鲁东大学学报(自然科学版)
Journal of Ludong University(Natural Science Edition) 2023,39(2):159—164

Doi:10.20062/j.cnki.CN 37-1453/N.2023.02.009

基于随机森林模型和遗传算法对抗乳腺癌药物的优化研究

任静莹,马成满,毕四旭,邵喜高

(鲁东大学 数学与统计科学学院,山东 烟台 264039)

摘要:乳腺癌是世界上常见且致死率高的癌症。本文在充分考虑化合物分子描述符之间的非线性关系的同时,建立了随机森林模型,对化合物的生物活性进行定量预测。为寻找最优分子描述符的取值,在轮盘赌策略的基础上采用遗传算法,对 ADMET 性质进行分类预测,通过预测结果提升拮抗剂生物活性的预测效率。研究结果表明:所建立的随机森林模型预测精度高,模型参考价值得到有效提升;通过多次迭代遗传算法,能够准确找到因变量的最优值,为抗乳腺癌药物的研究提供数据支撑和理论参考。

关键词:随机森林;遗传算法;乳腺癌药物;生物活性

中图分类号:R979.1 **文献标志码**:A **文章编号**:1673-8020(2023)02-0159-06

据 2018 年国际癌症研究机构(IARC)调查的最新数据显示,乳腺癌在全球女性癌症中的发病率为 24.2%,位居女性癌症首位。当前对乳腺癌的医学诊断并没有彻底遏制乳腺癌的复发与转移[1],甚至常引起患者心血管损伤、免疫力低下、药物敏感性降低等诸多不良反应,因此,筛选安全高效的天然抗乳腺癌活性成分对于抗乳腺癌药物的研发有着极其重要的作用[2]。

对抗乳腺癌药物生物活性的定量预测以及找到最优的分子描述符的取值,能够更好地帮助研究人员对药物展开研究,从而制定更有效的治疗方案[3]。李莉等[4]应用了支持向量机和人工神经网络算法进行预测,预测结果较接近实际值;李勇等[5]提出基于 C-AdaBoost 模型的集成学习算法,发现了判断乳腺癌是否复发、乳腺癌肿瘤是否为良性的最优特征组合,比机器学习分类器的准确率高 19.5%;王悦等[6]利用乳腺癌临床数据集构建 LightGBM 预测模型,模型准确率高达 97.14%;赖胜圣等[7]构建了基于 SFS-SVM 的乳腺癌预测模型,相对于单独 SVM 算法能够较好地对乳腺癌作出辅助治疗;李宁等[8]基于 2013 版超声乳腺影响报告系统(BI-RADA)对乳腺癌发病因素进行数据分析,所形成的 Logistic 回归模型能够筛查出更多具有治疗价值的指标;沈倩倩等[9]基于 XGBoost 算法构建乳腺癌预测模型,准确率达到 97.86%,且获得较高的 AUC 值;殷恺铭等[10]基于 LTP 算法提取的新型纹理特征预测精度较高,与常规特征融合后可进一步提高预测效能;董华等[11]通过机器学习建立三阴乳腺癌预测模型,预测精度超过 95.5%,并通过支持向量机特征消除算法使模型准确率达到 97.8%。这些成果利用机器学习、深度学习等算法进行研究,但算法的计算复杂度高,且对模型的预测精度一般较低。

本文根据数据之间的非线性相关性特点,用决策树、随机森林、KNN 三种模型弥补多元线性回归预测中的不足,选择均方误差最小的随机森林模型对生物活性进行预测,基于遗传算法寻找使得生物活性和 ADMET 性质均良好的分子描述符取值。

1 变量筛选与降维处理

为了研究乳腺癌治疗靶标 ERα,本文对临床试验得到的 1974 个化合物的 729 个分子描述符数据进行变量筛选。首先,对数据进行归一化处理,消除量纲影响;其次,用低方差滤波、高相关滤

收稿日期:2021-12-29;修回日期:2022-12-15

基金项目:山东省社会科学规划研究项目(20CSDJ10)

通信作者简介:邵喜高(1978—),男,讲师,硕士研究生导师,博士,研究方向为应用统计、纵向数据分析。E-mail:shaoxg1011@163.com

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

波对变量进行降维,剔除无意义及相关性强的变量;最后使用随机森林模型对剩余变量进行降维,以基尼系数作为特征重要性评估和分析的准则,筛选出对生物活性影响显著的 10 个分子描述符。

2 化合物对 ERα 生物活性的定量预测模型

目前在乳腺癌药物研发中,基于乳腺癌相关的靶标(本文选取 ERα)收集作用于该靶标的化合物及其生物活性数据,创建化合物的定量结构,建立化合物活性预测模型——活性关系模型;在高效率和低成本的情况下,预测具有更好生物活性的新化合物分子,或者指导已有活性化合物的结构优化[12]。

首先假设化合物的生物活性值与一系列分子结构描述符呈线性关系,得到多元线性回归模型为:

$$\hat{y} = 6.02 + 0.92x_1 - 0.05x_2 - 1.49x_3 + 1.09x_4 + 1.21x_5 - 0.99x_6 + 0.25x_7 + 0.89x_8 + 0.56x_9 + 1.21x_{10}, \quad (1)$$

其中,$x_i(i = 1,2,\cdots,10)$ 为排名位于前 10 的分子描述符。假设某数据集包含观察值 $y_1,y_2,\cdots,y_n$,相对应的模型预测值分别为 $f_1,f_2,\cdots,f_n$,定义残差 $e_i = y_i - f_i$,则平均观察值为:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i。 \quad (2)$$

于是得到总平方和为:

$$S_{\text{tot}} = \sum_{i=1}^{n} (y_i - \bar{y})^2, \quad (3)$$

回归平方和与残差平方和分别为:

$$S_{\text{reg}} = \sum_{i=1}^{n} (f_i - \bar{y})^2, \quad (4)$$

$$S_{\text{res}} = \sum_{i=1}^{n} (y_i - f_i)^2。 \quad (5)$$

由此,判定系数定义为:

$$R^2 = 1 - \frac{S_{\text{res}}}{S_{\text{tot}}}。 \quad (6)$$

经计算,多元线性回归模型的判定系数 $R^2 = 0.372\,4$,模型拟合效果差,说明线性预测模型不符合本数据的特点,无法解决 ERα 生物活性预测问题。因此,为适应数据变化且提高精度,下面建立非线性定量预测模型。

2.1 非线性定量预测模型

2.1.1 决策树模型

决策树模型是常见的分类模型,通过叶节点的推理原则实现对新数据的分类或回归预测,从而判定目标所属类别。常见的决策树节点划分依据为 ID3 决策树、C4.5 决策树、CART 决策树,本文采用 CART 决策树。

对于回归树模型,叶节点输出变量的均值与式(2)一致,其输出变量为数值型,异质性度量通常采用方差。为方便与其他模型对比,此处选用均方误差 $H(y)$ 作为度量指标,其数学表达式为:

$$H(y) = \frac{1}{N}\sum_{i\in N} (y - \bar{y})^2, \quad (7)$$

其中,N 为当前节点所有训练样本个数,$\bar{y}$ 为当前节点样本变量的平均值。

对于由排名前 10 的分子描述符构成的数据集,调节决策树模型进行初步拟合。通过网格搜索法得到,当决策树最大深度为 18,决策树节点继续分割的最小样本量为 8,每棵决策树叶节点的最小样本量为 6 时,模型最优,此时均方误差 $H(y) = 0.969$。

2.1.2 随机森林模型

本节将多个 CART 决策树进行整合,建立多层次的随机森林模型。由于模型是多层次、多维度的,因此将面临如何选择切分变量、切分点的问题。为得到最优切分变量和切分点,采用穷举法来研究每个特征及其取值;同时,以切分后节点的不纯度来衡量切分变量和切分点的好坏[13],得到各个子节点不纯度的加权和 $G(x_i,v_{ij})$,其计算公式为:

$$G(x_i,v_{ij}) = \frac{n_1}{N_s}H(X_1) + \frac{n_r}{N_s}H(X_r), \quad (8)$$

其中:x_i,v_{ij} 分别表示第 i 个切分变量及其切分值;n_1,n_r 分别为切分后左子节点、右子节点的训练样本个数,N_s 代表当前节点所有训练样本个数;X_1,X_r 分别为左、右子节点的训练样本集合;$H(X)$ 为衡量节点不纯度 X 的函数。由于本文主要解决回归问题,因此采用均方误差作为不纯度函数,其计算见式(8)。

与其他分类算法相比,随机森林同时生成多个预测模型,通过汇总模型的结果使得分类准确率通常更高[14]。在此模型中,设定包含 200 个决策树,均方误差为 0.611,按重要性对 10 个分子

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

结构描述符进行排序,结果见图 1。由图 1 得到,minsssN 对于 ERα 生物活性的影响最大。

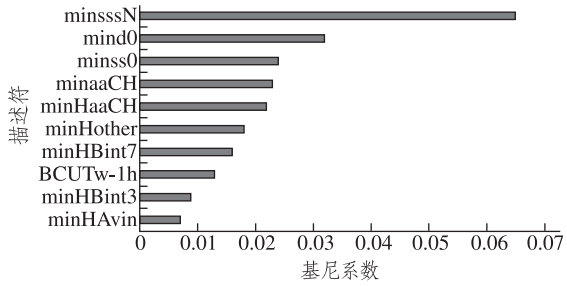

图 1　分子结构描述符的重要性排序

Fig.1 Order of importance of molecular structure descriptors

2.1.3　KNN 预测模型

KNN 算法属于惰性算法,不会预先生成一个预测或分类模型,而是将模型的构建与预测同时进行。对于连续性的因变量预测,使用 K 个已知的样本均值预测未知样本。

将训练集合的 N 个样本观测设为 p 维,根据 X_0 的 K 个近邻的样本 $y_1, y_2, \cdots, y_k$ 计算 $\hat{y}_0$。即:

$$\hat{y}_0 = \frac{1}{K} \sum_{X_i \in N_K(X_0)} y_i, \tag{9}$$

式中,$N_K(X_0)$ 表示 X_0 的 K 个近邻集合。本节选择均方误差作为度量原则,针对排名前 10 的分子描述符构成的数据集,通过研究得到不同 K 值对应的均方误差,结果见图 2。

由图 2 可以看到,随着 K 值的增加,均方误差逐渐趋于收敛。通过比较,选取最小均方误差对应的 K 值作为最佳近邻个数,得出最佳近邻个数 $K = 10$。进一步对化合物的生物活性进行KNN 预测,其均方误差为 0.720。

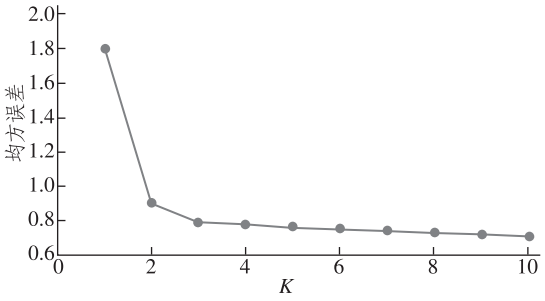

图 2　不同 K 值对应的均方误差

Fig.2 Mean squared error for different K values

2.2　模型对比

将三种模型的均方误差进行汇总,得到表 1。从表 1 可以看出,决策树模型的均方误差为 0.969,随机森林模型的均方误差为 0.611,KNN 预测模型的均方误差为 0.720,说明随机森林模型的预测性能优于 KNN 和决策树。接下来,采用随机森林模型预测 ERα 生物活性,部分预测结果见表 2。

表 1　三种模型对应的均方误差

Tab.1 Mean squared errors corresponding to the three models

模型	均方误差
决策树模型(h=18,n=6)	0.969
随机森林模型(i=200)	0.611
KNN 预测模型(K=10)	0.720

注:h 表示最大深度,n 表示叶节点的最小样本量,i 表示决策树个数。

表 2　部分生物活性预测结果

Tab.2 Partial biological activity prediction results

nM

化合物结构式	y_i	y_p
CC(=O)\C=C\c1ccc(cc1)C2=C(CCOc3cc(F)ccc23)c4ccc(O)cc4	699.429	6.155
OC(=O)\C=C\c1ccc(cc1)C2=C(CCOc3ccccc23)c4ccc(O)cc4	626.515	6.203
COc1ccc2C(=C(CCOc2c1)c3ccc(O)cc3)c4ccc(\C=C\C(=O)O)cc4	780.637	6.108
OC(=O)\C=C\c1ccc(cc1)C2=C(CCOc3cc(F)ccc23)c4ccc(O)cc4	684.851	6.164
OC(=O)\C=C\c1ccc(cc1)C2=C(CCSc3cc(F)ccc23)c4ccc(O)cc4	3 082.950	5.511

在表 2 中:y_i 是化合物对 ERα 的生物活性值,y_i 值越小代表生物活性越大,说明该化合物对抑制 ERα 活性越有效;y_p 是由 y_i 取负对数转化来的,y_p 值越大说明生物活性越高。因此,表 2 结果显示,这几种化合物的生物活性较低,对 ERα 的抑制效果差。

3　ADMET 性质的分类预测模型

3.1　多目标函数模型建立

合格的候选药物不仅具备良好的内在生物活

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

性,还需要较强的外在响应能力,例如人体吸收快、代谢速度适中、毒性小等,同时受试者服用后依旧具备良好的药代动力学性质和安全性,这些性质统称为 ADMET 性质[12],可以检测化合物是否具有遗传毒性。本文在进行药物选择时,需要充分考虑 ADMET 性质,并对其进行优化。

首先建立关于生物活性以及 5 种 ADMET 性质的多目标函数。将 10 个分子描述符的取值范围[0,1]代入随机森林预测模型,并将 y_i 和 5 种性质的取值代入遗传算法进行计算分析,输出最优解;基于轮盘赌的选择策略优化遗传算法[15],使得个体被选中的概率与其适应度成正比。具体操作步骤如下:

1) 计算个体的适应度 $f(x_i)(i=1,2,\cdots,M)$,M 为个体数量;

2) 计算个体被遗传到下一代群体中的概率,即

$$P(x_i)=\frac{f(x_i)}{\sum_{j=1}^{i} f(x_j)}; \tag{10}$$

3) 计算个体的累积概率:

$$q_i=\sum_{j=1}^{i} P(x_j); \tag{11}$$

4) 在区间[0,1]内产生一个均匀分布的伪随机数 r;

5) 若 $r<q_1$,则选择个体 1;否则,选择个体 k 使得 $q_{k-1}<r$ 成立;

6) 重复步骤 4)、5) M 次。

为进行遗传算法的运算,本文取 $M=20$,迭代次数为 80,变量上、下限分别为 0 和 1。

3.2 多目标函数模型求解

遗传算法是通过模拟生物进化染色体上基因的交叉、变异等随机性过程,采用随机方法[15]对问题进行求解和优化的算法。本文针对多目标求最优解的问题,找到使生物活性和 ADMET 性质同时满足条件的分子描述符的取值,利用遗传算法本身的高效性、快速性及高容错性等特点获得优化结果[16]。

创建 $M=20$ 的初始种群进行适应度计算,如果个体的累积概率大于生成的伪随机数,则以最优个体输出;否则,进行交叉、变异操作,继续进行适应度计算[17]。循环上述操作,具体流程图见图 3。

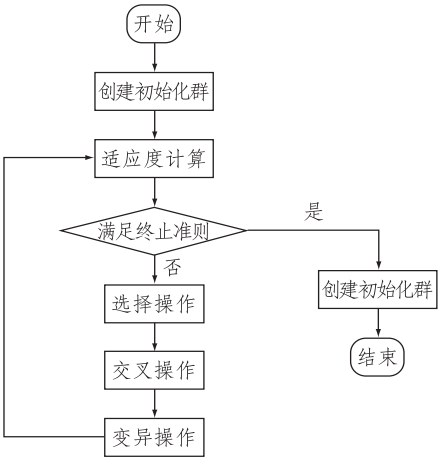

图 3 遗传算法流程图

Fig.3 Graph of genetic algorithm flow

利用包含 5 种 ADMET 性质的数据集进行遗传算法预测研究,遗传算法迭代与误差收敛结果见图 4。由图 4 可以看出,随着迭代次数的不断增加,当迭代到第 10 次时,算法逐渐收敛于最优解。因此,本文采用第 10 次的迭代结果预测分子描述符的取值。

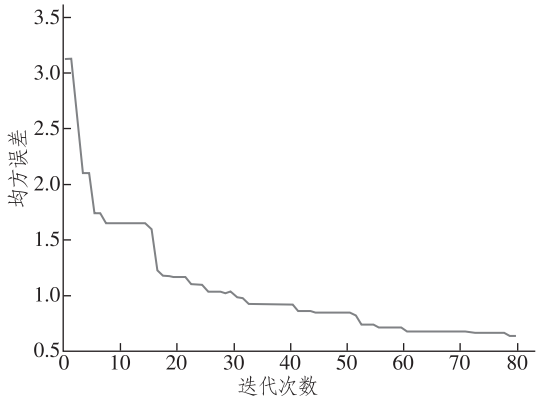

图 4 遗传算法迭代与误差收敛

Fig.4 Genetic algorithm iteration and error convergence

对抑制 ERα 活性排名前 10 的分子描述符进行预测,使得化合物具有更好生物活性的同时具有更好的 ADMET 性质。通过分析得到分子描述符的最优取值,结果见表 3。

化合物的分子描述符是一系列用于描述化合物的结构和性质特征的参数,其中包括物理化学性质(如分子量)、拓扑结构特征(如氢键供体数量、氢键受体数量)等。由表 3 得到,当它们取值分别为-0.424、0.319、0.321、1.531、0.494 时,可以保证化合物具有更好的生物活性和 ADMET 性质[18]。

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

表 3　分子描述符最优取值

Tab.3 Optimal values of molecular descriptor

描述符	描述符说明	数值
minHBint7	Minimum E-State descriptors of strength for potential Hydrogen Bonds of path length 7	-0.424
minHaaCH	Minimum atom-type H E-State:;CH:	0.319
minHother	Minimum atom-type H E-State:H on aaCH,dCH2 or dsCH	0.321
minaaCH	Minimum atom-type E-State:;CH:	1.531
minsssN	Minimum atom-type E-State:>N-	0.494

4　结论

本文以均方误差作为衡量模型优良的准则,通过建立非线性定量预测模型对 ERα 生物活性进行预测。结果表明,随机森林模型的预测性能优于 KNN 算法和决策树模型,同时得到生物活性较低的几种化合物。此外,对于多目标函数求最优解的问题建立遗传算法,预测抑制 ERα 活性排名前 10 的分子描述符,使得化合物同时具有更好的生物活性和 ADMET 性质。

参考文献:

[1] 孙少康,黄勇,李志明,等.生物活性多糖抗乳腺癌作用研究进展[J].世界中医药,2021,16(18):2798-2805.

[2] 郑莹.中国乳腺癌患者生活方式指南[J].全科医学临床与教育,2017,15(2):124-128.

[3] 梁永琴,杨喜花,赵莉莉,等.苏木抗乳腺癌活性成分的筛选及其作用效果研究[J].山西大学学报(自然科学版),2022,45(2):465-472.

[4] 李莉,汪咏,陆宁,等.基于多分类算法混合比较的乳腺癌预测[J].控制理论与应用,2021,38(10):1503-1510.

[5] 李勇,陈思萱,贾海,等.基于 C-AdaBoost 模型的乳腺癌预测研究[J].计算机工程与科学,2020,42(8):1414-1422.

[6] 王悦,王延博,王辛格.基于 LightGBM 的乳腺癌预测模型[J].智慧健康,2019,5(29):39-41.

[7] 赖胜圣,刘虔铖,余丽玲,等.基于 SFS-SVM 的乳腺癌预测模型的构建[J].中国医学物理学杂志,2019,36(7):826-829.

[8] 李宁,周伟.基于 2013 版超声乳腺影像报告数据系统的乳腺癌预测因素 Logistic 回归分析[J].中国妇幼保健,2019,34(14):3321-3324.

[9] 沈倩倩,邵峰晶,孙仁诚.基于 XGBoost 的乳腺癌预测模型[J].青岛大学学报(自然科学版),2019,32(1):95-100.

[10] 殷恺铭,闫士举,宋成利.基于改进局部三元模式的乳腺癌预测模型[J].中国医学影像技术,2018,34(4):616-620.

[11] 董华,马岚.基于机器学习的三阴乳腺癌预测模型[J].云南大学学报(自然科学版),2017,39(S1):111-115.

[12] 卢皎玲,谢沁沁.基于 K-MEANS 算法的抗乳腺癌候选药物 ERα 活性优化研究[J].信息技术与信息化,2021(12):45-48.

[13] 高东岳.基于机器学习方法的超声导波结构健康监测研究[J].纤维复合材料,2020,37(3):3-8.

[14] 冯晓荣,瞿国庆.基于深度学习与随机森林的高维数据特征选择[J].计算机工程与设计,2019,40(9):2494-2501.

[15] 魏士伟,邓维.基于多精英协同进化遗传算法的云资源调度[J].计算机应用与软件,2021,38(5):274-280.

[16] HAMDIA K M,ZHUANG X,TIMON R. An efficient optimization approach for designing machine learning models based on genetic algorithm [J]. Neural Computing and Applications, 2021, 33 (6): 1923-1933.

[17] CHEN Y,ZHENG W,LI W,et al.Large group activity security risk assessment and risk early warning based on random forest algorithm [J]. Pattern Recognition Letters,2021,144:1-5.

[18] 佘为为,黄清荣,李延敏,等.MicroRNAs 在乳腺癌发生、发展和转移中的作用[J].鲁东大学学报(自然科学版),2020,36(4):338-352.

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

Optimization of Anti-breast Cancer Drugs Based on Random Forest Model and Genetic Algorithm

REN Jingying, MA Chengman, BI Sixu, SHAO Xigao

(School of Mathematics and Statistics Science, Ludong University, Yantai 264039, China)

Abstract: Breast cancer is one of the most common and deadly cancers in the world. In this paper, a stochastic forest model is established to quantitatively predict the biological activity of compounds while considering the nonlinear relationship among molecular descriptors. In order to find the optimal value of molecular descriptors, the genetic algorithm was used to classify and predict the properties of ADMET based on the roulette strategy, which provides a prediction service for optimizing the biological activity of antagonists. The study result shows that: the established random forest model with appropriate has high prediction accuracy, and the reference value of the model is effectively improved; the optimal value of the dependent variable is found accurately through several iterations of the genetic algorithm, which provides theoretical reference and data support for the research of anti-breast cancer drugs.

Keywords: random forest; genetic algorithm; breast cancer; biological activity

(责任编辑 顾建忠)

(上接第158页)

Abstract ID: 1673-8020(2023)02-0153-EA

Security Portfolio Strategy Based on Economic Model Predictive Control

MA Xiaohan, LIU Xiaohua, GAO Rong

(School of Mathematics and Statistics Science, Ludong University, Yantai 264039, China)

Abstract: The purpose of security portfolio is to maximize returns and minimize risks. For investors with different risk attitudes, the problem of multi-objective security portfolio is studied by using economic model predictive control. Utopia-tracking method was introduced to solve this problem, and a multi-objective security portfolio strategy considering both return maximization and risk minimization under different risk attitudes is ensured. Finally, the effectiveness of the proposed strategy was verified by the simulation.

Keywords: multi-objective optimization; security portfolio; economic model predictive control(EMPC); Utopia-tracking method; risk aversion coefficient

(责任编辑 顾建忠)

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net