Modeling and Simulation 建模与仿真, 2023, 12(1), 130-139
Published Online January 2023 in Hans. https://www.hanspub.org/journal/mos
https://doi.org/10.12677/mos.2023.121013

基于数据挖掘的抗乳腺癌候选药物优化建模

严　俊，徐金府，刘天胤

上海理工大学机械工程学院，上海

收稿日期：2022年11月14日；录用日期：2023年1月9日；发布日期：2023年1月16日

摘　要

目前，治疗乳腺癌的候选药物是能够抗结ERα活性的化合物合成的，但由于化合物定量结构复杂、药代动力学性质(ADMET)不稳定，导致药物研发成本较高。本文通过相关性分析得出对ERα活性影响较高的20个分子描述符，并基于数据挖掘技术和机器学习算法，建立了相关化合物定量结构-ERα活性以及定量结构-ADMET性质的定量预测模型，对药物研发具有一定帮助。

关键词

ERα，相关性分析，多元回归分析，BP神经网络，随机森林

Optimal Modeling of Anti-Breast Cancer Drug Candidates Based on Data Mining

Jun Yan, Jinfu Xu, Tianyin Liu

School of Mechanical Engineering, University of Shanghai for Science and Technology, Shanghai

Received: Nov. 14th, 2022; accepted: Jan. 9th, 2023; published: Jan. 16th, 2023

Abstract

At present, compounds with anti-junction ERα activity are drug candidates for the treatment of breast cancer, but due to the complex quantitative structure of the compound and the unstable pharmacokinetic properties (ADMET), the drug development cost is high. In this paper, 20 molecular descriptors with high influence on ERα activity are obtained through correlation analysis, and based on data mining technology and machine learning algorithm, a quantitative prediction model of quantitative structure-ERα activity and quantitative structure-ADMET properties of related compounds is established, which is helpful for drug development.

文章引用: 严俊, 徐金府, 刘天胤. 基于数据挖掘的抗乳腺癌候选药物优化建模[J]. 建模与仿真, 2023, 12(1): 130-139.
DOI: 10.12677/mos.2023.121013

Keywords

ERα, Correlation Analysis, Multiple Regression Analysis, BP Neural Network, Random Forest

1. 引言

乳腺癌是目前女性常见三大癌症之一，据统计 2018 国内乳腺癌确诊人数占女性癌症确诊总人数的19.2%且乳腺癌确诊率呈现逐年上升趋势[1] [2]。根据癌细胞内蛋白分子的不同，乳腺癌可以分为雌激素受体、孕激素受体、人表皮生长因子-2 三类，其中，约 70%的乳腺癌患者表现为雌激素受体 α (Estrogen Receptorα, ERα)阳性[3] [4]。雌激素受体 α 是一种转录因子核受体，其活性主要通过与雌激素结合来调控，该受体的活性受到雌激素的影响，研究发现该受体长期与雌激素结合是乳腺癌产生的因素之一[5]。

目前，对于 ERα 表达的乳腺癌患者的常规治疗是采用抗激素疗法，主要是通过限制雌激素受体的活性，从而达到控制体内激素水平的目的。抑制 ERα 成为了治疗乳腺癌的重要手段，因此在选择治疗乳腺癌的临床药物上，能够拮抗 ERα 活性的化合物成为了首选。当下，在药物研发过程中，建立化合物预测模型成为了筛选化合物活性的主要方法。另外，化合物具备良好的药代动力学性质和安全性，合称为 ADMET 性质，只有具备良好生物活性和 ADMET 性质的化合物，才能成为候选药物。

本文旨在根据提供的 ERα 拮抗剂信息，通过机器学习方法构建化合物生物活性的定量预测模型和 ADMET 性质的分类预测模型，从而达到为优化 ERα 拮抗剂的生物活性定量预测和 ADMET 性质分类预测提供服务的目的。

2. 化合物对 ERα 生物活性的定量预测模型

2.1. 分子描述符筛选

本文采用了斯皮尔曼相关系数法评价化合物各分子描述符对生物活性的相关程度，以找出要求的前 20 个影响最为显著的分子描述符；同时使用基于 BP 神经网络的 MIV 平均影响值算法，将 729 个分子描述符数据作为输入，pIC50 预测值作为输出，计算得到各分子描述符所代表的相关系数，找出前 20 个影响最为显著的分子描述符。

Spearman 相关系数能够确切的表明两个变量之间相关的程度，所以化合物的分子描述符对其生物活性的显著影响可由相关系数表征。本文中化合物的分子描述符属于等级数据，独立变量 X 和依赖变量 Y 之间无明显的正态分布和线性关系，故 Spearman 相关系数方法是用来求解该问题的有效方法。

针对化合物的分子描述符，Spearman 相关系数可简化为：

$$\rho = 1 - \frac{6\sum d_i^2}{n\left(n^2 - 1\right)} \tag{1}$$

其中，d_i 为变量之间的等级差，一个数的等级，就是将其所在的一列数按照从小到大排序后这个数所在的位置，可以证明：ρ位于−1 和 1 之间。Spearman 相关系数的计算采用的方式不是取值本身，而是采用取值的等级。

平均影响值(MIV：mean impact value)用在神经网络中评价变量的相关性，该方法能够反映出自变量对输出神经元的影响[6]。本文将 MIV 结合 BP 神经网络，先训练好网络再将一个输入减少 10%和增加 10% (其它输入保持不变)，将两组数据都输入到网络中，查看该输入的落差会引起网络输出多少的落差。如果引起的落差很少，则说明该输入对输出影响很小，否则，则认为对输出影响较大。

BP 神经网络主要由三个层次组成，主要包含输入层，隐含层，输出层。首先数据通过输入层进入网络中，在隐含层进行网络处理，在输出层得到最后的结果。当训练得到的输出层的结果与预期的结果相差较大时，此时数据在神经网络中反向传播，在该阶段中通过调整网络权值，使得最终的输出结果与给定的预期结果满足一定的条件。本文建立的 BP 神经网络模型结构含有 729 个输入、1 层隐含层以及 1 个输出，用于分析化合物对 ERα 生物活性的影响。

输入输出层：将 729 个分子描述符数据作为输入，pIC50 预测值作为输出。故输入层神经元个数 $n = 729$，输出层神经元个数 $m = 1$。

隐层：确定神经元的个数是网格设计过程中重要的环节，当神经元个数过多的时候，网络计算量大，计算冗余，当网络神经元个数较少的时候，会影响网络的稳定性，计算误差大，无法达到预期的效果。隐含层网格神经元的个数主要与问题的复杂度和输出层预测结果与预期结果的误差条件有直接的关系。本文出于保证精度的考虑，经反复调试，最终确定隐层神经元个数为 9 个。神经网络结构示意如图 1 所示。

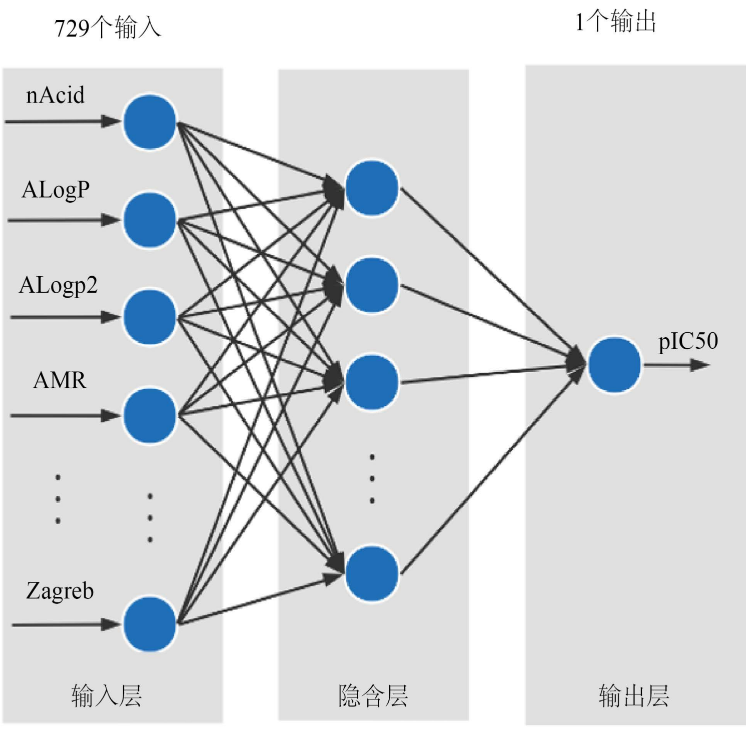

Figure 1. Schematic diagram of the neural network structure based on MIV BP
图 1. 基于 MIV 的 BP 的神经网络结构示意图

由 Spearman 相关系数得到的各个分子描述符对化合物生物活性的相关系数如图 2 所示。

由 MIV 计算得到的各个分子描述符对化合物生物活性的相关系数如图 3 所示。

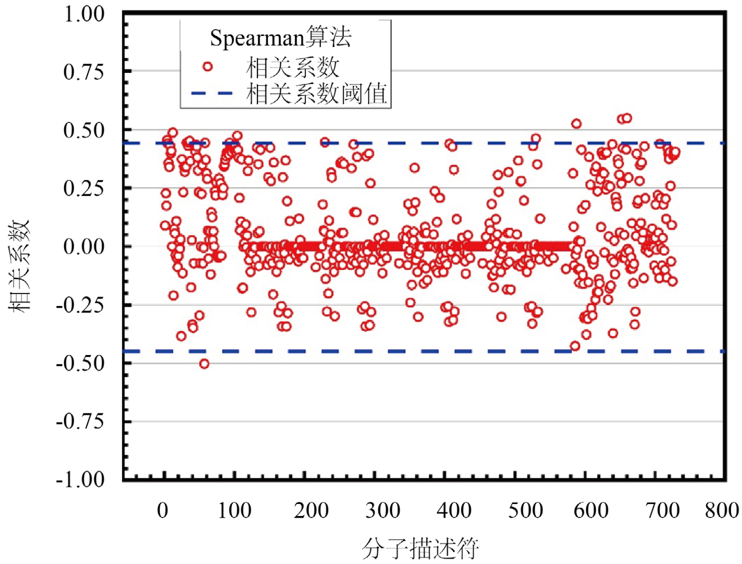

Figure 2. Schematic diagram of the distribution of the coefficients of correlation for calculating the molecular descriptor Spearman
图 2. 计算分子描述符 Spearman 相关系数分布示意图

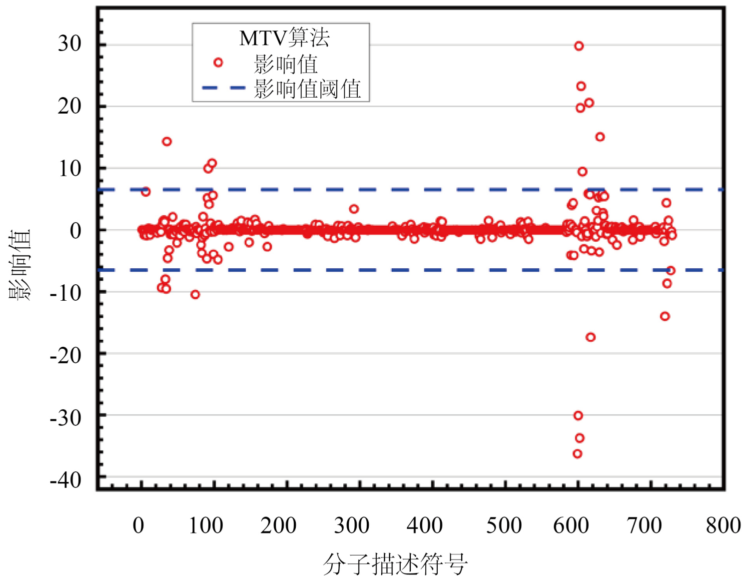

Figure 3. Schematic diagram of the effect of molecular descriptors on biological activity values for MIV calculations
图 3. MIV 计算分子描述符对生物活性影响值示意图

图中蓝色划线表示相关度阈值，对比 spearman 相关系数与 MIV 平均影响值算法得出的结果，MIV 计算分子描述符对生物活性影响值大多数都集中在 0 附近，说明大多数分子描述符对生物活性值影响不大，只有少部分对生物活性影响较大；spearman 相关系数计算得出的每一个分子描述符对生物活性影响较为离散，数据对比更加明显。本文认为相关系数的大小表征了分子描述符对生物活性的影响程度，因此选择对 spearman 相关系数计算出的结果进行排序，由此得到前 20 个对生物活性最具有显著影响的分子描述符及其相关系数，如表 1 所示。

Table 1. Significantly affects the table of correlation coefficients for molecular descriptors
表 1. 较显著影响分子描述符相关系数表

序号	前 20 个对生物活性最具有显著影响的分子描述符
1	MDEC-23
2	MLogP
3	LipoaffinityIndex
4	C1SP2
5	nC
6	CrippenLogP
7	maxsOH
8	AMR
9	ATSp5
10	SwHBa
11	ATSp4
12	ATSp2
13	ATSp1
14	C2SP2
15	SP-5
16	apol
17	minsssN
18	nT6Ring
19	fragC
20	SaaCH

2.2. 多元回归的预测模型

本文在选定好前 20 个对生物活性最具有显著影响的分子描述符的基础上，构建化合物对 ERα 生物活性的定量预测模型，通过模型对 IC50 值和对应的 pIC50 值进行预测。主要对 50 个化合物进行 pIC50 值预测，进而得到对应的 50 个 IC50 的预测值。

2.2.1. 数据处理

本文在建模中为了后续方便观察，先单独提取出这 20 个分子描述符所在的组对应的所有数据。由于这些数据都是化合物对 ERα 的生物活性采集的真实数据样本，所以这里不用对数据进行清洗，只需根据数据特征和后续各个模型的数据输入格式进行适当的分析和预处理即可，在此基础上对数据做了以下两个方面的处理：

1) 建立 IC50 与 pIC50 的负对数关系函数

本文已知 IC50 与 pIC50 的值成负对数关系，在现有的 1974 组 IC50 与对应的 pIC50 数值条件下，建立出 IC50 与 pIC50 的负对数关系函数，具体函数解析式如下所示：

$$y = -0.434\ln(x) + 9 \tag{2}$$

式中：x 为 IC50；y 为 pIC50。

2) 归一化处理

由于得到的二十组分子描述符数据之间、各组数据与 pIC50 数据的量纲以及数量级可能存在差异，这种差异可能会对模型结果造成影响，为消除数据之间的影响，本文将对所有输入进行归一化处理。当原始数据经过数据标准化处理之后，经过归一化处理之后数据的指标量级相同，对于这种量级相同的数据适合进行综合对比评价，同时还可以提高模型的预测精度[7]。归一化目前有两种方法，一种是把数据修正为(0, 1)之间的小数，另一种是把有量纲表达式修正为无量纲表达式。本文将采用(0, 1)标准化方法，其计算公式如下：

$$x_{\text{normalization}} = \frac{\mathrm{x}-\min}{\max-\min} \tag{3}$$

式中：x 为变量的值，min 为该类变量中的最小值，max 为该类变量中的最大值，$x_{\text{normalization}}$ 表示归一化的值。

2.2.2. 多元回归预测模型构建

研究一个因变量与两个或者两个以上的自变量的回归称为多元回归，多元回归主要在于反映一种现象或事物的数量，能够随着多种现象或者事物的数量的变动而发生相应变动的现象[8]。本文中由于分子描述符变量与 pIC50 值的线性或非线性关系未知，所以分别建立四个多元回归模型，四个模型为两个多元线性回归模型以及两个多元非线性回归模型。两个线性回归模型分别为原始数据线性回归以及稳健回归(使用加权最小二乘)；两个非线性回归模型分别为最高项为二次多元非线性回归以及最高项为三次多元非线性回归。

多元回归预测模型程序最终选用 Matlab 软件进行编写，本文将数据按照分子描述符随机打乱，选取前 80%作为训练集，后 20%作为测试集。将训练集数据带入运行程序可得四个多元回归模型，其中两个线性回归模型中二十个分子描述符与 pIC50 值的回归关系式如下所示，

$$\begin{aligned} y_1 = {} & 0.50349+0.22444x_1+0.24878x_2+0.45882x_3-0.9154x_4+0.822x_5-0.42634x_6 \\ & +0.17855x_7+1.0655x_8-0.75221x_9-0.56858x_{10}-0.85332x_{11}-15.367x_{12}+21.981x_{13} \\ & -0.31854x_{14}+1.4427x_{15}-8.6921x_{16}+0.065924x_{17}-0.3471x_{18}+1.0771x_{19}+0.23666x_{20} \end{aligned} \tag{4}$$

$$\begin{aligned} y_2 = {} & 0.39612+0.31757x_1+0.52051x_2+0.44571x_3-0.96281x_4+0.74686x_5-0.24961x_6 \\ & +0.15347x_7+1.9803x_8-0.4776x_9-0.55135x_{10}-1.0785x_{11}-13.311x_{12}+19.851x_{13} \\ & -0.30645x_{14}+1.019x_{15}-10.722x_{16}+0.06831x_{17}-0.28744x_{18}+5.0337x_{19}+0.17824x_{20} \end{aligned} \tag{5}$$

本文得出的四个回归模型如图 4 所示。

2.3. 回归模型验证

将测试集导入回归预测模型中，将预测得到的结果与其真实值进行对比，结果如图 5 所示。

可见多元回归预测值与其对应的真实值相近，说明回归模型具有较好回归结果，可较为真实地反应 pIC50 预测值。

图 6 计算各样本的预测残差百分比，可见原始数据线性回归以及稳健回归模型预测 pIC50 值残差百分比最大为 0.5%左右，最高项为二次的非线性回归模型预测 pIC50 值残差百分比最大为 0.23%左右，其中三次非线性回归模型预测结果最为准确，样本中预测的误差百分比均控制在 0.2%以内。

通过计算可以得出原始数据线性回归预测模型决定系数为 0.524，稳健回归预测模型决定系数为 0.562，最高项为二次的非线性回归预测模型决定系数为 0.722，最高项为三次的非线性回归预测模型决定系数为 0.958。在多元回归预测模型中，最高项为三次的多元非线性回归预测模型最为准确。

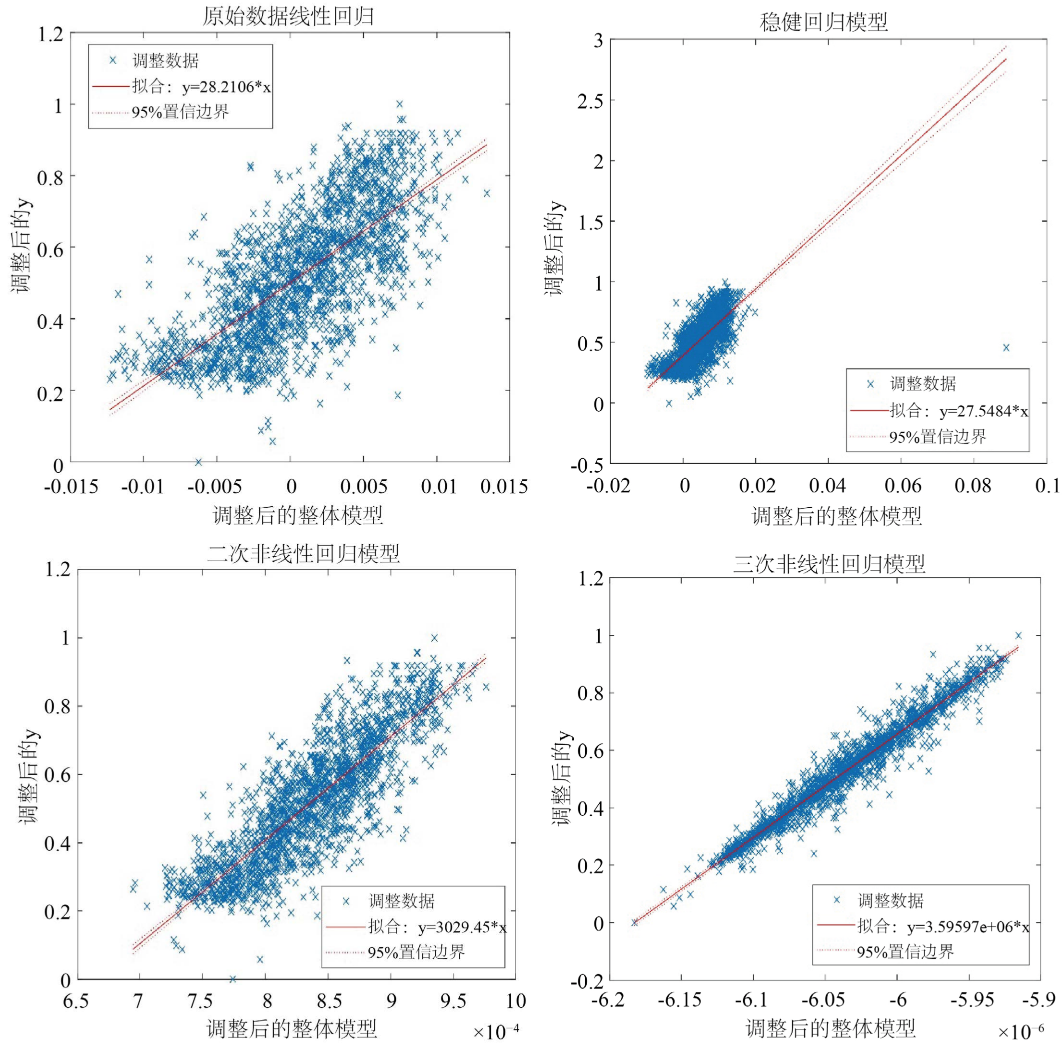

Figure 4. Regression model diagram
图 4. 回归模型图

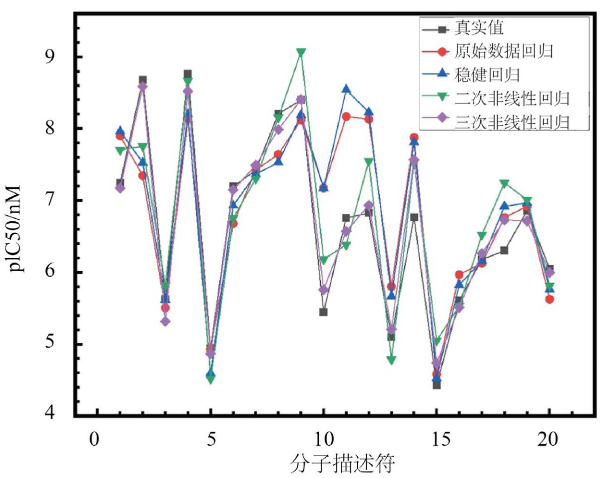

Figure 5. The prediction results of the four models of pIC50 are compared with the true values
图 5. pIC50 四个模型预测结果与真实值对比

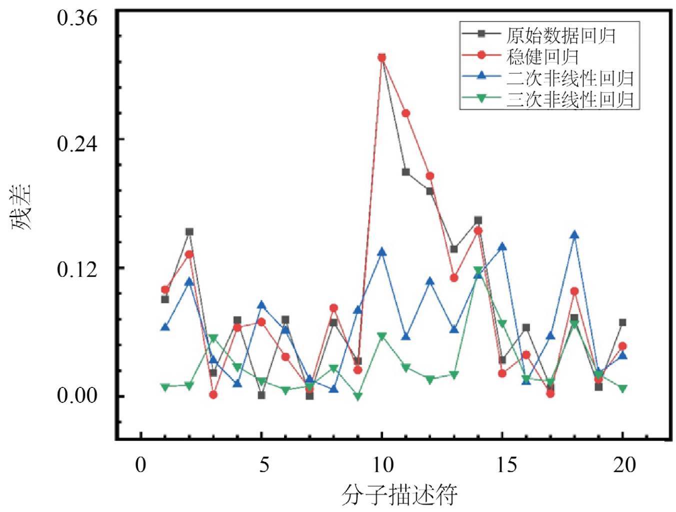

Figure 6. The four models of pIC50 predict the percentage of residual of the result
图 6. pIC50 四个模型预测结果残差百分比

3. 分类预测模型

本文使用 729 个分子描述符变量数据训练出来的模型，构建五种(Caco-2、CYP3A4、hERG、HOB、MN)分类预测模型，以实现对 50 种化合物的 ADMET 性质的预测。由于可以采用 0、1 两种形式来描述一种化合物的优劣两种属性，所以可简化为一个建立 0/1 分类预测模型的问题。难点在于涉及的变量及变量数据较多，不仅对有 729 个自变量，并且每组自变量有对应的 1974 个数据。筛选数据训练模型使其能够对应变量进行准确的 0/1 预测，针对这一难点，普通的求极值方法往往会陷入得到的结果是局部最优解的困境。为解决该问题本文采用广义回归神经网络(GRNN)和随机森林算法，利用数据训练模型，对比准确度选择最优模型对化合物进行相应的预测。

3.1. 广义回归神经网络(GRNN)与随机森林分类预测模型

广义回归神经网络(GRNN)是基于非线性回归理论的前馈式神经网络模型，作为径向基函数(RBF)网络的一种，是通过概率密度函数进行预测，其神经网络结构与 BP 神经网络结构相似，分为输入层、模式层、求和层和输出层[9]。

随机森林由多个决策树构成，决策树彼此之间互不关联。在进行分类预测的时候，每一棵决策树都要参与分类和判断，每一棵决策树都会产生独属于自己的分类结果，随机森林的最终结果是由决策树中分类最多的结果决定的[10] [11]。

3.2. 分类预测模型模型验证

本文将数据按照分子描述符随机打乱，选取前 80%作为训练集，后 20%作为测试集。利用训练集对模型调试，对测试集样本进行求解，选择最优模型对化合物 ADMET 性质进行相应的预测。

将预测模型分别带入测试集进行模型验证，预测得到的结果与其真实值进行对比计算正确率，首先测试 Caco-2 分类预测模型，结果如图 7，图 8 所示。

GRNN 分类预测模型预测正确率达到 94.59%；随机森林预测模型预测正确率为 81.89%。下面按照同样的方法对 CYP3A4、hERG、HOB、MN 分类预测模型的预测结果进行正确率计算，计算结果如表 2 所示。

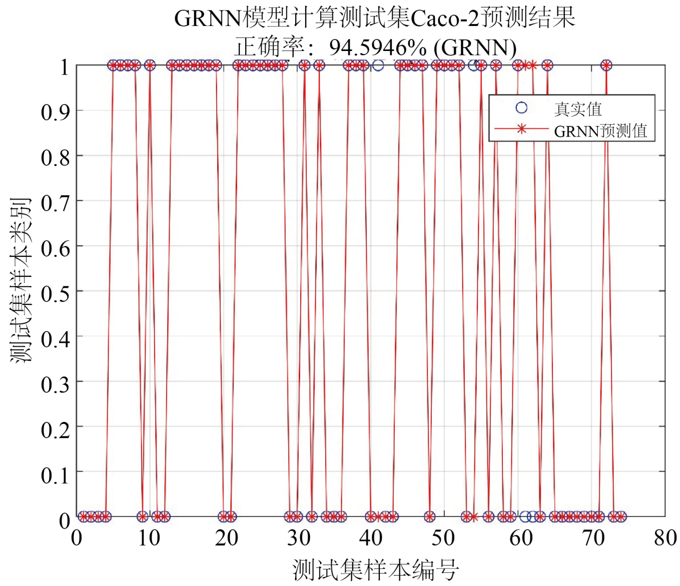

Figure 7. The GRNN model calculates the accuracy of the prediction results of the test set Caco-2
图 7. GRNN 模型计算测试集 Caco-2 预测结果正确率

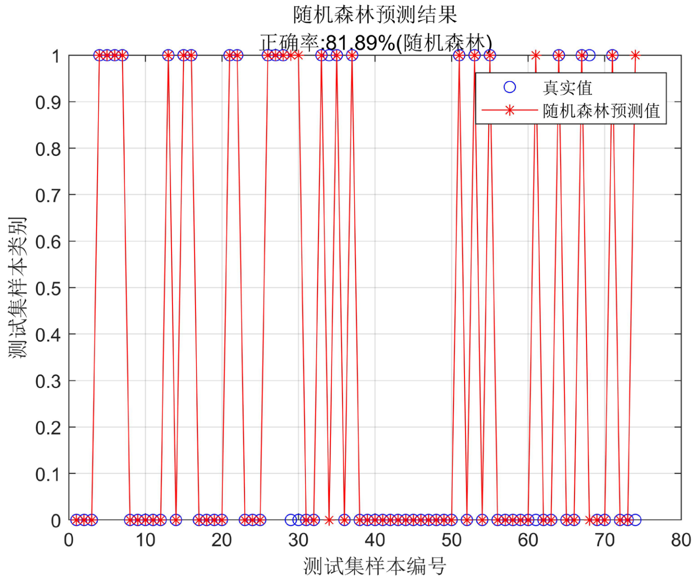

Figure 8. The random forest model calculates the accuracy of the prediction results of the test set Caco-2
图 8. 随机森林模型计算测试集 Caco-2 预测结果正确率

Table 2. Comparison of the accuracy of prediction results
表 2. 预测结果正确率的对比

分类预测模型 / ADMET 性质	GRNN	随机森林
Caco-2	94.59%	81.89%
CYP3A4	93.24%	82.43%

Continued

hERG	95.94%	90.54%
HOB	93.24%	79.72%
MN	97.30%	87.84%

可见通过 GRNN 预测模型预测对(Caco-2、CYP3A4、hERG、HOB、MN)的预测都能达到 93%以上，随机森林预测正确率相对较低，可见在实际研究过程中利用 GRNN 预测模型能够有效预测所给 50 种化合物的 ADMET 性质。

4. 结论

本文充分利用非线性相关分析、多元回归分析、GRNN 神经网络模型、随机森林算法等数据挖掘技术以及机器学习技术，建立了相关化合物定量结构-ERα 活性以及定量结构-ADMET 性质的定量预测模型。得出以下结论：

1) 利用 Spearman 相关系数计算方法得到 729 个分子描述符对生物活性影响程度，并对其进行排序，能得到影响程度排名前二十的变量。

2) 最高项为三次的多元非线性回归预测模型精确度最高，R 方为 0.958，残差百分比都在 0.2%以内，使用此方法建立了相关化合物定量结构-ERα 活性定量预测模型，能提高 IC50 值和对应的 pIC50 值预测的正确率。

3) 利用广义回归神经网络(GRNN)分类预测模型，该模型进行优化后预测正确率均在 93%以上，使用此模型建立了相关化合物定量结构-ADMET 性质的定量预测模型，提高预测的准确性。

参考文献

[1] Feng, R.M., Zong, Y.N., Cao, S.M. and Xu, R.H. (2019) Current Cancer Situation in China: Good or Bad News from the 2018 Global Cancer Statistics? *Cancer Commun* (*Lond*), **39**, 22.

[2] Wilkinson, K.D. (2004) Ubiquitin: A Nobel Protein. *Cell*, **119**, 741-745.

[3] Deroo, B.J. and Korach, K.S. (2006) Estrogen Receptors and Human Disease. *Journal of Clinical Investigation*, **116**, 561-570.

[4] 吴俊. 新型氧桥双环庚烯类选择性雌激素受体调节剂功能与作用机制研究[D]: [博士学位论文]. 武汉: 武汉大学, 2019. https://doi.org/10.27379/d.cnki.gwhdu.2019.000491

[5] 黄楚怡. 紫铆因通过调控雌激素受体 α 降解抑制乳腺癌细胞生长[D]: [硕士学位论文]. 广州: 广州医科大学, 2020. https://doi.org/10.27043/d.cnki.ggzyc.2020.000259

[6] 徐龙博, 王伟, 张滔, 杨莉, 汪少勇, 李煜东. 基于神经网络平均影响值的超短期风电功率预测[J]. 电力系统自动化, 2017, 41(21): 40-45.

[7] 曹丽君, 吴湘华. k 均值聚类算法归一化处理前后效果研究比较[J]. 电子制作, 2014(16): 50-51. https://doi.org/10.16589/j.cnki.cn11-3571/tn.2014.16.104

[8] 孙众, 宋洁, 吴敏华, 骆力明. 教学干预: 提升混合课程质量的关键因素[J]. 中国电化教育, 2017(4): 90-96.

[9] 简书强. 萤火虫群优化算法与 GRNN 神经网络并行集成学习研究[D]: [硕士学位论文]. 合肥: 合肥工业大学, 2020. https://doi.org/10.27101/d.cnki.ghfgu.2020.000563

[10] 方匡南, 吴见彬, 朱建平, 谢邦昌. 随机森林方法研究综述[J]. 统计与信息论坛, 2011, 26(3): 32-38.

[11] 黄海新, 吴迪, 文峰. 决策森林研究综述[J]. 电子技术应用, 2016, 42(12): 5-9. https://doi.org/10.16157/j.issn.0258-7998.2016.12.001