Modeling and Simulation 建模与仿真, 2022, 11(4), 1119-1130
Published Online July 2022 in Hans. http://www.hanspub.org/journal/mos
https://doi.org/10.12677/mos.2022.114102

Hans汉斯

基于BP神经网络的抗乳腺癌药物的选择与寻优

程　攀[1]，胡　琪[2]

[1]上海理工大学机械工程学院，上海
[2]上海理工大学光电信息与计算机工程学院，上海

收稿日期：2022年5月15日；录用日期：2022年7月11日；发布日期：2022年7月18日

摘　要

研究表明，雌激素受体α亚型(Estrogen receptors alpha, ERα)在乳腺发育过程中至关重要，ERα被认为是治疗乳腺癌的重要靶标。本文基于1974个与ERα的生物活性有关的化合物，采用机器学习，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。随后选用基于样本相关系数的检验对729个分子描述符对生物活性影响的重要性进行排序，最终得到了前20个最具影响的分子描述符。采用具有非线性映射能力的BP神经网络来建立生物活性预测模型。同时从线性模型与非线性模型两个角度出发来构建模型，计算得出两种模型寻找的20个主要分子描述符及获得的生物活性和ADMET性质。

关键词

相关性分析，线性模型，非线性模型，BP神经网络

Selection and Optimization of Anti-Breast Cancer Drugs Based on BP Neural Network

Pan Cheng[1], Qi Hu[2]

[1]School of Mechanical Engineering, University of Shanghai for Science and Technology, Shanghai
[2]School of Opto-Electronic Information and Computer Engineering, University of Shanghai for Science and Technology, Shanghai

Received: May 15th, 2022; accepted: Jul. 11th, 2022; published: Jul. 18th, 2022

Abstract

Studies have shown that estrogen receptors alpha (ERα) is crucial in mammary gland development, and ERα is considered an important target for breast cancer treatment. Based on 1974 compounds related to the biological activity of ERα, this paper uses machine learning to build a quantitative prediction model of compound biological activity and a classification prediction

文章引用: 程攀, 胡琪. 基于 BP 神经网络的抗乳腺癌药物的选择与寻优[J]. 建模与仿真, 2022, 11(4): 1119-1130.
DOI: 10.12677/mos.2022.114102

model of ADMET properties. Then, the test based on the sample correlation coefficient was used to rank the importance of the influence of 729 molecular descriptors on biological activity, and finally the top 20 most influential molecular descriptors were obtained. A BP neural network with nonlinear mapping ability was used to establish a biological activity prediction model, and a multiple linear regression model and a gradient boosting regression tree model were established for comparison and verification. At the same time, the model was constructed from the perspective of linear model and nonlinear model, and the 20 main molecular descriptors sought by the two models and the obtained biological activities and ADMET properties were calculated.

Keywords

Correlation Analysis, Linear Model, Nonlinear Model, BP Neural Network

1. 引言

乳腺癌是发生于乳腺上皮或导管上皮的恶性肿瘤，病因尚不完全清楚，其在全球女性癌症中的发病率为24.2%，位居首位，其中52.9%发生在发展中国家。研究表明，雌激素受体 α 亚型[1] (Estrogen receptor alpha, ERα)在乳腺发育过程中至关重要。现阶段，抗激素治疗方法经常被用来帮助ERa表达的患者，它可以调节雌激素的受体活性从而达到控制体内雌激素水平的目的。ERα 被认为是治疗乳腺癌的重要靶标，能够拮抗ERα活性的化合物可能是治疗乳腺癌的候选药物。

在药物研发中，通常采用建立化合物活性预测模型的方法来筛选潜在活性化合物。现收集一系列作用于ERa靶标的化合物及其生物活性数据，然后以一系列分子结构描述符作为自变量，化合物的生物活性值作为因变量，构建化合物的定量结构－活性关系(Quantitative Structure-Activity Relationship, QSAR)模型，然后使用该模型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。

一个化合物想要成为候选药物，需具备的性质合称为ADMET (Absorption 吸收、Distribution 分布、Metabolism 代谢、Excretion 排泄、Toxicity 毒性)性质。为方便建模，本文仅考虑化合物的5种ADMET性质，分别是：1) 小肠上皮细胞渗透性(Caco-2)；2) 细胞色素P450酶(Cytochrome P450, CYP) 3A4亚型(CYP3A4)；3) 化合物心脏安全性评价(human Ether-a-go-go Related Gene, hERG)；4) 人体口服生物利用度(Human Oral Bioavailability, HOB)；5) 微核试验(Micronucleus, MN)。

本文基于2021年华为杯中国研究生数学建模竞赛D题提供的数据集，数据集内含1974个化合物数据、1974个化合物的729个分子描述符信息、1974个化合物的5种ADMET性质，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，从而为同时优化ERα拮抗剂的生物活性和ADMET性质提供预测服务。

2. ERα 生物活性数据的选择

要求变量对生物活性影响的重要性对729个分子描述符进行变量选择，可以认为是特征选择问题。因此本文选用基于样本相关系数的检验对729个分子描述符对生物活性影响的重要性进行打分，为防止分子描述符之间存在耦合关系，使用皮尔森相关性检验对按照重要度排序的分子描述符进行线性相关检验，最终选取前20个最具影响的分子描述符，其流程图如图1所示：

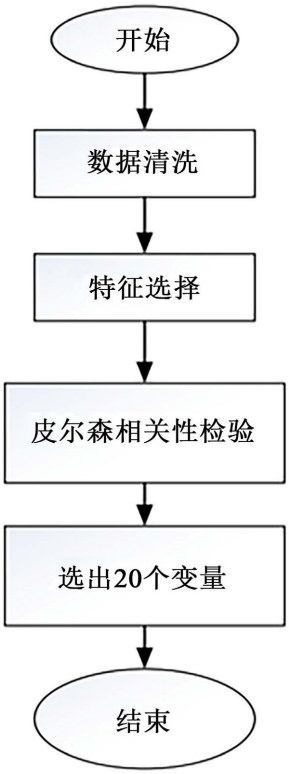

Figure 1. Operation flow chart
图 1. 操作流程图

2.1. 皮尔森相关性分析

皮尔森相关系数(皮尔森 Correlation Coefficient)常用来衡量两个数据集合是否在一条线上面，它用来衡量定距变量间的线性关系。本文假设 729 个分子描述符之间相互独立且对生物活性影响为线性，故皮尔森相关系数适用。

皮尔森相关系数可以定义为：

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

式中： d_i 表示两个变量之间的差值， n 表示样本大小。

皮尔森相关系数衡量的是线性相关关系，若 $|\rho| = 0$，说明两者无线性相关关系； $|\rho|$ 越大，相关性越强， $|\rho|$ 越接近于 0，相关度越弱。统计学中对相关性强弱有着如下约定，如表 1 所示：

Table 1. Relevance scale
表 1. 相关程度度量表

相关性	相关系数
极强相关	0.8~1.0
强相关	0.6~0.8
中相关	0.4~0.6
弱相关	0.2~0.4
极弱或无相关	0~0.2

将皮尔森相关系数的阈值设为 0.6，再基于前面筛选下的变量，计算两两变量间的皮尔森相关系数，然后判断两者的相关系数是否大于所设阈值。若大于，则说明两变量相关性高，对两变量之间贡献值排名靠后的那一变量采取删除操作。若小于，则说明两变量间并不强相关，则将两变量都暂时予以保留。之后重复上述操作，直至遍历结束。完成后，根据分子描述符含义解释，对强耦合变量进行手动剔除，最后保留对生物活性最具影响的前 20 个分子描述符见表 2，以及它们之间的皮尔森相关系数热力图见图 2：

Table 2. Final selection of molecular descriptors
表 2. 最终选择的分子描述符

排名	名称	排名	名称
1	MDEC-23	11	C2SP2
2	MLogP	12	hmin
3	LipoaffinityIndex	13	AMR
4	maxsOH	14	SwHBa
5	minsOH	15	maxsssN
6	nC	16	MDEC-22
7	nT6Ring	17	SP-5
8	n6Ring	18	SaaCH
9	minsssN	19	CrippenLogP
10	BCUTp-1h	20	maxHsOH

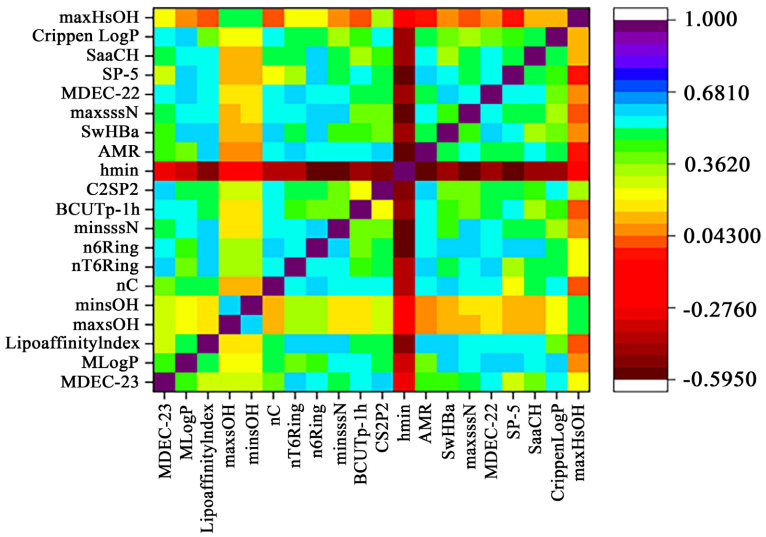

Figure 2. Correlation of selected molecular descriptors relative to biological activity
图 2. 选择的分子描述符相对生物活性的相关性

2.2. 基于遗传算法优化的 BP 神经网络预测模型

2.2.1. 线性预测与非线性预测模型

现结合问题 1 选择不超过 20 个分子描述符构建的化合物对抑制 ERα 生物活性的预测模型，50 个化合物进行 IC50 值和对应的 pIC50 值预测。

一个化合物想要成为候选药物，一般需要具备良好的生物活性，常用做法是针对与疾病相关的某个靶标(此处为 ERα)，收集一系列作用于该靶标的化合物及其生物活性数据，然后以一系列分子结构描述符作为自变量，化合物的生物活性值作为因变量，构建化合物的定量结构－活性关系(Quantitative Structure-Activity Relationship, QSAR)模型，然后使用该模型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。而由于分子描述符众多，且涉及到的方面较复杂，因此考虑同时采用线性和非线性的预测模型来建立以分子描述符构建的化合物对生物活性的预测模型，根据最终表现来评估模型，其思路如图 3 所示：

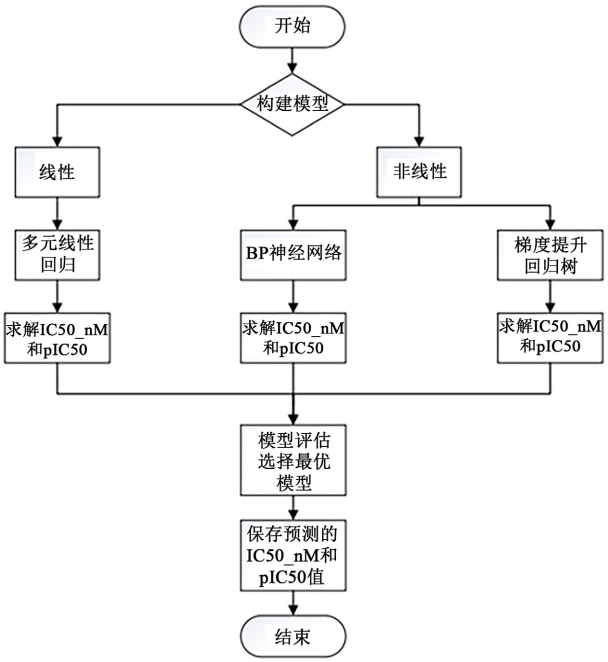

Figure 3. Experimental flow chart
图 3. 实验流程图

2.2.2. BP 神经网络模型

BP 神经网络属于前向神经网络，强调网络采用误差反向传播的学习算法。其包括一个输入层、若干隐含层和一个输出层组成。其核心思想是通过样本训练集，不断修正神经网络的权值和阈值，逐步逼近期望输出值。其输入层与输出层关系如图 4：

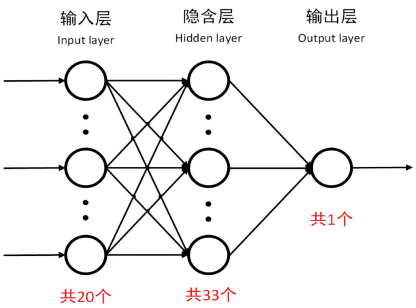

Figure 4. Neural network I/O layer diagram
图 4. 神经网络输入输出层关系图

根据本题数据，建立的 BP 神经网络模型初始输入为 20 个变量，输出层为预测值，含有 5 个隐藏层，神经元个数分别为 50，100，50，100，50，其中输入层值为 X，即 20 个变量。

目前主要有三种常用的激励函数：Sigmoid、Thah 和 ReLU 激励函数。ReLU 使得 SGD 的收敛速度比 Sigmoid 和 Thah 快很多，使过程计算量减少，此外还解决了梯度消失问题。出于此种考虑，本文最后采用 ReLU (Rectified Linear Uni)激励函数作为本神经网络的激励函数。

选用 Python 编写程序进行神经网络训练。设置最大迭代次数为 1000 次，期望误差与学习速率均设置为软件自适应寻优，该网络训练完成后，只需将各项主要变量值输入网络即可得到预测数据。

2.2.3. BP 神经网络模型求解

建立 BP 神经网络模型后，经过求解得出 50 个化合物进行 IC50 和对应的 IC50 值。本文将数据分为两个部分，75%用于训练，25%用于模型测试，测试结果及训练 LOSS 如图 5 所示，拟合结果如图 6 所示：

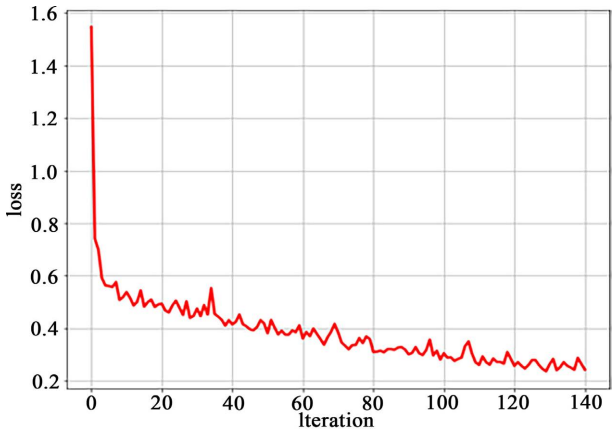

Figure 5. Neural network LOSS diagram
图 5. 神经网络 LOSS 图

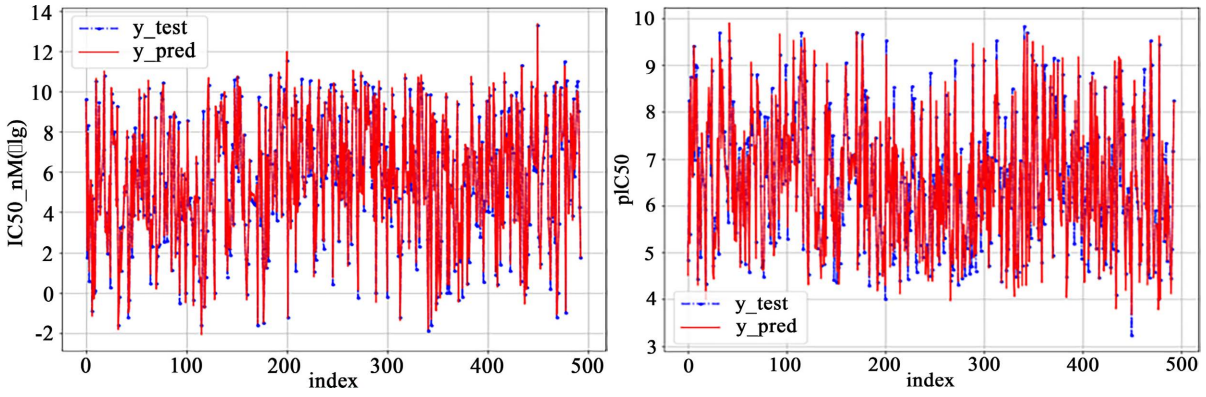

Figure 6. Curves of true and fitted values (left IC50_nM (take l g) right pIC50)
图 6. 真实值和拟合值曲线(左 IC50_nM (取 l g)右 pIC50)

通常用神经网络结果的均方误差(MSE)和真实值 R 来判断神经网络模型的预测精度。均方误差(MSE)是预测值和真实值之差的平方和的平均值，MSE 值越小，该模型的预测精度就越高。真实值 R 反映了预测模型的拟合程度，R 越接近 1 效果越好。分析图 6 可得误差分析结果可得 BP 神经网络的均方差等于 0.03212，其判定系数 R 为 0.9605，是想要的预测结果。

3. 基于支持向量机(SVM)的 ADMET 性质分类预测模型

现需要根据 729 个分子描述符，针对提供的 1974 个化合物的 ADMET 数据，分别构建化合物 Caco-2、CYP3A4、hERG、HOB、MN 的分类预测模型，50 个化合物进行相应的预测。基于此，本文建立 K 近邻分类、决策树分类、随机森林分类三种分类预测模型进行分类预测，并在运行中进行评分，从而保证结果的准确性。

K 近邻[2] (K-Nearest Neighbour, KNN)是一种基本分类方法，通过测量不同特征值之间的距离进行分类。K 近邻的思路是：如果一个样本在特征控件中的 k 个最相似(即特征空间中最邻近)的样本中的大多数属于某一个类别，则该样本也属于这个类别，其中 k 通常是不大于 20 的整数。

决策树[3] (Decision Tree)是一种基本的分类与回归方法，当决策树用于分类时称为分类树，用于回归时称为回归树。分类树是一种描述对实例进行分类的树形结构。在使用分类树进行分类时，从根结点开始，对实例的某一特征进行测试，根据测试结果，将实例分配到其子结点。这时，每一个子结点对应着该特征的一个取值。如此递归地对实例进行测试并分配，直至达到叶结点。最后将实例分到叶结点的类中。随着划分过程不断进行，决策树的分支结点所包含的样本尽可能属于同一类别，即结点的“纯度”(purity)越来越高。

随机森林[4]就是通过集成学习的思想将多棵树集成的一种算法，它的基本单元是决策树，而它的本质属于机器学习的一大分支——集成学习(Ensemble Learning)方法。每棵决策树都是一个分类器，那么对于一个输入样本，N 棵树会有 N 个分类结果。而随机森林集成了所有的分类投票结果，将投票次数最多的类别指定为最终的输出，这就是一种最简单的 Bagging 思想。

最后采用 k 近邻分类、决策树分类和随机森林分类对 Caco-2、CYP3A4、hERG、HOB、MN 进行的分类结果如表 3 所示。

Table 3. Classification results of three models for Caco-2, CYP3A4, hERG, HOB, MN
表 3. 三种模型对 Caco-2、CYP3A4、hERG、HOB、MN 分类结果

	Caco-2			CYP3A4			hERG			HOB			MN		
	①	②	③	①	②	③	①	②	③	①	②	③	①	②	③
1	0	0	0	1	1	1	1	1	1	0	0	0	1	1	1
2	0	0	0	1	1	1	0	1	1	0	0	0	1	1	1
3	0	0	0	1	1	1	1	1	1	0	0	0	1	1	1
4	0	0	0	1	1	1	0	1	1	0	0	0	1	1	1
5	0	0	0	1	1	1	1	0	1	0	0	0	1	1	1
6	0	0	0	1	1	1	1	1	1	0	0	0	1	1	1
7	0	0	0	1	1	1	1	1	1	0	0	0	1	0	1
8	0	0	0	1	1	1	1	1	1	0	0	0	1	1	1
9	0	0	0	1	1	1	1	0	1	0	0	0	1	1	1
10	0	0	0	1	1	1	0	1	1	0	0	0	1	0	1
11	0	0	0	1	1	1	1	0	1	0	0	0	1	0	1
12	0	0	0	1	1	1	1	1	1	0	0	0	1	0	1
13	0	0	0	1	1	1	1	1	1	0	0	0	1	0	1
14	0	0	0	1	1	1	1	1	1	0	0	0	1	0	1
15	0	0	0	1	1	1	1	1	1	0	0	0	1	1	1

Continued

16	0	0	0	1	1	1	1	1	1	0	0	0	1	1	1
17	0	0	0	1	1	1	1	1	1	0	0	0	1	1	1
18	0	0	0	1	1	1	1	1	1	0	0	0	1	1	1
19	0	1	0	1	1	1	0	0	1	1	1	0	1	1	1
20	0	0	0	1	0	0	1	0	0	0	0	0	1	1	1
21	0	0	0	1	1	1	1	1	1	0	0	0	1	1	1
22	0	0	0	1	1	1	0	0	1	0	0	0	1	1	1
23	1	1	1	1	1	0	0	1	0	1	0	1	0	0	0
24	1	1	1	1	1	0	0	1	0	0	0	1	0	0	0
25	1	1	1	0	0	0	1	1	1	0	0	1	1	1	0
26	1	1	1	0	0	0	1	1	1	0	0	1	1	1	0
27	0	1	0	1	1	1	1	1	1	0	0	0	0	0	0
28	0	0	0	1	1	1	1	1	1	0	0	0	1	1	1
29	0	0	0	1	1	1	1	0	1	0	0	0	1	1	1
30	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1
31	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1
32	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1
33	1	0	1	1	1	1	1	1	1	0	0	1	1	1	1
34	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1
35	0	1	0	1	1	1	1	1	1	0	0	0	1	1	1
36	0	0	0	1	1	1	0	1	0	0	1	0	1	1	0
37	0	0	0	1	1	1	0	1	0	0	1	0	1	1	1
38	0	0	0	1	1	1	1	1	1	0	0	0	0	0	0
39	0	0	0	1	1	1	0	1	1	0	1	0	1	1	1
40	0	0	0	1	1	1	0	1	1	0	1	0	1	1	1
41	0	0	0	1	1	1	0	1	1	0	1	0	1	1	1
42	0	0	0	1	1	1	0	1	1	0	1	0	1	1	1
43	0	0	0	1	1	1	0	1	1	0	1	0	1	1	1
44	0	0	0	1	1	1	0	1	1	0	1	0	1	1	1
45	0	0	0	1	1	1	0	1	1	0	1	0	1	1	1
46	1	0	1	1	1	1	1	1	1	0	1	0	1	0	1
47	0	0	0	1	1	1	1	1	1	0	1	0	1	0	0
48	0	0	0	1	1	1	1	1	1	0	1	0	1	0	1
49	0	0	0	1	1	1	1	1	1	0	1	0	1	1	1
50	0	1	0	1	1	1	1	1	1	0	0	0	0	0	0

注：① 表示 k 近邻分类；② 表示决策树分类；③ 表示随机森林分类。

结果如表 3 所示，三种模型的结果非常接近，其中不同结果的按照模型的正确率作为主要条件来综合评估。

4. 模型建立及寻优求解

4.1. 优化多目标问题

现优化目标为，需要找出一些分子描述符在某定值或者取值范围内，既能够使化合物对抑制 ERα 具有更好的生物活性，同时具有更好的 ADMET 性质(给定的五个 ADMET 性质中，至少三个性质较好)。本题属于回归问题，可以理解为既要保证好的生物活性，同时至少要具有三个好的 ADMET 性质，即遴选出合格的分子描述符，在 Caco-2、CYP3A4、hERG、HOB、MN 五个指标至少有三个为 1 的前提下，生物活性尽可能好，由于生物活性主要看 pIC50 值，pIC50 值越大说明生物活性越高，因此把这个值看做寻优目标，其求解过程如图 7：

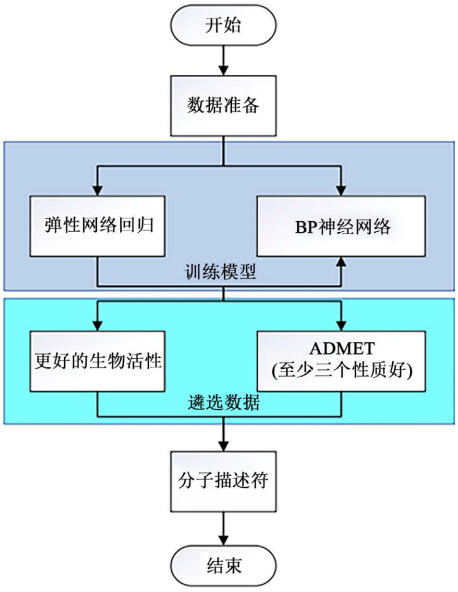

Figure 7. Optimization flow chart
图 7. 寻优流程图

现遴选出 20 个相对目标较为重要的变量，同时同线性模型与非线性模型两个角度出发来构建模型，选择弹性网络回归(ElasticNet)与 BP 神经网络来对比解题。

4.2. 模型建立

考虑到建模目标发生改变，普通线性回归模型可能因为没有加入惩罚项问题而产生较差拟合效果，而弹性网络回归则解决了这一问题，原因在于线性回归损失函数基于最小二乘法的：

$$\min_{\theta}\frac{1}{2m}\sum_{i=1}^{m}\left(\theta^{\mathrm{T}}x^{(i)}-y^{(i)}\right)^{2}$$

在无法权衡到底是 L1 还是 L2 正则化对参数更新更为有利的时候，便需要在 ElasticNet 中加入 L1 和 L2 正则化项：

$$\min_{\theta}\frac{1}{2m}\left[\sum_{i=1}^{m}\left(\theta^{\mathrm{T}}x^{(i)}-y^{(i)}\right)^{2}+\lambda_{1}\sum_{j=1}^{n}\|\theta\|+\lambda_{2}\sum_{j=1}^{n}\|\theta\|_{2}^{2}\right]$$

考虑到数据是具有非线性的，于是仍使用表现良好的 BP 神经网络同时求解该题。

4.3. 模型求解与评价

本文利用 Python 科学计算库自带的网格搜索(GridSearch CV)函数来求解 ElasticNet 回归模型，它存在的意义就是自动调参，在小数据集下，只要把参数输进去，就能很快给出最优化的结果和参数。它其实是一种贪心算法：拿当前对模型影响最大的参数调优，直到最优化；再拿下一个影响最大的参数调优，如此下去，直到所有的参数调整完毕。这个方法的缺点就是可能会调到局部最优而不是全局最优，但是省时间省力，GridSearch CV 用于系统地遍历多种参数组合，通过交叉验证确定最佳效果参数。在此基础上，最终通过计算，发现模型在学习率为 0.01 的情况下效果最好。

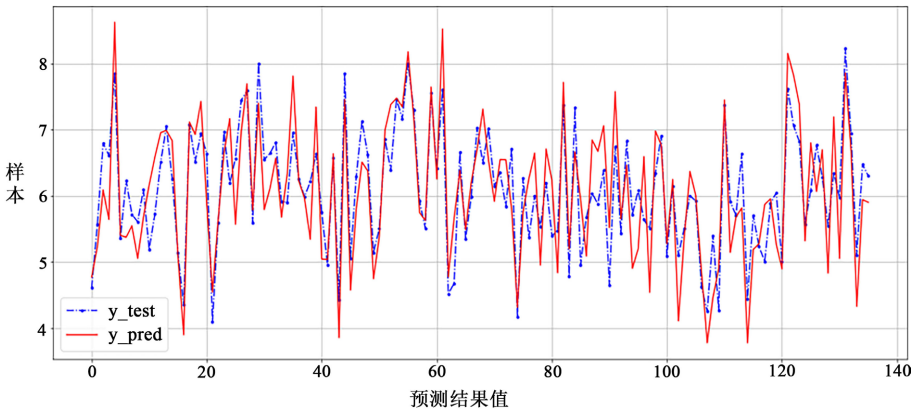

Figure 8. ElasticNet model fit vs true
图 8. ElasticNet 模型拟合值与真实值

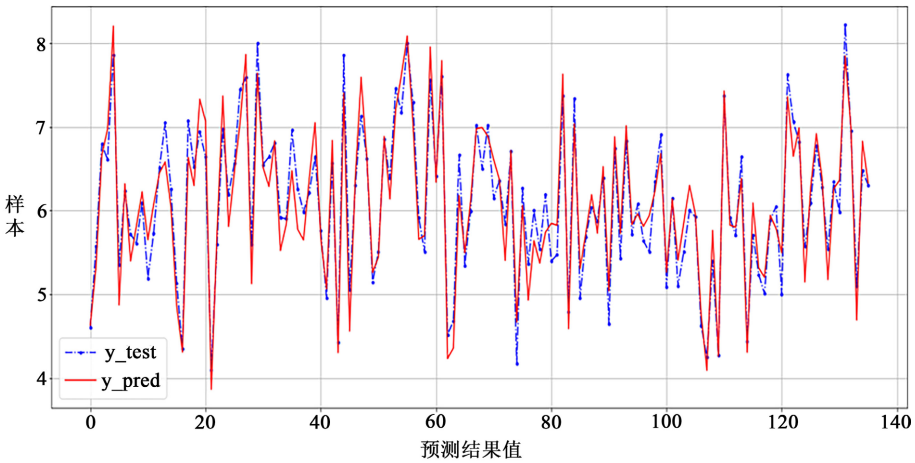

Figure 9. Fitted value and true value of BP neural network model
图 9. BP 神经网络模型拟合值与真实值

图 8 和图 9 分别为 ElasticNet 模型和 BP 神经网络模型的拟合值与真实值，蓝色点划线为从样本变量中预留的 25%的数据，红色为模型的拟合值，从图中可以看出两种模型均取得了较好的成绩，通过表 4 对比两种模型的均方误差、平均绝对误差以及判定系数 $r2$ 值，发现两种模型的均方误差均较小，符合要求，ElasticNet 回归模型平均绝对误差远大于 BP 神经网络模型，且 BP 神经网络判定系数高于 ElasticNet 回归模型，说明 BP 神经网络模型拟合度更好。因此，相比于 ElasticNet 回归模型 BP 神经网络模型表现更好。

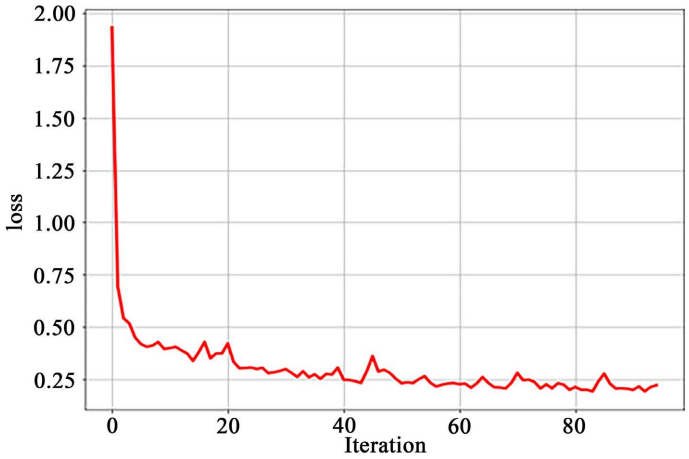

Figure 10. BP neural network model loss function
图 10. BP 神经网络模型损失函数

Table 4. Two model evaluation forms
表 4. 两种模型评估表

预测模型	平均绝对误差(MAE)	均方差(MSE)	判定系数(r2)
BP 神经网络	0.05973	0.07930	0.9228
ElasticNet	0.29281	0.12928	0.8292

如表 5 中所示，给出两种模型寻找的 20 个分子描述符在表 6 中的取值获得的生物活性以及 ADMET 性质，并且可以认为当化合物的分子描述符在这两种结果的取值范围内时，不仅能够使化合物对抑制 ERα 具有更好的生物活性，还能同时至少具有三个好的 ADMET 性质，满足要求。

Table 5. Calculation results for both models (select molecular descriptors)
表 5. 两种模型计算结果(选择分子描述符)

	ElasticNet 回归	BP 神经网络
MDEC-23	30.715194	31.615754
MLogP	3.33	3.44
Lipoaffinity Index	5.752637	6.763017
maxsOH	10.208578	9.873607
minsOH	9.528334	9.524219
nC	21	21
nT6Ring	3	3
n6Ring	3	3
minsssN	0	0
BCUTp-1h	12.463557	12.404744
C2SP2	15	15
hmin	0.093031	0.103031

Continued

AMR	106.0279	104.4224
SwHBa	20.999342	25.627035
maxsssN	0	0
MDEC-22	13.44782	15.857284
SP-5	7.26146	7.135826
SaaCH	18.352802	21.06564
CrippenLogP	3.63558	3.92998
maxHsOH	0.566152	0.566152

Table 6. Two model evaluation forms
表 6. 两种模型评估表

	pIC50	Caco-2	CYP3A4	hERG	HOB	MN
BP 神经网络	11.0237	1	1	1	0	1
ElasticNet	12.1762	0	1	1	0	1

5. 总结与分析

1) 本文首先根据样本相关系数的检验对 729 个分子描述符对生物活性影响的重要性进行打分，并使用皮尔森相关性检验对按照重要度排序的分子描述符进行线性相关检验，最终得到了前 20 个最具影响的分子描述符。

2) 筛选出 20 个对生物活性影响最重要的分子描述符，数据最大为 1974 × 20，数据量相对中等，故采用具有非线性映射能力的 BP 神经网络来建立生物活性预测模型针对主要变量与生物活性之间复杂的关系，选用 BP 神经网络机器学习算法，同时数据集与测试集分开，所得到的结果最大相对误差很小，所建立的预测模型精度表现优秀。

3) 第三步选用 K 近邻分类、决策树分类、随机森林分类三种分类预测模型进行分类预测，从而保证结果的准确性。结果显示，三种模型的结果非常接近，其中不同的结果均以模型的正确率作为主要条件来综合评估。

4) 最后同时从线性模型与非线性模型两个角度出发来构建模型，经过对比分析，最终选择弹性网络回归与问题二中表现良好的 BP 神经网络来对比解题。结果显示，两种模型均取得了较好的成绩，通过对比两种模型的均方误差、平均绝对误差以及判定系数值，发现 BP 神经网络模型表现更好，计算得出两种模型寻找的 20 个主要分子描述符及获得的生物活性和 ADMET 性质，并且认为当化合物的分子描述符在这两种结果的取值范围内时，获得较好的分子活性和至少 3 个好的 ADMET 性质。

参考文献

[1] 史海龙, 李军, 郭新荣, 党琳, 张红. 基于雌激素受体从传统中药库中筛选治疗乳腺癌的小分子拮抗剂[J]. 辽宁中医药大学学报, 2017, 19(5): 85-89.

[2] 皮亚宸. K 近邻分类算法的应用研究[J]. 通讯世界, 2019, 26(1): 286-287.

[3] Dantas de Jesus Ferreira, J.A. and Secco, N.R. (2021) Decision Tree Classifiers for Unmanned Aircraft Configuration Selection. *Aircraft Engineering and Aerospace Technology*, **93**, 1122-1132. https://doi.org/10.1108/AEAT-03-2021-0074

[4] 吕红燕, 冯倩. 随机森林算法研究综述[J]. 河北省科学院学报, 2019, 36(3): 37-41.