathrm{T}}(X_4-X_3)}{\|X_1-X_3\|\,\|X_4-X_3\|}\right\}\arccos\left\{\frac{(Y_1-Y_3)^{\mathrm{T}}(Y_4-Y_3)}{\|Y_1-Y_3\|\,\|Y_4-Y_3\|}\right\}\right]+\\ &\quad E\left[\arccos\left\{\frac{(X_1-X_3)^{\mathrm{T}}(X_4-X_3)}{\|X_1-X_3\|\,\|X_4-X_3\|}\right\}\arccos\left\{\frac{(Y_2-Y_3)^{\mathrm{T}}(Y_5-Y_3)}{\|Y_2-Y_3\|\,\|Y_5-Y_3\|}\right\}\right]-\\ &\quad 2E\left[\arccos\left\{\frac{(X_1-X_3)^{\mathrm{T}}(X_4-X_3)}{\|X_1-X_3\|\,\|X_4-X_3\|}\right\}\arccos\left\{\frac{(Y_2-Y_3)^{\mathrm{T}}(Y_4-Y_3)}{\|Y_2-Y_3\|\,\|Y_4-Y_3\|}\right\}\right]\end{aligned} \tag{3}$$

其中，$\|\cdot\|$代表 L2 范数。观察上式可知，投影协方差的一个显著特征是它只使用形式为$(X_k-X_l)/\|X_k-X_l\|$的式子进行运算，使得对于任意维数的随机向量，消除了距离相关(Distance Correlation)[8-9]所要求的对(X,Y)矩的限制。

接着，定义随机向量 X 和 Y 的投影相关系数为下式的平方根：

$$\{PC(X,Y)\}^2=\frac{\{P\ \mathrm{cov}(X,Y)\}^2}{P\ \mathrm{cov}(X,X)\ P\ \mathrm{cov}(Y,Y)} \tag{4}$$

并且当 $P\ \mathrm{cov}(X,X)=0$ 或 $P\ \mathrm{cov}(Y,Y)=0$ 时，定义 $PC(X,Y)=0$。在一般情况下 $0\leqslant PC(X,Y)\leqslant 1$，当且仅当 X 和 Y 相互独立时，$PC(X,Y)=0$，并且有 $PC(X,X)=0$ 当且仅当 $X=E(X)$成立。上述性质表明，投影相关系数一般适用于作为衡量随机变量间相关性的指标。

利用 V 统计量可给出 $PC(X,Y)$的样本估计，经过计算和推导证明得到，投影相关检验在不需要任何力矩条件的情况下，对所有依赖项都是一致的，具体运算过程和推导细节见文献[6]。

2.2 基于投影相关系数的两阶段随机森林模型

本文将投影相关系数应用于特征与类别之间的相关性计算，并与随机森林算法相融合，建立用于高维数据处理的两阶段随机森林模型。

随机森林(Random Forest，RF)属于机器学习的一大分支——集成学习(Ensemble Learning)算法，是通过集成学习中 Bagging 的思想将多棵决策树集成的一种算法。集成学习是一种近年来十分流行的有监督机器学习算法，通过组合使用多种学习算法来获得比单独使用任何一种算法更为优秀的训练表现和预测性能，其中 Bagging 思想是将一些单个分类器作为权重相同的基分类器，针对同一模型数据集中有放回抽取的训练集进行投票表决，最终的分类结果由所有分类器投票所得，投票次数最多的类别被指定为最终的输出。一般情况下，采取 Bagging 得到的分类器比单个分类器得到的结果更准确，并且在过拟合问题和有噪声数据的训练上鲁棒性更强。

随机森林算法是 Bagging 的一种拓展方法，以决策树作为基分类器的同时引入随机属性选择，即在决策树的每个结点随机地从可选属性组合中抽取部分属性，再从中选择一个最优属性用于划分。其应用前景很广泛，在欺诈检测、市场营销、金融决策和疾病风险预测等领域都有发挥空间，对于分类、回归、异常点检测等问题的解决都可提供帮助。但随机森林算法在应用于高维数据分类预测问题时，存在计算量大、准确率有待提高等问题。为降低计算复杂度，提高预测准确率，将投影相关系数作为特征选择指标融入随机森林算法进行改进，构建基于投影相关系数的两阶段随机森林模型——Projection Correlation-Random Forest(PC-RF)模型。模型构造过程如下：

设 $y=(Y_1,\cdots,Y_q)^{\mathrm{T}}$ 为样本的 q 维响应变量，$x=(X_1,\cdots,X_p)^{\mathrm{T}}$ 为样本的 p 维特征，在高维数据集中，样本维数 p 远大于样本量 n，特征选择方法可以去除与响应变量无关或是相关程度低的特征，将数据的特征维数降低到样本量

以下。定义活跃特征集：

$$A(\delta)=\{k:PC(X_k,Y)^2\geqslant\delta,1\leqslant k\leqslant p\} \tag{5}$$

特征集中的特征必属于也仅属于 $A(\delta)$或 $A^c(\delta)$($A(\delta)$的补集)中的一个，其中 δ 是预先设定好的阈值，参考文献[1]，设定 δ 使得 $A(\delta)$中的特征数量为$[n/\log n]$([]为取整符号，即取不超过[]中数字的最大整数)。

将本章模型的第一阶段称为 PC 筛选，由文献[10]可知，在选择适当 δ 的情况下，能够证明 PC 筛选具有一致性，即当概率接近 1 时，所有活跃特征都包含在活跃特征集 $A(\delta)$中。

设 $\omega_k\equiv PC(X_k,y)^2$ 是第 k 个特征和相应变量之间的总体投影相关系数的平方，接着有条件：

$$\exists\, c>0,0\leqslant\kappa<1/2 \quad \text{s.t.}\min_{k\in A}\omega_k\geqslant 2cn^{-\kappa} \tag{6}$$

式(6)可以看作是一种稀疏性假设，保证了活跃特征与非活跃特征的区别。

在式(6)成立的条件下，取 $\delta\leqslant\min_{k\in A}\omega_k/2$ 时，满足：

$$\Pr(A\subseteq A(\delta))\geqslant 1-O(s\exp\{-c_1n^{1-2\kappa}\}) \tag{7}$$

其中，c_1 是大于零的常数。即 PC 筛选具有一致性，且收敛速度快于 DC-SIS，详细证明过程见文献[10]。

经过第一阶段的特征选择，将所选活跃特征集中的特征及样本的响应变量整理成新的数据集，进入第二阶段，将新数据集放入基于随机森林算法构建的分类器中，对分类结果进行预测，并针对具体数据调整算法中的参数，最后评估分类结果，即完成本文模型的运算流程。

本文将投影相关系数作为特征选择方法，随机森林算法作为分类器，二者相融合构建基于投影相关系数的两阶段随机森林模型，并将其应用于基因微阵列数据。模型的算法流程图如图 1 所示。

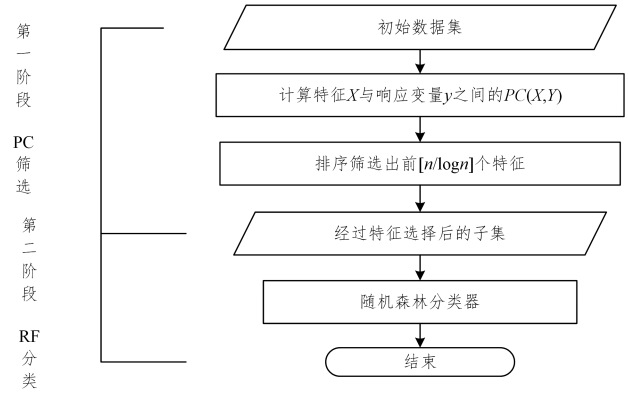


图 1　PC-RF 模型算法流程图

Fig. 1　Algorithm flow chart of PC-RF model

3　实验

3.1　实验环境及数据

本文的实验环境基于 python3.8，首先使用两类癌症的基因微阵列数据，数据来自博德基因研究所的公开数据网站[1)]。

结肠癌数据集(Colon)[11]由 Alon 等收集，包含 62 个样本，每个样本记录了 2000 个基因，其中 40 个样本为患病样本，22 个为非患病样本。随机抽取 80%的样本作为训练集，剩余 20%作为测试集。$[n/\log n]=34$，在模型的特征选择阶段选取 34 个特征进行分类。

白血病数据集(Leukemia)[12]由 Golub 等收集，包含 72 个急性白血病样本，每个样本记录了 7129 个基因，其中 47 个样本为 ALL(急性淋巴白血病)，25 个为 AML(急性髓细胞白血病)。随机抽取 80%的样本作为训练集，剩余 20%作为测试集。$[n/\log n]=38$，在模型的特征选择阶段选取 38 个特征进行分类。

在应用本文模型对基因微阵列数据进行研究预测时，将基因看作特征，受试者患病与否(或疾病的种类)看作响应变量，即 $Y_i(i=1,\cdots,q)$的取值为 0 或 1(代表是否患病或疾病的不同种类)，特征选择阶段的目的在于从所有检测的基因数据中选择出与疾病表现型相关的基因，降低后续分类预测阶段的计算复杂度。

原始数据在进入模型计算之前，先用式(8)进行标准化，以克服量纲对模型计算的影响。经过标准化，特征数据分布在[0,1]区间。标准化表达式为：

$$x_i=\frac{x_i-\bar{x_i}}{\sigma_{x_i}}(i=1,\cdots,p) \tag{8}$$

其中，x_i 为基因观测值，$\bar{x_i}$ 和 σ_{x_i} 分别是其均值和方差。

3.2　实验结果分析

采用分类结果的准确率作为评价指标，并与使用在相同数据集上的现有算法进行对比，结果如表 1 所列。

由表 1 容易看出本文模型在降维和预测方面的优越性，与其他现有算法相比，PC-RF 模型以更少的特征将准确率提升了 2.4%～6.5%。为了更进一步研究 PC-RF 模型在数据处理上的特性，以及特征数目选择对后续预测结果的影响，将继续使用维数更庞大的数据集进行实验。

实验使用的乳腺癌数据集(Breast)[17]由 Yau 等收集，包含 683 个样本，每个样本记录了 9168 个基因，其中 447 个样本为雌激素受体阴性 ER－，236 个为雌激素受体阳性 ER＋。

表 1　Leukemia 和 Colon 数据集在现有模型下的对比

Table 1　Comparison of Leukemia and Colon datasets in existing models

数据集	特征数量	准确率/%	特征选择方法	分类器	方法来源
Leukemia	50	97.06	组内组间平方和比率	二次判别分析	[13]
	50	95.94	偏最小二乘	Logistic 判别	[14]
	25～1000	88.24～94.12	Fisher 判别准则	SVM	[15]
	38	**100**	projection correlation	随机森林	本文
Colon	50	94.70	组内组间平方和比率	二次判别分析	[13]
	50	93.50	偏最小二乘(PLS)	Logistic 判别	[14]
	12	93.55	遗传算法	SVM	[16]
	34	**100**	projection correlation	随机森林	本文

1) https://www.broadinstitute.org/

针对 Breast 数据集使用随机森林模型与 PC 筛选阶段选定特征数为 240，120，80（对应$[n/\log n]$，$[n/\log n]/2$、$[n/\log n]/3$）的 PC-RF 模型做分类预测，在随机抽取 75%的样本作为训练集，剩余 25%作为测试集时，该数据集在实验下的分类准确率如表 2 所列。

表 2 Breast 数据集在随机森林模型与 PC-RF 模型下的对比

Table 2 Comparison of Breast data set under random forest model and PC-RF model

实验模型	随机森林模型	PC-RF 模型		
特征个数	9168	240	120	80
准确率/%	81.25	90.51	88.32	84.8

表 2 数据证实了 PC-RF 模型在高维数据预测的准确性上明显优于未经改进的随机森林模型。为探究投影相关系数在作为特征选择的指标时，与随机森林算法相融合是否能够达到最优效果，将 Breast 数据集经过本文模型的第一阶段 PC 筛选处理后，针对在 3 种选定的特征数下整理得到的新数据集，采取不同的分类算法进行分类预测，并在多项衡量分类效果的指标下做对比。

实验使用十折交叉得到的平均 auc 值（下文简称 auc 值）、75%训练集划分下计算得到的准确率、敏感性、特异性、平衡准确率（下文均做简称，不再重复说明表 3 数据是 75%训练集划分下得到），这 5 项指标作为衡量分类预测结果的评判标准。

表 3 中加粗字体为实验在 PC 筛选阶段选定不同特征数下各指标的最优结果。由此可以直观地看出，PC-RF 模型在大部分指标上都占有优势，特别是敏感性和平衡准确率上表现出显著优于其他分类器的效果。

表 3 PC 筛选后 Breast 数据集在不同分类器下的效果对比

Table 3 Effect comparison of Breast data set filtered by projection correlation under different classifiers

特征数	分类器	auc 值	准确率	敏感性	特异性	平衡准确率
240	adaboost	0.860	0.848	0.874	0.789	0.831
	bagging	0.808	0.818	0.837	0.784	0.811
	beyes	0.854	0.591	0.575	0.621	0.598
	ctree	0.869	0.848	0.842	0.860	0.851
	logistic	0.636	0.591	0.575	0.621	0.598
	nnet	0.870	0.883	0.783	**0.920**	0.851
	PC-RF	**0.878**	**0.905**	**0.914**	0.893	**0.903**
120	adaboost	0.856	0.844	0.883	0.750	0.817
	bagging	0.855	0.842	0.895	0.702	0.799
	beyes	0.856	0.778	0.816	0.702	0.759
	ctree	0.788	0.836	0.868	0.781	0.825
	logistic	0.813	0.719	0.774	0.607	0.691
	nnet	0.830	0.784	0.726	**0.808**	0.767
	PC-RF	**0.870**	**0.883**	**0.951**	0.700	**0.825**
80	adaboost	0.521	0.567	0.682	0.361	0.521
	bagging	0.583	0.550	0.655	0.345	0.500
	beyes	0.855	0.830	0.840	0.815	0.828
	ctree	0.828	0.830	0.834	0.662	0.748
	logistic	0.848	0.790	0.846	0.667	0.756
	nnet	0.822	0.854	0.786	0.887	0.836
	PC-RF	**0.870**	**0.848**	**0.857**	**0.935**	**0.896**

敏感性是指分类中被正确预测为正例的样本个数在实际为正例的样本中占有的比率，应用在本节使用的基因微阵列数据中，就是指在患病的病例中能检测出来多少真实患者，即诊断方法对疾病的敏感程度、识别程度，也可称其为不漏诊率。敏感性的表现对于应用在疾病诊断领域的算法至关重要。

特异性是指正确预测为负例的样本个数在实际为负例的样本中占有的比率，即在测验的阴性结果中，有多少是真阴性。样本的健康特征与发病特征是有区别的，特异性表示了对这种区别的利用程度，可理解为不误诊率。

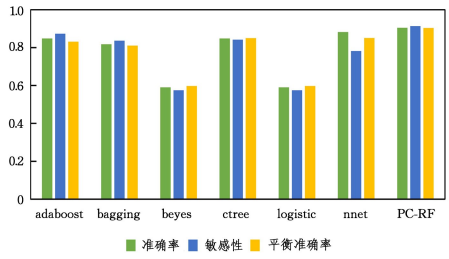
敏感性与特异性在实验中往往呈现相反的变化趋势。PC-RF 模型在敏感性上表现明显优越，在特异性上没能全部达到最优，使用这两项指标的算术平均值——平衡准确率，可以更全面地体现出本文模型的优越性。

表 3 中 auc 值这一指标，充分排除了实验随机抽取训练集和测试集对不同分类模型的影响，有力体现出随机森林分类器在与投影相关系数作为特征选择指标结合时的分类效果显著优于其他分类器，证实了本文提出的 PC-RF 模型在分类效果上的优越性。

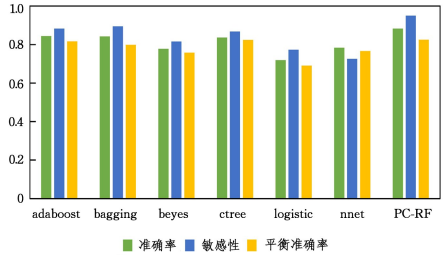
3.3 讨论

将准确率、敏感性、平衡准确率这 3 个指标在不同特征数下的实验结果用柱形图表示出来，如图 2 所示。从中可以直观看出，在模型第一阶段 PC 筛选出不同特征数的情况下，随机森林算法与 PC 筛选融合得到的效果最好也最稳定。

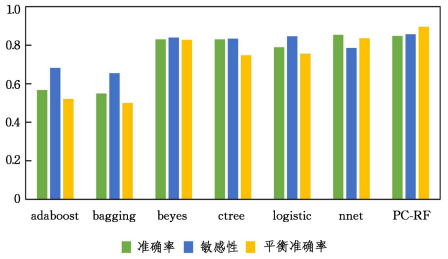
此外，本文模型在敏感性和平衡准确率上的良好效果也体现了将模型应用在患病诊断时的实际意义。高敏感性模型意味着将其应用于医疗诊断时，患病者被正确诊断的概率大，能够有效保护患者在就医时得到及时治疗，保障了患者的生命安全。平衡准确率是综合考虑敏感性和特异性的指标，该指标的效果良好，说明本文模型的特异性高，也就是没有患病的样本被正确判定为健康状况的比例高。综合来说，本文模型在诊断正确率上的效果良好且稳定，应用在临床上时，能够为医护人员的判断提供真实可靠的科学依据。



(a)所选特征数为 240 时 3 种指标的情况



(b)所选特征数为 120 时 3 种指标的情况



(c)所选特征数为 80 时 3 种指标的情况

图 2　不同选定特征数下 3 种指标的情况

Fig. 2　Three indicatorswith different selected feature numbers

图 3 给出了经过本文模型第一阶段 PC 筛选后，选定不同特征数下各分类器的准确率情况。能够看出，PC-RF 模型在选定不同特征数下的分类效果明显达到了最为稳定和优良的状态，证实随机森林算法在与 PC 筛选相融合时能够达到最佳效果。

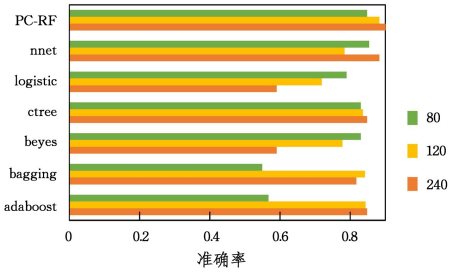
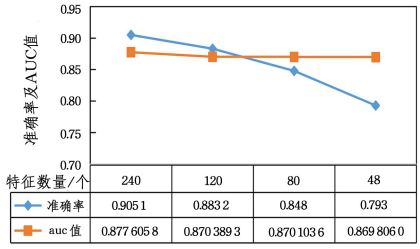


图 3　不同特征数下各分类器的准确率

Fig. 3　Accuracy of each classifier with different feature numbers

从表 3 中还可以观察到，在选定不同特征数的情况下，分类准确率的变化幅度并不明显。为了寻找到包含原始数据有效信息较多且特征数较小的情况，图 4 给出了 PC-RF 模型在选定不同特征数下，准确率、auc 值这两大重要指标的变化趋势。可以看出，auc 值随特征个数的变化幅度不明显，而准确率存在相对明显的下降趋势。由图 4 可以直观看出，在本实验中，既能包含数据集中大部分有效信息，又能使预测准确率保持在 80%以上，在减少计算量的情况下可以选择的特征数为$\lceil n/\log n \rceil/3=80$。由此可以得知，针对具体的数据情况，在保持分类准确率的前提下，可指定不同的特征数以降低计算复杂度，使模型更加精简。



特征数量/个	240	120	80	48
准确率	0.9051	0.8832	0.848	0.793
auc 值	0.8776058	0.8703893	0.8701036	0.8698060

图 4　PC-RF 模型第一阶段筛选特征数与分类准确率的关系

Fig. 4　Relationship between the first stage screening feature number and classification accuracy of PC-RF model

结束语　本文提出了一种基于投影相关系数的两阶段随机森林模型——PC-RF 模型，并将其应用于不同类型癌症的基因微阵列数据，进行患病诊断与疾病分类的实证分析，结果证实本文模型的预测效果显著优于传统随机森林模型，且在不同规模的高维数据集上均取得良好效果。使用投影相关系数作为特征选择的指标，消除了对所需处理数据的矩的限制，两阶段算法拓宽了模型的适用范围；选择随机森林算法作为分类器，经过与其他分类器对比证实了其与 PC 筛选相融合达到的分类效果最优。

本文模型还可以与其他算法相结合，比如与其他树模型、其他分类算法相结合，或考虑基因与环境相互作用。本文模型在诊断正确率上的效果良好且稳定，应用在临床上时，能够为医护人员的判断提供真实可靠的科学依据。

参 考 文 献

[1] FAN J Q, LV J C. Sure Independence Screening for Ultrahigh Dimensional Feature Space[J]. Journal of the Royal Statistical Society. Series B(Statistical Methodology), 2008, 70(5): 849-911.

[2] LI G R, PENG H, ZHANG J, et al. Robust Rank Correlation Based Screening[J]. The Annals of Statistics, 2012, 40(3): 1846-1877.

[3] FAN J, FENG Y, SONG R. Nonparametric Independence Screening in Sparse Ultra-High Dimensional Additive Models [J]. Publications of the American Statistical Association, 2011, 106(494): 544-557.

[4] NIU Y, LI H P, LI Y H, et al. Review of feature screening methods for ultra-high dimensional data[J]. Applied Probability Statistics, 2021, 37(1): 69-110.

[5] HE S M, WANG X. Ultra-high-dimensional feature screening method based on maximum marginal utility[J]. Statistics and Decision, 2021, 37(15): 38-43.

[6] ZHU L P, XU K, LI R Z, et al. Projection correlation between two random vectors[J]. Biometrika, 2017, 104(4): 829-843.

[7] ESCANCIANO J. A Consistent Diagnostic Test For Regression Models Using Projections [J]. Econometric Theory, 2006, 22(6): 1030-1051.

[8] DAVID S, MATTESON, RUEY S. Tsay. Independent Component Analysis via Distance Covariance[J]. Journal of the Ameri-

can Statistical Association, 2017, 112(518): 623-637.

[9] LI R, ZHONG W, ZHU L. Feature Screening via Distance Correlation Learning[J]. Am Stat Assoc., 2012, 107(499): 1129-1139.

[10] LIU W J, KE Y, LIU J Y, et al. Model-free Feature Screening and FDR Control with Knockoff Features[J]. Journal of the American Statistical Association, 2020, 117(537): 428-443.

[11] ALON U, NOTTERMAN D A. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays[J]. Proceedings of the National Academy of Sciences, 1999, 96(12): 6745-6750.

[12] GOLUB T R, SLONIM D K. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression[J]. Science, 1999, 286(5439): 531-537.

[13] ANTONIADIS A, LAMBERT-LACROIX S, LEBLANC F. Effective dimension reduction methods for tumor classification using gene expression data[J]. Bioinformatics, 2003(5): 563-570.

[14] NGUYEN D V, ROCKE D M. Tumor classification by partial least squares using microarray gene expression data[J]. Bioinformatics, 2002, 18(1): 39-50.

[15] FUREY T S, CRISTIANINI N, DUFFY N, et al. Support vector machine classification and validation of cancer tissue samples using microarray expression data[J]. Bioinformatics, 2000, 16(10): 906-14.

[16] PENG S, XU Q, LING X B, et al. Molecular classification of cancer types from microarray data using the combination of genetic algorithms and support vector machines[J]. FEBS LETTERS, 2003, 555(2): 358-362.

[17] YAU C, ESSERMAN L, DAN H M, et al. A multigene predictor of metastatic outcome in early stage hormone receptor-negative and triple-negative breast cancer[J]. Breast Cancer Research: BCR, 2010, 12(5): R85.

HAN Yimei, born in 1998, postgraduate. Her main research interests include data mining and machine learning.

LI Dongxi, born in 1982, Ph.D, associate professor. His main research interests include data analysis, data mining, machine learning, biostatistics and biomathematics.