文章编号：2095-2163(2021)08-0130-05 中图分类号：TP181 文献标志码：A

基于改进随机森林优化算法在医疗数据中的应用研究

朱 城[1]，苏前敏[1]，郭晶磊[2]，沈宙锋[1]
（1 上海工程技术大学 电子与电气工程学院，上海 201620；2 上海中医药大学，上海 201203）

摘 要：本文针对临床疾病预测过程中医疗临床数据特征因子相关性较强，模型调参工作较为复杂的问题，提出一种多模型复合优化方法。考虑到过多的强相关性特征因子很容易降低模型的运行效率，利用 SelectKbest 变量筛选算法对临床数据进行筛选，采用遗传算法对随机森林分类器模型的参数选择进行优化。最后，以乳腺癌的临床数据为例，实验证明通过以上方法优化后算法模型的精准值、召回率、*F*1 分值、*AUC* 值等方面均有提高，该提出的超参数调优方法为具有强共线性的临床数据处理和疾病预测提供了一种新思路。

关键词：遗传算法；kbest 变量筛选；随机森林；超参数调优

Research on application of improved random forest optimization algorithm in medical data

ZHU Cheng[1], SU Qianmin[1], GUO Jinglei[2], SHEN Zhoufeng[1]
(1 School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai 201620, China;
2 Shanghai University of Traditional Chinese Medicine, Shanghai 201203, China)

【Abstract】 In the process of clinical disease prediction, the medical clinical data feature factors are strongly correlated and the model adjustment work is more complicated. A multi-model compound optimization method is proposed. Considering that too many strong correlation feature factors can easily reduce the operating efficiency of the model, the SelectKbest algorithm is used to screen the clinical data, and the genetic algorithm is used to optimize the parameter selection of the random forest classifier model. Finally, taking the clinical data of breast cancer as an example, experiments prove that the accuracy of the algorithm model after optimization by the above method, recall rate, *F*1 score, *AUC* value and other aspects are improved. The proposed method provides a new idea for clinical data processing and disease prediction with strong collinearity.

【Key words】 genetic algorithm; kbest variable selection; random forest; hyper parameter tuning

0 引 言

国外研究人员早在 2007 年就将大数据运用在流感预测上，根据用户网络搜索的关键词来判别流感是否爆发、流感预测，比当时的疾控中心还能早两周就预测出流感发病率[1]。精准智能医疗可以实现对未知病人是否患病情况进行智能预测，常用于肿瘤疾病的预测[2]。如华大基因等公司最近更是推出了自主研究的肿瘤基因检测服务，通过采取患者样本，对患者的癌组织进行相关基因分析，实现乳腺癌等癌症患者的早期检测[3]。

当前，众多学者、专家在医疗数据领域运用多种机器学习算法进行了分析和预测。提出了使用支持向量机建立模型对乳腺癌数据进行处理分析，发现基于 K-medoids 聚类和支持向量机的改进算法（KD-SVM）分类精准率优于 H-SVM 算法[4]；提出了决策树机器学习算法在乳腺癌诊断中的应用，预测准确率达 96%[5]；研究了联合决策树及 logistic 回归，建立乳腺癌相对风险预测模型，在保证模型检测准确率的同时通过 logistic 组合模型能够对患病的危险特征进行捕捉和判断[6]；提出了基于随机森林的乳腺癌计算机辅助诊断研究，相比决策树，使用随机森林算法构建模型，最终的检测准确率高达 96.93%，又有了新的提升[7]。

本文应用了对医疗数据预测率较高的随机森林模型，通过改进减少特征维数的方式和将遗传算法的理论应用于参数调优，提出一种改进的随机森林多模型复合优化算法。

1 算法优化简介

基金项目：上海市 2017 年度“科技创新行动计划”基金资助项目（17401970900）。

作者简介：朱 城（1993-），男，硕士研究生，主要研究方向：大数据分析与应用；苏前敏（1974-），男，博士，副教授，主要研究方向：智能信息处理、大数据分析、软件工程。

收稿日期：2021-01-11

1.1 SelectKbest 变量筛算法

SelectKbest 变量筛算法是在 n 堆数据中寻求价值最优的 k 类数据[8]，每堆数据中有一定的特性 v_i 和 w_j，可令某组数据的特定值 $S = \{i_1, i_2, i_3, \cdots, i_k x\}$，表示为式(1)：

$$s(S) = \frac{\sum_{j=1}^{k} V_{ij}}{\sum_{j=1}^{k} W_{ij}}, \tag{1}$$

输入中包括 n 堆数据，并且含有 k 类需要保留的数据，符合条件为：$w_j(1 \leqslant k \leqslant n \leqslant 100\ 000)$，数据特性 v_i 符合条件为 $(0 \leqslant v_i \leqslant 10^6, 0 \leqslant w_j \leqslant 10^6)$，$v_i$ 和 w_j 的总和不超过10^7。

$$\frac{\sum v}{\sum w} = X_t, \tag{2}$$

$$\frac{k_1 + k_2 + \cdots k_k}{k} = X. \tag{3}$$

对比 X_t 的值，当所计算的值不再大于 X 时停止，而最终的 X_t 的值就是选取 k 类数据所计算出的值。

1.2 随机森林分类模型

随机森林以获取最后阶段的最优输出为目的，对数据集进行重复采样的优化模型。随机放回的抽样模式与传统不放回的样品抽样有相似点，但却并不独立，所以有式(4)：

$$E\left[\frac{\sum X_i}{K}\right] = E[X_i], \tag{4}$$

随机森林算法属于 bagging 算法的一种，也属于 bagging 算法的一种加强算法[9]，样本的数据集输入为式(5)：

$$E = \{(x_1 y_1), (x_2 y_2), \cdots, (x_m y_m)\}, \tag{5}$$

迭代次数为 t 次，即是对训练集进行 $t = 1, 2, \ldots, t$ 次分别采样，得到最终的集合 E_t，所得集合的算术平均值就是最后的模型输出。 随机森林建模过程如下：

$f(x) = A \quad f(y) = B$

For $i = 1$ to B

{ T_i = bootstrap sample from A

$C_x = \{T_i\}_1^B$

}

$$\mathrm{Exp}(x) = \frac{1}{B}\sum_{i=1}^{B} T_i(x)$$

Exp

输入训练集 A 和测试集 B，B 为储存样本。在 B 的样本量中选择一定量的特征因子，借助决策树来获取最合适的分割位置，并不断重复。将重复的结果存储到 Cx 中，在所有的结果中得到最终的样本预测值 Exp。

1.3 基于遗传算法的超参调优方法

本文以遗传函数的思想对参数进行优化，参数的优化问题可以定义为多目标优化问题，即可以用数学模型(6)规划。

$$\begin{cases} V - \min f(x) = [f_1(x), f_2(x), f_3(x), \cdots, f_p(x)]^{\mathrm{T}}, \\ \mathrm{s.t.}\ x \in X, \\ X \subseteq R^{\mathrm{m}}. \end{cases} \tag{6}$$

式中，$V - \min$ 表示向量极小化，即向量目标 $f(x)$ 中的各个子目标函数都尽可能极小化。

遗传函数的参数寻优首先就要进行编码，编码方式采用二进制，根据模型参数的类型选择染色体的长度，计算二进制所对应的十进制数[10]，式(7)：

$$x = \min + \frac{\text{十进制数} \times (\max - \min)}{2^{\text{染色体长度}} - 1}. \tag{7}$$

产生初始化群体，计算适应度即衡量交叉验证评价标准的值，进入繁衍迭代循环，根据不同的交叉和变异比例进行繁衍复制，若达到设定的迭代次数，则终止繁衍，否则继续迭代繁衍，在最终的“族群”中得出最优解。

2 实验与验证

2.1 数据集

本文所使用的实验数据来自威斯康星州诊断性乳腺癌数据库的乳腺癌数据集，在主流的癌症肿瘤智能医疗探索中，大多数研究是针对肿瘤的细胞核的特征进行分析，包括乳腺癌细胞核光滑性、凹陷点数标准差等数十个量化特征因子[11]。

癌症肿瘤数据之所以难以分析，是由于每一个特征连续型变量的变化都与其它的很多指标相联系，一个特征的变化又会导致更多指标发生变化。在对任何一个数据集进行分析处理的时候，过多的特征因子很容易降低模型的运行效率，而且很容易导致过拟合[12]，每一个变量又是连续变化的，所以所得出模型的结果也会随之变化。

2.2 特征因子的筛选

通过 spss 工具对乳腺癌数据集进行因子分析。对数据集进行 KMO 检验和巴特利特检验，KMO 检验是用于比较变量相关系数和偏相关系数的指标，KMO 值越接近于 1，证明变量间的相关性越强。巴

特利特球形检验是一种检验各个变量之间相关性程度的检验方法。一般在做因子分析之前都要进行巴特利特球形检验,用于判断变量是否适合用于做因子分析。通过检验乳腺癌数据集变量的 *KMO* 值为 0.832,说明 30 个变量间共线性很强,见表 1。

表 1 本文使用数据集 KMO 和巴特比特检验

Tab. 1 The data set KMO and Butterbite test used in this article

KMO 取样适合切性量数	0.832
巴特利特球形度检验近似卡方	39 362.121
巴特利特球形度检验自由度	435
巴特利特球形度检验显著性	0.000

在碎石图 1 中可以清楚地看到 30 个变量因子的特征值变化趋势。根据已设定的特征值提取参数,最终提取了 6 个特征值大于 1 的特征变量,对 6 个特征因子特征值进行旋转后,方差和特征根发生变化,这 6 个因子累计可以解释 88.759%的方差因素,解释力度非常强。

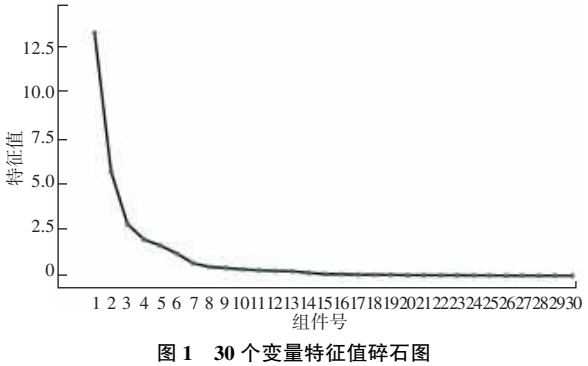

图 1 30 个变量特征值碎石图

Fig. 1 Crushed stone map of 30 variable eigenvalues

为了提高模型的表现力,对数据集进行数据筛选,找到能够达到模型效果最大化的变量筛选方法和筛选方案。在选择变量筛选方法时,需要对变量筛选的方法充分理解,对数据集样本的大小、实现的难易程度有所评估[13]。现在常用的变量筛选方法有比例法、方差法、SelectKbest 变量筛选法和模型筛选法,见表 2。针对乳腺癌数据集变量间存在的特殊性共线性关系,选用比例法和方差法时,模型得分反而降低。而选用 SelectKbest 变量筛选法可以自主控制想保留的变量因子个数,通过穷举法判别保留不同的变量因子个数时模型的表现性,找到变量筛选的最佳参数。当保留的变量因子从 1~6 个逐渐上升时,模型的得分呈现大幅度提高,符合因子分析的结果,通过调节 *K* 值的参数分析得出,保留 15 个变量因子时,模型得分超过不进行变量筛选时的模型得分,而当保留 16 个变量因子时,模型的得分达到最高,如图 2 所示。

表 2 常见变量筛选与无特征筛选乳腺癌分类器模型得分

Tab. 2 Common variable screening methods and five-feature screening of breast cancer classifier model scores

分类器模型	得分
无特征筛选的乳腺癌分类器模型得分	0.956 378
模型法筛选后乳腺癌分类器模型得分	0.947 368 4
比例法筛选后乳腺癌分类器模型得分	0.929 824 5
方差法筛选后乳腺癌分类器模型得分	0.947 255 3

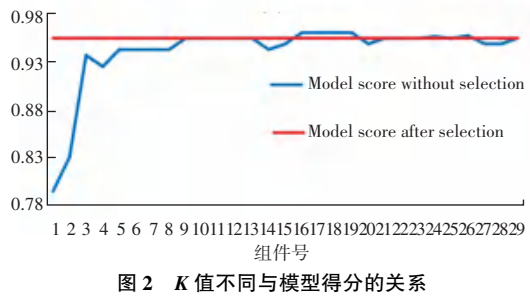

图 2 *K* 值不同与模型得分的关系

Fig. 2 The relationship between different *K* values and model scores

2.3 参数调优

随机森林分类器的结果作为输入值送入遗传算法超参数调优的模型中,首先产生足够数量染色体的种群,在本文中,一条染色体代表一组超参数,而每组超参数中的任意一个参数就是染色体上的基因。通过定义交叉验证的评价指标,计算超参数优化的适应性函数。整个迭代过程遗传算法主要运用两种方式来创建新一代,一是交叉的方式,二是变异的方式。通过改变随机森林中的树木数量和深度及叶子数在每一次迭代中删除表现型最差的参数组合,选择表现最好的机组参数组合进入下一次迭代,经过足够的迭代次数后,最终选择出最优的参数组合[14]。

通过遗传函数对随机森林的参数进行优化,流程如图 3 所示。随机生成一个基因序列产生第一个染色体,再生成一个基因序列,产生第二个染色体,重复到生成指定个染色体之后进行交叉和变异;对结果进行评估,选择最优的几个染色体进行下一次的迭代;重复进行交叉和变异,达到迭代次数后停止计算。设定迭代次数为 10 次,在第六次迭代之后就达到了模型的最大得分,第七次迭代之后模型得分不再变化。遗传函数 10 次迭代的模型最终参数设置如下:

input_matrix,

booststrap = *gini*

max_features = 0.1

min_samples_leaf = 1

min_samples_spit = 2

n_estimators = 100

遗传函数 10 次迭代的模型最佳得分见表 3。为了展示实验结果的科学性，分别与单个参数网格调参、多参数网格调参、随机网格调参的方式对精确率、召回率、*F*1 分值、*AUC* 值进行比较。见表 4 经过结果对比发现，通过遗传函数对模型进行优化的结果在交叉验证精准值、召回率、*F*1 分值、*AUC* 值等方面的准确率均超出使用单个网格、多个网格及随机网格等调参方法对模型进行优化的结果。

表 3　遗传函数 10 次迭代模型最佳得分

Tab. 3　The best model for ten iterations of genetic function

迭代次数	最佳得分
1	0.981 667 349 524 921 4
2	0.981 737 685 789 065 3
3	0.982 468 567 953 696 4
4	0.982 668 056 713 928 3
5	0.983 502 919 099 249 5
6	0.983 502 919 099 249 5
7	0.983 502 919 099 249 5
8	0.983 502 919 099 249 5
9	0.983 502 919 099 249 5
10	0.983 502 919 099 249 5

表 4　不同的参数调优性能指数

Tab. 4　Different parameter tuning performance indexes

参数调参	精准率	召回率	*F*1 分值	*AUC* 值
遗传函数	0.980 264 900 662	0.961 558 441 558	0.960 819 672 131	0.969 020 709 02
单个网格	0.951 454 563 345	0.932 476 037 408	0.928 314 539 721	0.921 780 372 48
多网格调参	0.963 216 150 387	0.931 421 110 850	0.929 104 883 018	0.923 572 847 50
随机网格调参	0.963 216 837 057	0.931 422 490 074	0.929 084 752 859	0.923 470 750 26

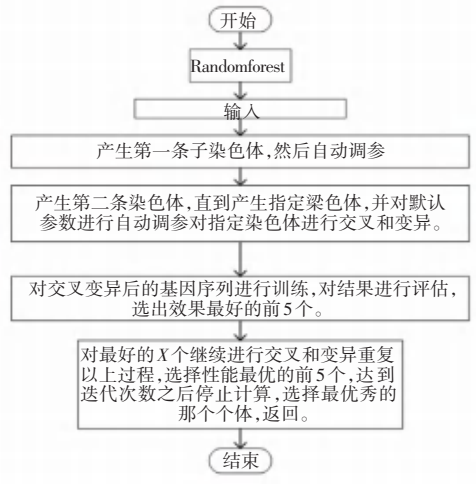

图 3　参数调优流程图

Fig. 3　Parameter tuning flowchart

根据精准率和召回率分别构建以遗传函数进行参数优化和以多网格调参优化后的模型 PR 曲线图和两个模型最后的 ROC 曲线如图 4 所示，可以看到遗传函数参数优化后的模型的 P-R 曲线完全将多网格调参后的模型 P-R 曲线覆盖，而从 ROC 曲线图来看，遗传函数参数优化后的模型所表现的 ROC 曲线面积更大。由此分析可以得出，通过遗传函数进行参数优化后的模型性能要高于多网格调参后的模型性能。

对模型进行 KS 检验，结果得出乳腺癌分类器的 KS 值达到 0.85，如图 5 所示，证明模型鉴别癌细胞与正常细胞的能力很强。

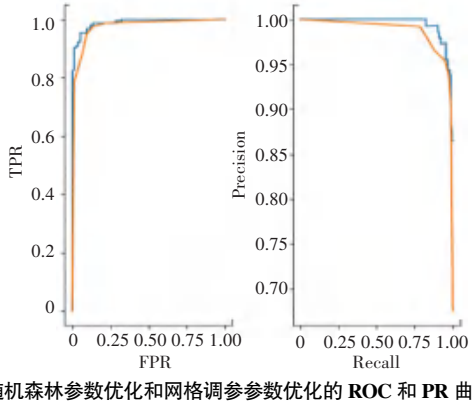

图 4　随机森林参数优化和网格调参参数优化的 ROC 和 PR 曲线对比

Fig. 4　Comparison of ROC and PR curves of random forest parameter optimization and grid parameter optimization

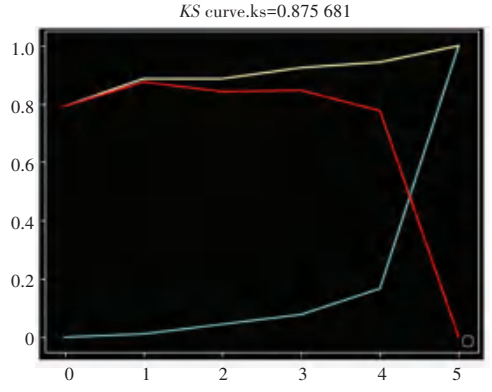

图 5　模型的 KS 验证

Fig. 5　KS verification of the model

3　结束语

(1)对于一般的医学临床数据集，由于特征因子之间共线性较强，通过试验发现选用比例法和方

差法以及模型验证法时,模型得分反而降低。而选用 Kbest 变量筛选法可以自主控制想保留的变量因子个数,减少了因子间共线性强对模型的影响,通过对比不同特征因子模型的表现性的变化来筛选特征,找到变量筛选的最佳参数;

(2)采用遗传算法对随机森林分类模型进行超参数调优后,目标模型的精准率达到 0.980 2,相比于运用多网格的调参方法提升了 0.17,不仅提高了模型的预测准确度,从模型的召回率、*F*1 分值、*AUC* 值进行比较分析,也提升了模型的鲁棒性。

优化后的随机森林算法模型能够最大限度程度上让模型的表现性达到最大,便于对强共线性特征数据的分析,优化参数寻优结构,为临床数据处理和疾病预测提供了一种新思路。

参考文献

[1] 李晓洁,丛亚丽. 从“谷歌流感趋势”预测谈健康医疗大数据伦理[J]. 医学与哲学,2019,40(14):5-8.

[2]俞碧莹.人工智能下的医学的发展应用[J]. 中国多媒体与网络教学学报(上旬刊),2019(10):28-29.

[3]施慧琳,苏燕,许丽,王玥.高通量测序行业现状与发展趋势分析[J]. 生物产业技术,2018(3):6-12.

[4] 郑雅文. 基于特征选择和支持向量机的乳腺癌诊断研究[D]. 太原: 太原理工大学,2019.

[5] 翁天乐.决策树机器学习算法在乳腺癌诊断中的应用[J]. 通讯世界,2018(10):224-226.

[6] 宋祖玲,刁莎,严兰平,等. 联合决策树及 logistic 回归建立乳腺癌相对风险预测模型[J]. 现代预防医学,2019,46(7):1156-1160,1175.

[7] 全雪峰. 基于随机森林的乳腺癌计算机辅助诊断[J]. 软件,2017,38(3):57-59.

[8] 杨佳琳,仝怡. 一种基于 MIMO 的改进型信号检测 K-Best 算法[J]. 现代导航,2017,8(2):142-146.

[9] 黄光成,周良,石建伟,等. 机器学习算法在疾病风险预测中的应用与比较[J]. 中国卫生资源,2020,23(4):432-436.

[10]莫鸿强. 遗传算法搜索能力和编码方式研究[D]. 广州: 华南理工大学,2001.

[11]郑存芳,洪文学,王金甲.基于偏序结构图的乳腺癌诊断规则提取方法[J]. 计算机工程与设计,2016,37(6):1599-1603,1686.

[12]闫云凤. 基于决策森林的回归模型方法研究及应用[D]. 杭州: 浙江大学,2019.

[13]郭慧玲,曾辉,闫柏屹,等. 常用主变量筛选方法及其应用特性分析[J]. 江西中医学院学报,2013,25(5):50-52.

[14]郦少将. 基于改进遗传算法的 HBV 水文模型参数优化[D]. 郑州: 华北水利水电大学,2018.

(上接第 129 页)

[2] SPAAN M T J, VEIGA T S, LIMA P U. Decision-theoretic planning under uncertainty with information rewards for active cooperative perception[J]. Autonomous Agents and Multi-Agent Systems, 2015, 29(6):1157-1185.

[3] 马庆平. 多 PTZ 主动摄像头的类目标检测定位系统[D]. 成都: 电子科技大学, 2014.

[4] SONG Bi, DING Chong, KAMAL A T, et al. Distributed camera networks[J]. IEEE Signal Processing Magazine, 2011, 28(3): 20-31.

[5] 王洪亮. 基于多摄像机协同的飞机着陆过程视频监测方法[D]. 哈尔滨:哈尔滨工业大学, 2019.

[6] YANG Jianbo, LIU Jun, WANG Jin, et al. Belief rule-base inference methodology using the evidential reasoning approach-RIMER [J]. IEEE Transactions on systems, Man, and Cybernetics-part A: Systems and Humans, 2006, 36(2):266-285.

[7] YANG Jianbo, LIU Jun, XU Dongling, et al. Optimization models for training belief-rule-based systems [J]. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 2007, 37(4):569-585.

[8] 周志杰,唐帅文,胡昌华,等. 证据推理理论及其应用[J]. 自动化学报, 2021, 47(5): 970-984.

[9] YANG Longhao, WANG Yingming, FU Yanggeng. A consistency analysis-based rule activation method for extended belief-rule-based systems[J]. Information Sciences, 2018, 445:50-65.

[10] CHANG Leilei, ZHOU Zhijie, CHEN Yuwang, et al. Belief rule base structure and parameter joint optimization under disjunctive assumption for nonlinear complex system modeling [J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2018, 48(9): 1542-1554.

[11] ZHOU Zhijie, CAO You, HU Guanyu, et al. New health-state assessment model based on belief rule base with interpretability [J]. Science China-Information Sciences, 2021, 64(7):172214.

[12] LI Bin, WANG Hongwei, YANG Jianbo, et al. A belief-rule-based inference method for aggregate production planning under uncertainty [J]. International Journal of Production Research, 2013, 51(1):83-105.

[13] KONG Gailing, JIANG Zhijie, YIN Xiaofei, et al. Combining principal component analysis and the evidential reasoning approach for healthcare quality assessment [J]. Annals of operations research, 2018, 271(2): 679-699.

[14] LI Gailing, ZHOU Zhijie, HU Changhua, et al. A new safety assessment model for complex system based on the conditional generalized minimum variance and the belief rule base[J]. Safety Science, 2017, 93:108-120.

[15] ZHOU Zhijie, HU Guanyu, ZHANG Bangdong, et al. A model for hidden behavior prediction of complex systems based on belief rule base and power set[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2018, 48(9):1649-1655.