DOI:10.13878/j.cnki.jnuist.2023.01.006

许美贤[1]　郑琰[1]　李炎举[1]　吴伟豪[1]

基于 PSO-BP 神经网络与 PSO-SVM 的抗乳腺癌药物性质预测

摘要

通过实验筛选研发新药的过程非常缓慢且需耗费大量的人力物力,而利用计算机辅助预测药物的分子性质可极大地节省药物研发时间和成本.因此,为了能够使抗乳腺癌候选药物对抑制 ERα 具有良好的生物活性和 ADMET 性质,针对收集到的 1 974 种化合物,首先利用随机森林分类器筛选出前 20 个对生物活性最具显著影响的分子描述符,并以此和 pIC50 值作为特征数据建立 QSAR 模型.其次,基于 PSO 优化 BP 神经网络对 50 个新化合物的生物活性值进行预测,模型拟合度为 0.833 7,根均方误差为 0.731 5,比优化前的 BP 神经网络预测值更贴合实际.随后为提高药物研发的成功率,依据已有的 ADMET 性质数据利用 PSO 优化 SVM 构建 ADMET 分类预测模型,算法交叉验证 CV 准确率达到 94.076 7%,5 个指标模型的预测准确率均在 79%以上.结果表明,所建立的模型比基准模型的预测性能更好,采用的预测策略是有效的,可为抗乳腺癌药物的研发提供借鉴.

关键词

抗乳腺癌药物;生物活性;ADMET 性质;粒子群优化算法;BP 神经网络;支持向量机

中图分类号 TP183

文献标志码 A

收稿日期 2021-12-06

资助项目 国家自然科学基金(71701099,71501090);江苏省高等学校自然科学研究项目(17KJB580008)

作者简介

许美贤,女,硕士生,主要从事人工智能辅助药物设计、数据挖掘的研究.xumeixain3210@163.com

郑琰(通信作者),女,博士,副教授,主要从事计算生物物理学、人工智能辅助生物分子结构预测的研究.ZhengYan3210@163.com

1 南京林业大学 汽车与交通工程学院,南京,210037

0 引言

美国癌症中心 2018 年的癌症数据报告显示,乳腺癌是目前全球女性最高发的恶性肿瘤,它严重威胁着女性的身心健康[1].乳腺癌已经成为一个世界性的医疗保健问题,治疗方案既要有选择性也要考虑有效性的概率.为解决这个问题,药用化学领域对大量的候选药物进行了研究分析.通过对雌激素受体 α 亚型(ERα)基因缺失小鼠的实验结果表明,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.

抗乳腺癌候选药物从研发到投入使用需要拥有良好的生物活性,同时其药代动力学性质和安全性也要符合相关政策法规的要求.而如果仅仅采用实验的方式去评估化合物的生物活性、药代动力学性质和安全性,需要花费的时间和成本将不可计量,其中药代动力学性质和安全性合称为 ADMET(Absorption(吸收)、Distribution(分布)、Metabolism(代谢)、Excretion(排泄)、Toxicity(毒性))性质.而且在试验动物身上获取的数据与临床数据并不完全重合,因此不能满足现代药物研究的需求[2].为了节约时间和成本,研究机构通常选择把体外研究技术和计算机运算模型结合起来建立化合物活性预测模型,筛选潜在活性化合物.即通过收集一系列作用于 ERα 的化合物及其生物活性数据,并选取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(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.此外,除了考虑生物活性,药物代谢动力学性质和毒性(ADMET)也是决定药物研发成功与否的重要因素.一个化合物的活性再好,如果其 ADMET 性质不佳,比如很难被人体吸收,或者体内代谢速度太快,或者具有某种毒性,那么其仍然难以成为药物,因而还需要进行 ADMET 性质优化.

而在如今药物数量剧增的情况下,最经济合理的研究方式是利用计算机辅助的人工智能算法对药物生物活性和 ADMET 性质进行预测分析.顾耀文等[3]从多个公共数据库中收集到了大量的药物 ADMET 数据,经过有效的数据清洗后提出利用图神经网络模型来进行药物研发的虚拟筛选,研究结果表明所建模型预测性能较好,可进行

泛化使用.谢良旭等[4]考虑到浅层和深层神经网络的精度和拟合度问题,选择把数个神经网络和堆叠法等结合起来预测药物分子性质,融合模型预测准确性和可靠性较高.秦洁[5]为有效预测药物先导化合物分子生物活性值,深入研究了矩阵补全算法在标记配体特征中的学习,算法比深度学习展现出更强的优势,预测的最优值更贴合实际.贾聪敏[6]采用随机森林、支持向量机、人工神经网络 3 种机器学习算法进行药物靶点定量预测模型的构建,对比分析 3 种算法的预测结果,表明其构建的最优模型能够客观地从分子振动角度筛选出有效的分子描述符.沈杰[7]在经典遗传算法的基础上吸入精英仓库策略建立小分子 ADMET 的 QSAR 预测模型,同时基于信息增益来评估化合物分子结构,验证了所建模型可推广应用至药物代谢、毒性评估等方面.

回顾文献[1-7]可知,利用人工智能方法预测药物的生物活性和 ADMET 性质显然已成为研究的热点.研究表明利用人工智能算法开展对药物生物活性和 ADMET 性质的预测分析可显著地降低研发成本,提高研发成功几率,且更有利于对候选药物在生物体内发挥的作用进行探索,有效避免因药物产生的副作用和毒性导致的人体疾病,可指导临床治疗时的合理用药[8].由此可见,使用计算机辅助的人工智能算法进行理论预测抗乳腺癌候选药物的生物活性和 ADMET 性质是极具现实意义的.

本文从加拿大阿尔伯塔大学的 DrugBank 药物分子数据库中获取 1 974 种化合物对乳腺癌治疗靶标 ERα 的生物活性和 ADMET 性质数据,利用所收集到的信息从化合物分子描述符角度出发建立定量预测模型,基于粒子群优化 BP 神经网络算法来预测新化合物的 IC_{50}和 pIC_{50}值.同时构建分类预测模型,基于粒子群优化支持向量机来预测化合物的 5 种 ADMET 性质,分别是 Caco-2、CYP3A4、hERG、HOB、MN,从而寻找到能满足化合物活性较高且尽可能使得 ADMET 性质较好的化合物分子描述符,以加快抗乳腺癌候选药物的研发进程.

1 数据收集

针对乳腺癌治疗靶标 ERα,从阿尔伯塔大学的 DrugBank 药物分子数据库中获取了 1 974 个化合物对 ERα 的生物活性数据、729 个分子描述符信息数据、5 种 ADMET 性质数据[9].DrugBank 数据库拥有独特的生物信息学和化学信息学资源,它将详细的药物数据和全面的药物目标信息结合起来,以便科学家们研究药物机制和探索新型药物.本文收集到的数据中包含了化合物的 SMILES 结构式、化合物对 ERα 的生物活性值 IC_{50}和 pIC_{50}值、729 个分子描述符信息(自变量)、分子描述符含义解释,以及采用 0-1 二分类法提供相应取值的 Caco-2、CYP3A4、hERG、HOB、MN 等 5 种药代动力学性质和毒性.

2 筛选主要的分子描述符

2.1 数据预处理

针对收集到的 729 个分子描述符信息进行观察,对数据进行处理发现 1 974 个有机化合物中有些描述符全为 0,例如分子描述符 *n*B(硼原子数)全为 0.大量为“0”的数据并不是缺失,而是化合物的分子描述符就是“0”这个数字[10],这对制药研究是有实际意义的,故在数据预处理时不需要把全为 0 的描述符行列剔除.因此可直接利用原有的 1 974 个化合物的 729 个分子描述符数据作为自变量,生物活性值作为因变量构建定量结构-活性关系(QSAR)模型.

在收集到的数据集中,化合物对 ERα 的生物活性值用 IC_{50}表示.IC_{50}为实验测定值,单位是 nmol/L,该值越小代表生物活性越大,对抑制 ERα 活性越有效.参考文献[7-10]及利用分子描述符计算的专用软件 PaDEL-Descriptor 试验可知,pIC_{50}值通常由 IC_{50}转化而得到(即 IC_{50}值的负对数),而 pIC_{50}值通常与生物活性具有正相关性,即 pIC_{50}值越大表明生物活性越高.在实际 QSAR 理论建模中,一般采取 pIC_{50}值来表示生物的活性值.首先需要针对 1 974 个化合物的 729 个分子描述符进行变量选择,根据各变量对生物活性影响的重要性进行排序,得出前 20 个对生物活性最具显著影响的分子描述符(即自变量).由于收集到的分子描述符数据为二维数据,即对应分子的溶解度、表面积等信息,需要筛选出对结果影响最大的几个特征,以此作为建立模型时的特征数据.而常见的求解方法有主成分分析法、LASSO、随机森林等,但是主成分分析法和 LASSO 这类经典算法对 729 个变量指标进行特征提取和降维时会带来模糊性,使得原始变量含义失去了清晰确切性[11].因此选择利用随机森林(RF)算法对特征重要性进行评估,筛选出对活性值影响大的分子描述符.

2.2 基于随机森林筛选分子描述符

随机森林基于 Bagging 算法的集成思想为每棵

决策树生成独立的同分布训练样本集,所有决策树的投票将决定最终的分类结果.基于随机森林模型把收集到的分子描述符数据输入 MATLAB 软件中进行运算,第 i 次和第 j 次程序运行结果分别如图 1 和图 2 所示.

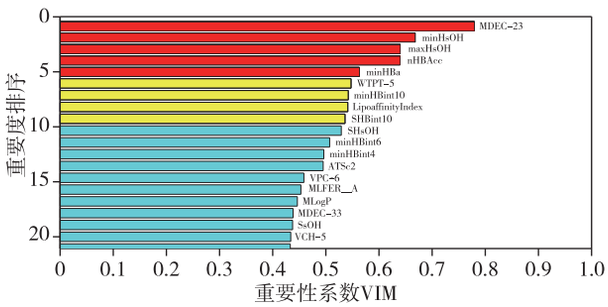

图 1 第 i 次实验分子描述符(变量)的相对重要度

Fig. 1 Relative importance of molecular descriptors (variables) in the i-th experiment

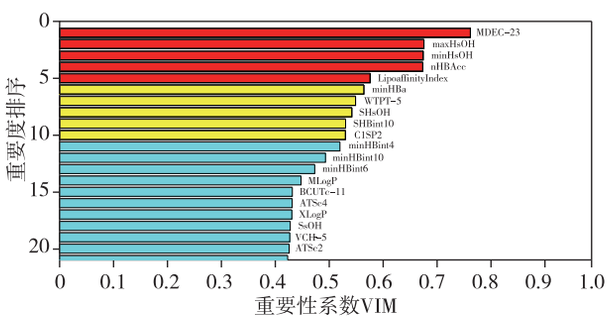

图 2 第 j 次实验分子描述符(变量)的相对重要度

Fig. 2 Relative importance of molecular descriptors (variables) in the j-th experiment

由于每次训练都是随机抽样,程序运行后分子描述符重要度排名结果有所差异,则设计 10 次实验对分子描述符重要性进行统计.设 VIM 为重要度系数,则 VIM_j^i 分别表示第 j 次实验的第 i 名分子描述符的重要度系数.通过统计 10 次实验排名前 20 所出现过的分子描述符,然后计算统计的分子描述符的平均重要性系数,记为 $\overline{VIM}$.最后根据 $\overline{VIM}$ 对所统计的分子描述符进行排序,取平均重要性系数前 20 的为最具显著影响的分子描述符.统计 10 次实验分子变量符中出现的次数如表 1 所示.由表 1 可知 27 个变量出现次数排序,理论出现次数高的其重要性系数也相对较大.通过统计这 27 个变量的平均重要性系数,可得 10 次实验中平均重要性系数排序,如图 3 所示.根据图 3 可得出这 20 个分子描述符来尽可能地描述化合物的生物活性.

3 基于 PSO 优化 BP 神经网络的 QSAR 模型预测分析

在对分子描述符数据进行降维处理后,大大减少了数据量.鉴于 BP 神经网络模型的自适应、泛化及容错能力较强,且可以通过数据逼近任意线性连续的函数,这一特点与分子描述符数据性质对候选药物影响方式的特点相吻合.因此可以选择 BP 神经网络进行训练学习,并对 50 个化合物进行 IC_{50} 值和对应的 pIC_{50} 值预测.本节分析基于 BP 神经网络的生物活性值预测方法,并通过引入具备运行速度较快、全局寻优能力较好的粒子群算法(PSO)来避免传统 BP 神经网络易陷入局部最优解的问题.

3.1 BP 神经网络生物活性值预测模型

采用包含着输入层、隐含层和输出层共 3 层的神经网络进行训练和预测.如图 4 所示,设定输入数据为前文筛选得出的 20 个分子描述符,即输入层神经元节点数为 20,输出层神经元节点数设置为 1[12].隐含层神经元节点数可根据经验公式(1)进行确定数量范围在 4~14,本节设置隐含层神经元节点数为 10:

表 1 10 次实验中 27 个变量出现的次数

Table 1 Number of occurrences of the 27 variables in 10 experiments

序号	变量	次数	序号	变量	次数	序号	变量	次数
1	MDEC-23	10	10	minHBint10	10	19	MLogP	7
2	minHsOH	10	11	SHBint10	10	20	VCH-5	5
3	nHBAcc	10	12	minHBint4	10	21	MDEC-33	5
4	maxHsOH	10	13	minHBint6	10	22	SsOH	4
5	minHBa	10	14	ATSc2	9	23	ATSc4	3
6	C1SP2	10	15	BCUTc-1l	8	24	MLFER_BH	2
7	WTPT-5	10	16	VPC-6	8	25	SPC-6	2
8	LipoaffinityIndex	10	17	XLogP	8	26	ndssC	1
9	SHsOH	10	18	MLFER_A	7	27	ETA_Shape_Y	1

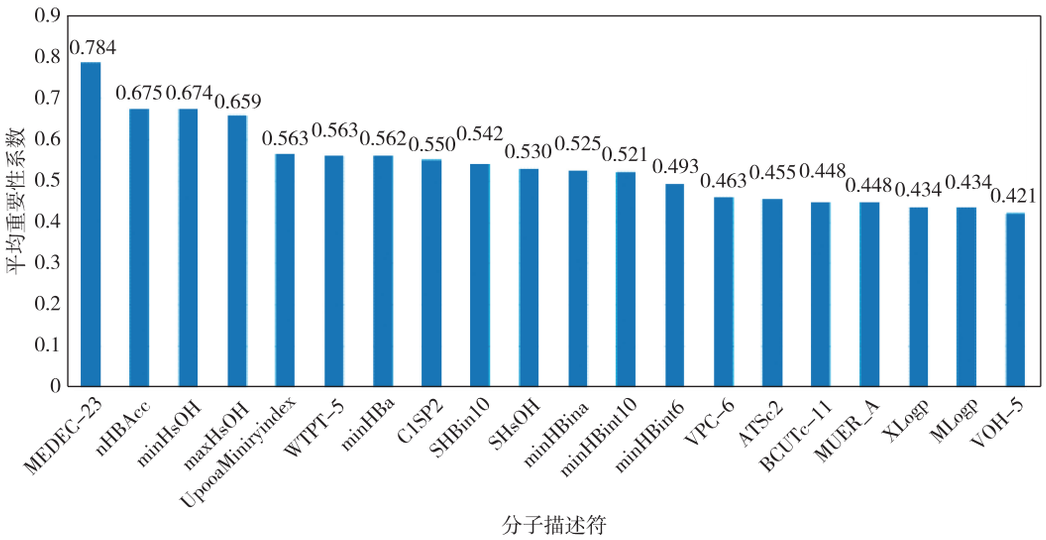

图 3　20 个分子描述符的平均重要性系数

Fig. 3　Average importance coefficients of 20 molecular descriptors

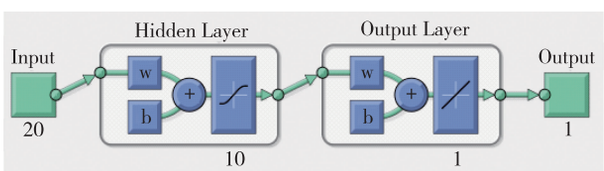

图 4　设定的 3 层 BP 神经网络结构

Fig. 4　Three-layer BP neural network structure

$$q = \sqrt{k + l} + a, \tag{1}$$

式(1)中:q 是隐含层神经元的个数;k 是输入层神经元的个数;l 是输出层神经元的个数;a 是一个固定的常数值,取值范围在 0~10 之间[13].

BP 神经网络中隐含层的激活函数为 sigmoid,输出层的激活函数为 relu,用函数式(2)、(3)表示:

$$\text{sigmoid}(z) = \frac{1}{1 + e^{-z}}, \tag{2}$$

$$\text{relu}(z) = \begin{cases} z, & z > 0, \\ 0, & z \leqslant 0. \end{cases} \tag{3}$$

用 $S_j^{[l]}$ 来表示第 l 层中第 j 个神经元的激活函数输出,ω_{jk}^{l} 表示从网络第$(l-1)$层 k 个神经元指向第 l 层第 j 个神经元的连接权重[14].用 σ 表示激活函数.

从输入层到隐含层的计算公式为

$$S_j^l = \sigma\left(\sum_{p=1}^{P} \omega_{pl} x_p + b_1\right),$$
$$p = 1,2,\cdots,P; l = 1,2,\cdots,L. \tag{4}$$

由隐含层到输出层的计算公式为

$$S_m = \sigma\left(\sum_{l=1}^{L} \omega_{lm} f_1(S_j^l) + b_2\right),$$
$$l = 1,2,\cdots,L; m = 1,2,\cdots,M. \tag{5}$$

式(4)、(5)中:b_1 和 b_2 为阈值;ω_{pl} 和 ω_{lm} 为连接权值;隐含层输出结果为 $f_1(S_l)$,f_1 为 relu 激活函数;输出层输出结果为 $f_2(S_m)$,f_2 为输出层的输出函数.

3.2　BP 神经网络求解结果分析

基于传统 BP 神经网络模型,按 8∶2的比例将 1 974 个样本数据划分成训练集和测试集,用训练集训练模型,再用训练好的模型在测试集上验证效果,训练回归结果如图 5 所示.观察可得该模型计算的拟合度为 0.820 62,其训练和测试数据较为集中.测试预测结果误差如图 6 和图 7 所示.由图 6 可知选取的 50 组测试集进行预测有所波动,出现个别误差较大的情况,但主要集中在 0.1~0.3 范围内,测试平均误差为 21.671 5%.由图 7 可知 50 组测试集所预测的 pIC_{50} 值与实际测试值有误差,其均方根误差 RMSE 为 1.416 4,决定系数 $R^2 = 0.466\ 69$.可以发现单纯通过 BP 神经网络进行模型预测虽然可以预测出一定的 pIC_{50} 值,但并不准确,应该通过相关算法对模型进行优化从而减少误差.由于粒子群优化算法(PSO)不依赖于问题信息,采用实数求解,算法的通用性强[15],容易实现且收敛速度快,因此,在追求误差较小的基础上,可以通过基于粒子群算法来优化 BP 神经网络模型进行预测.

3.3　PSO 优化 BP 神经网络模型

BP 神经网络会由于初始阈值与权值选取不合理,而导致陷入局部最优解.同时若需要进行大量的训练,极容易造成过度拟合,将在一定程度上影响泛

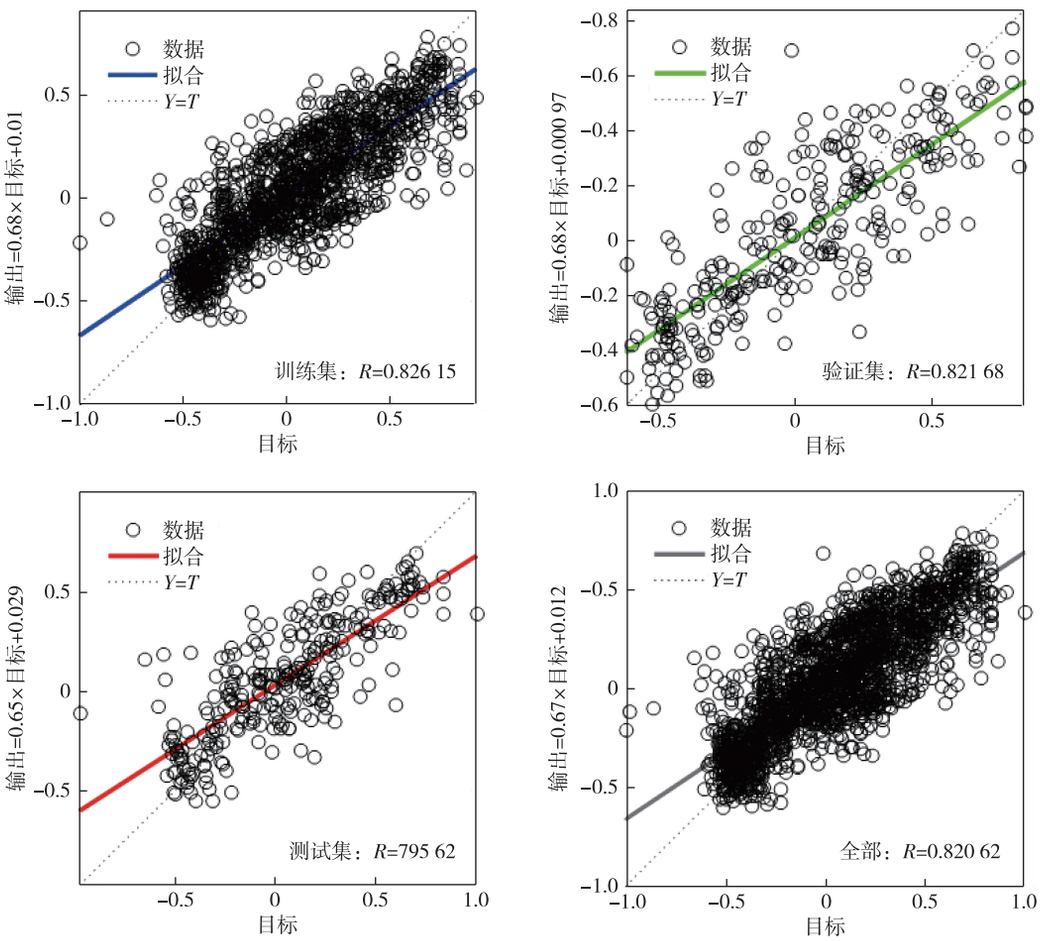

图 5　BP 神经网络训练回归结果

Fig. 5　Regression results of BP neural network training

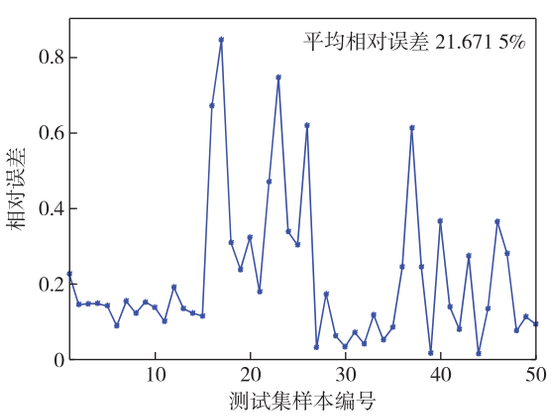

图 6　预测值和真实值的相对误差

Fig. 6　Relative error between predicted values and real values

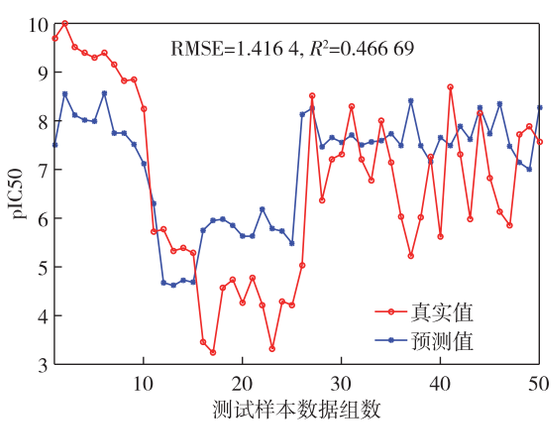

图 7　测试集预测值与真实值的对比

Fig. 7　Comparison between predicted values and real values of test set

化能力.针对 BP 神经网络的缺点,可以考虑使用遗传算法或粒子群算法对网络进行优化,本文考虑到 PSO 算法采用实数编码,比采用二进制编码的遗传算法运行速度更快,同时可利用遗传算法的变异思想增加变异算子和动态调整学习因子等来改进不足[16],避免陷入局部最优,保证种群多样化.使用的 PSO 优化 BP 神经网络算法流程如图 8 所示.

在更新粒子的速度和位置时,可以依据式(6)对粒子的位置和速度进行调整:

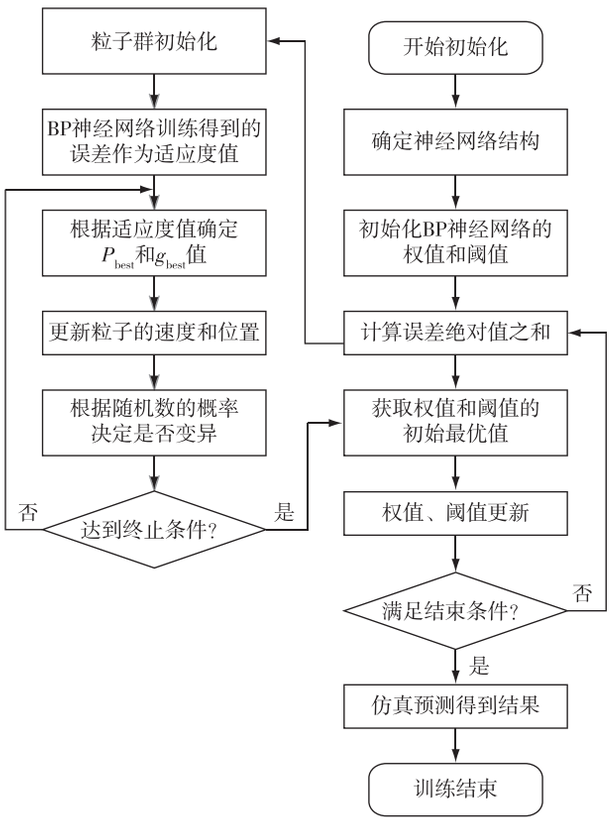

图 8　PSO 优化 BP 神经网络算法流程

Fig. 8　PSO optimized BP neural network algorithm flow

$$\begin{cases} V_i^{k+1} = \omega V_i^k + c_1(j) \times r_1 \times (P_{\text{best}} - X_i^k) + \\ \qquad c_2(j) \times r_2 \times (g_{\text{best}} - X_i^k), \\ X_i^{k+1} = X_i^k + V_i^{k+1}, \\ c_2(j) = c_{\max} - (c_{\max} - c_{\min}) \times \dfrac{(i_{t\max} - j)}{i_{t\max}}, \\ c_1(j) = 4 - c_2(j), \end{cases} \tag{6}$$

式(6)中:$c_1(j)$,$c_2(j)$ 表示进行第 j 次迭代产生的学习因子;i 表示迭代的次数;ω 表示权值系数;r_1,r_2 表示随机函数.

3.4　PSO 优化 BP 神经网络预测生物活性结果分析

基于 PSO 优化 BP 神经网络算法建立定量预测模型,同样将 1 974 个样本数据划随机分成 80%的训练集和 20%的测试集,用训练集训练,用测试集对模型进行检验.其预测结果如图 9 所示,训练集和测试集的拟合优度分别为 0.862 77 和 0.745 85,预测模型整体拟合优度为 0.833 7,比优化前的 BP 神经网络的拟合度有所提升.

PSO 优化 BP 神经网络算法测试预测结果误差如图 10 和图 11 所示.由图 10 可知测试集样本预测的平均相对误差为 9.491 3%,预测准确度有所提升,其测试集的数据相对集中.而图 11 表明均根方误差 RMSE 为 0.731 5,决定系数 R^2 = 0.740 92.相比未优化前的 BP 神经网络预测结果,其 RMSE 降低且 R^2 有所增加,说明优化后的网络预测得到的生物活性值数据更加贴近真实值,通过拟合度和误差分析论证了 PSO 优化 BP 神经网络的模型整体效果更好.

通过上文建立的化合物对 ERα 生物活性的定量预测模型,对 50 个化合物的生物活性值进行预测.在数据集中 IC_{50} 值的单位是 nmol/L,因此不能直接用 IC_{50} 值取负对数,应乘以 10 的−9 次方后再取负对数,所以 IC_{50} 与 pIC_{50} 的关系为 $IC_{50} = 10^{(9-pIC_{50})}$,而 pIC_{50} 是 IC_{50} 的转化值,并无单位.由此可得模型优化前后的预测值,但经过对比最终只选取 PSO 优化 BP 神经网络模型预测得到的 IC_{50} 值和对应的 pIC_{50} 值,详见表 2.

4　基于 PSO 优化 SVM 的 ADMET 性质预测模型分析

4.1　化合物 ADMET 性质分析及预测模型构建

一个化合物想要成为候选药物,除了需要具备良好的生物活性(即抗乳腺癌活性)外,还需要在人体内具备良好的药代动力学性质和安全性,合称为 ADMET(Absorption(吸收)、Distribution(分布)、Metabolism(代谢)、Excretion(排泄)、Toxicity(毒性))性质[17].其中,ADME 主要指化合物的药代动力学性质,描述了化合物在生物体内的浓度随时间变化的规律,T 主要指化合物可能在人体内产生的毒副作用.一个化合物的活性再好,如果其 ADMET 性质不佳,比如很难被人体吸收,或者在体内代谢速度太快,或者具有某种毒性,那么其仍然难以成为药物,因而还需要进行 ADMET 性质优化.由于建模优化的复杂程度,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),可度量化合物被人吸收的能力;2)细胞色素 P450 酶(Cytochrome P450,CYP)3A4 亚型(CYP3A4),这是人体内的主要代谢酶,可度量化合物的代谢稳定性;3)化合物心脏安全性评价(human Ether-a-go-go Related Gene,hERG),可度量化合物的心脏毒性;4)人体口服生物利用度(Human Oral Bioavailability,HOB),可度量药物进入人体后被吸收进入人体血液循环的药量比例;5)微核试验(Micronucleus,MN),是检测化合物是否具有遗传毒性的一种方法[18].为方便讨论,

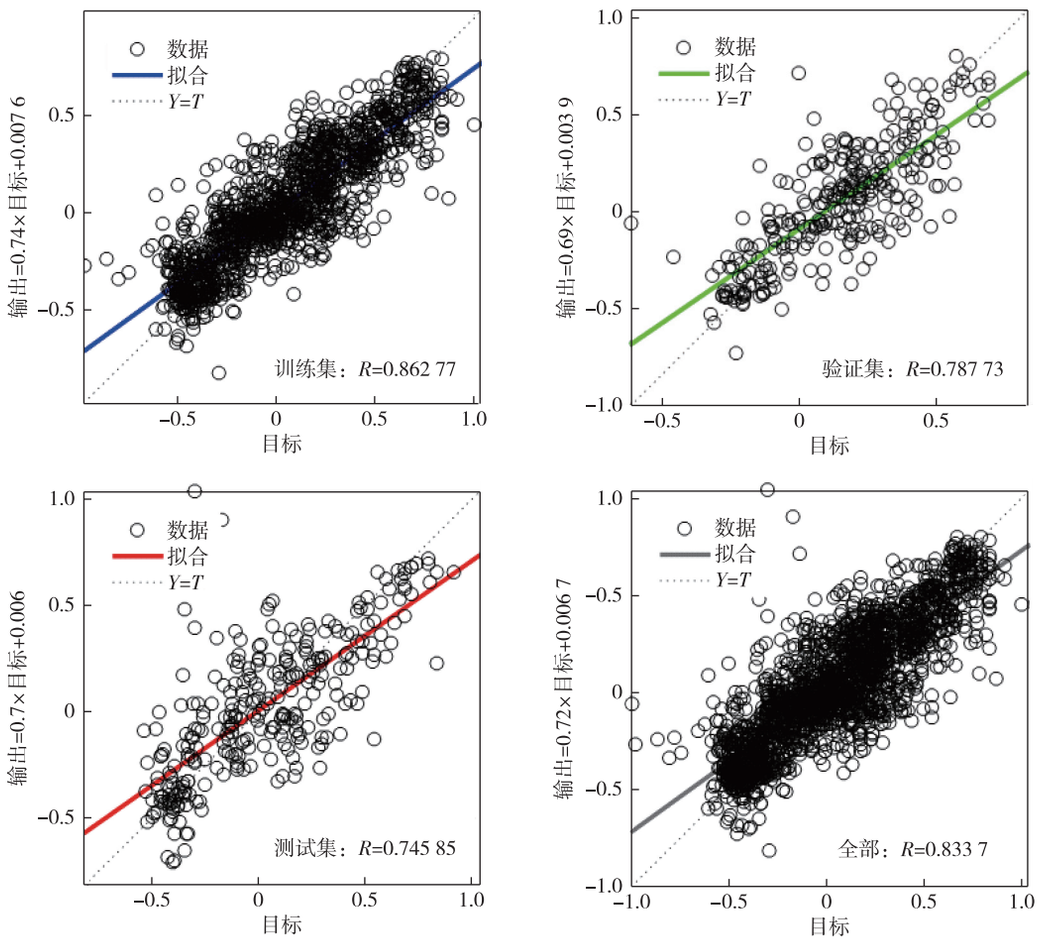

图 9　PSO 优化 BP 神经网络的训练回归结果

Fig. 9　Regression results of the PSO optimized BP neural network training

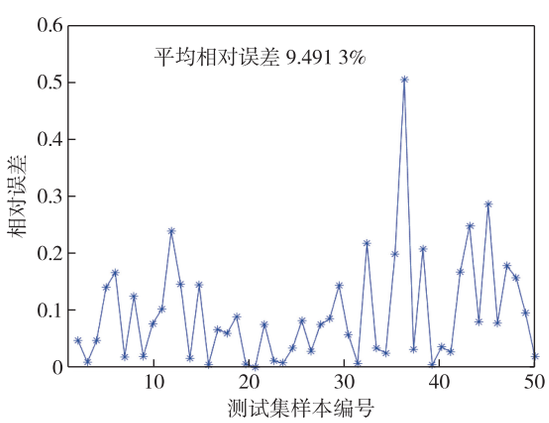

图 10　PSO 优化 BP 神经网络预测值和真实值的相对误差

Fig. 10　Relative error between predicted values and real values of the PSO optimized BP neural network

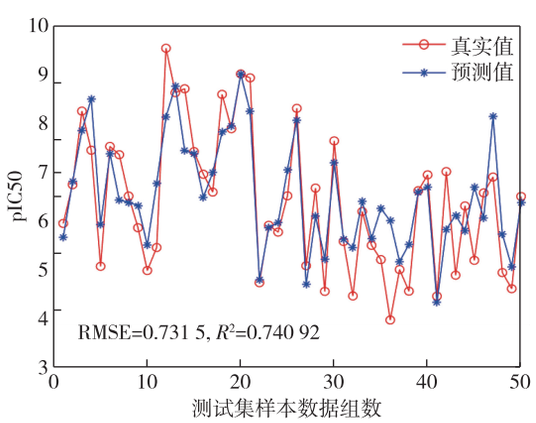

图 11　PSO 优化 BP 神经网络测试集预测值与真实值的对比

Fig. 11　Comparison between predicted values and real values of the PSO optimized BP neural network on test set

本文统一使用二分类法提供 ADMET 性质的相应取值,比如对于 Caco-2:“1”代表该化合物的小肠上皮细胞渗透性较好,“0”代表该化合物的小肠上皮细胞渗透性较差.其他 4 个的二分类法可依此类推.

由于收集到的 ADMET 性质数据样本量有限,且具有非线性及维数较多的特点,在收集过程中易受到操作环境等复杂因素的影响,使得数据具有较高的含噪性且容易出现缺失和错漏,因此在选用数据

表 2　IC_{50}值及对应 pIC_{50}值的最优预测结果

Table 2　Prediction result of IC_{50} values and corresponding pIC_{50} values

序号	IC_{50}/(nmol/L)	pIC_{50}	序号	IC_{50}/(nmol/L)	pIC_{50}
1	26.858 68	7.570 915	26	371.272 90	6.430 307
2	68.983 95	7.161 252	27	95.477 43	7.020 099
3	55.911 54	7.252 499	28	476.006 30	6.322 387
4	36.283 74	7.440 288	29	397.218 80	6.400 970
5	14.734 88	7.831 654	30	1 817.580 00	5.740 507
6	68.249 47	7.165 901	31	9 911.869 00	5.003 844
7	43.991 17	7.356 635	32	8 743.830 00	5.058 298
8	39.354 82	7.405 002	33	10 230.940 00	4.990 085
9	33.241 72	7.478 317	34	10 205.750 00	4.991 155
10	33.900 47	7.469 794	35	16 024.880 00	4.795 205
11	32.427 80	7.489 082	36	210.626 60	6.676 487
12	46.355 03	7.333 903	37	186.963 10	6.728 244
13	28.091 03	7.551 432	38	274.515 80	6.561 433
14	33.323 70	7.477 247	39	234.068 20	6.630 658
15	27.722 76	7.557 163	40	263.636 80	6.578 994
16	21.511 10	7.667 337	41	248.628 80	6.604 449
17	48.613 36	7.313 244	42	248.628 80	6.604 449
18	217.380 40	6.662 780	43	226.389 90	6.645 143
19	70.141 01	7.154 028	44	336.138 50	6.473 482
20	14.294 85	7.844 820	45	248.628 80	6.604 449
21	74.874 38	7.125 667	46	46.834 08	7.329 438
22	208.536 60	6.680 818	47	57.342 58	7.241 523
23	381.135 30	6.418 921	48	71.946 71	7.142 989
24	204.318 80	6.689 692	49	153.608 20	6.813 586
25	348.628 30	6.457 637	50	53.156 35	7.274 445

挖掘算法进行分析预测时,需考虑算法的适用性.经过比较分析几个常用算法发现:朴素贝叶斯算法对输入数据的表达形式很敏感,分类决策存在一定的错误率,其训练效率低且运算框架比较复杂,不适用于化合物的 ADMET 性质预测;决策树算法在处理特征关联性比较强的数据时表现一般,容易出现过拟合;支持向量机 SVM 算法的最终决策函数只由少数的支持向量所确定,计算的复杂性取决于支持向量的数目,而不是样本空间的维数,避免了“维数灾难”,且对非线性分类任务的可解释性强,能找出至关重要的关键样本,算法拟合精度较高,具有较好的鲁棒性,对化合物 ADMET 性质预测具有较强的适用性.故采用支持向量机建立出 5 种 ADMET 性质各自的 0-1 二分类模型.但该方法的参数核函数 g 和惩罚参数 c 的选取问题会限制其进一步的发展.根据现有研究表明,截至目前还没有一种比较好的、公认的、固定的参数选取方法.一般而言,经验估计法是研究中最常使用的方法,但是该方法在选取参数时比较随机,会产生较大的局限性.而粒子群算法(PSO)在参数寻优求解过程中拥有比较显著的优势,并且该方法的模型结构相对简单[19],因此,本文将选择使用粒子群算法优化支持向量机参数,其算法的运行流程如图 12 所示.

在使用 PSO 优化 SVM 方法来计算各粒子的适应度值时,适应度函数取均方误差(MSE),如式(7)所示:

$$\mathrm{MSE}=\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2, \tag{7}$$

式(7)中:y_i 是实际取值;$\hat{y}_i$ 是预测取值;n 是训练的样本个数.

4.2　基于 PSO 优化 SVM 的分类预测结果分析

基于上述 PSO 优化 SVM 算法构建化合物的 ADMET 预测模型,分别对 5 个指标进行预测分析,依次设立输出变量指标分别为 Caco-2、CYP3A4、

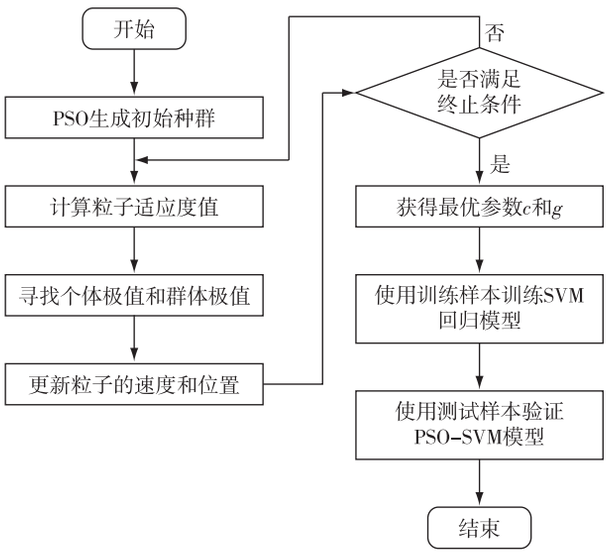

图 12　PSO 优化 SVM 算法流程

Fig. 12　Flow chart of PSO optimized SVM algorithm

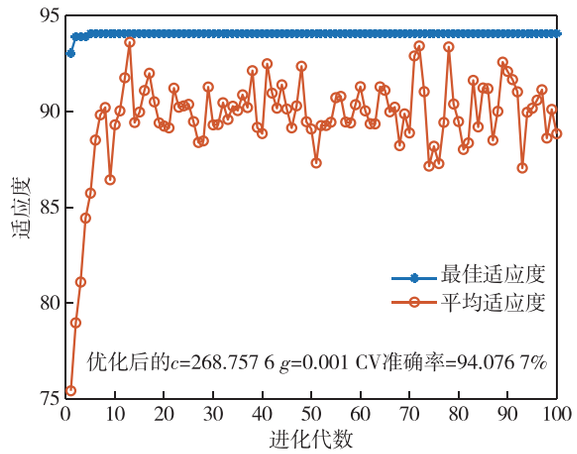

图 13　PSO 优化 SVM 的 Caco-2 迭代过程

Fig. 13　PSO optimizing the Caco-2 iterative process of SVM

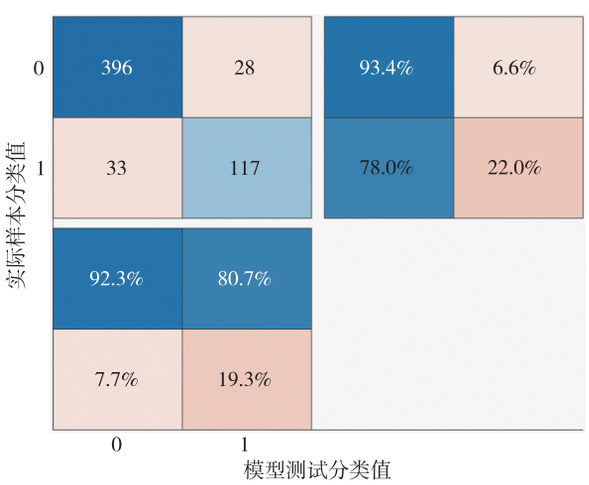

图 14　Caco-2 测试组数据的混淆矩阵

Fig. 14　Confusion matrix of Caco-2 test group data

hERG、HOB、MN,代入 MATLAB 软件中运行.

4.2.1　化合物的小肠上皮细胞渗透性 Caco-2 预测分析

对于指标 Caco-2 的预测,图 13 表示 PSO 优化 SVM 的 Caco-2 迭代过程,可得到优化后的惩罚参数 $c=268.757\ 6$ 和核参数 $g=0.001$,交叉验证 CV 准确率达到 94.076 7%,准确性较好,对 Caco-2 指标预测具备一定参考价值.图 14 表示 574 个测试数据的混淆矩阵,其中有 396 个化合物的实际样本分类值和模型预测分类值均为“0”,117 个化合物的实际样本分类值和模型预测分类值均为“1”,混淆矩阵的精确度为 80.7%,召回率为 78.0%,特异度为 93.4%.图 15 表示 PSO 优化 SVM 后的实际分类与预测分类对比情况[20],对于 574 个测试数据的 Caco-2 的真实值和预测值大部分相互吻合,其预测准确度为 89.372 8%.

4.2.2　化合物的代谢稳定性 CYP3A4 预测分析

对于指标 CYP3A4 的预测,图 16 表示其迭代过程,优化后的惩罚参数 $c=549.464\ 9$ 和核参数 $g=0.001$,交叉验证迭代过程中 CV 的准确率为 97.735 2%,具有较好的精度.图 17 表示 574 个测试数据的混淆矩阵,其中有 59 个化合物的实际样本分类值和模型预测分类值均为“0”,481 个化合物的实际样本分类值和模型预测分类值均为“1”,混淆矩阵的精确度为 97.0%,召回率为 96.2%,特异度为 79.7%.图 18 为预测 CYP3A4 指标时测试集的实际分类和预测分类结果[21],测试集的实际分类和预测分类也相对较高,其预测准确度为 94.076 7%.

4.2.3　化合物的心脏毒性 hERG 预测分析

对于指标 hERG 的预测,图 19 表示其迭代过程,优化后的惩罚参数 $c=891.311\ 9$ 和核参数 $g=0.001$,交叉验证迭代过程中 CV 准确率为 89.198 6%,精度一般.图 20 表示 574 个测试数据的混淆矩阵,其中有 93 个化合物的实际样本分类值和模型预测分类值均为“0”,390 个化合物的实际样本分类值和模型预测分类值均为“1”,混淆矩阵的精确度为 84.4%,召回率为 95.4%,特异度为 56.4%.图 21 为预测 hERG 指标时测试集的实际分类和预测分类结果[22],测试集的实际分类和预测分类也相对较高,其预测准确度为 84.146 3%.

4.2.4　化合物的 HOB 预测分析

对于指标 HOB 的预测,图 22 表示其迭代过程,

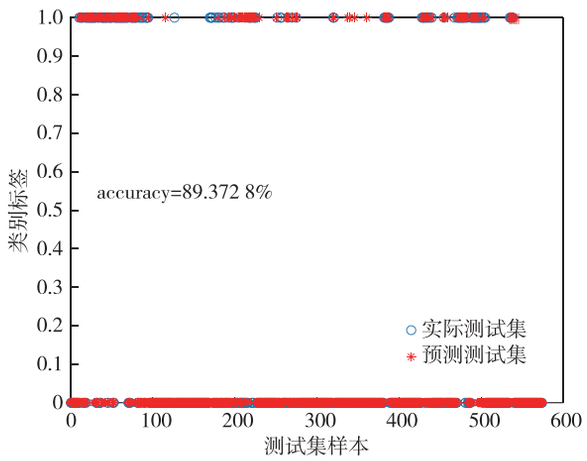

图 15　预测 Caco-2 指标时测试集的实际分类和预测分类

Fig. 15　The actual classification and predicted classification for the test set when predicting the Caco-2 indicator

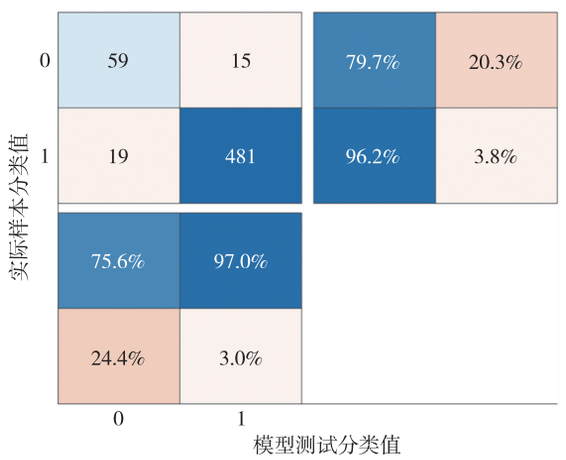

图 17　CYP3A4 测试组数据的混淆矩阵

Fig. 17　Confusion matrix of CYP3A4 test group data

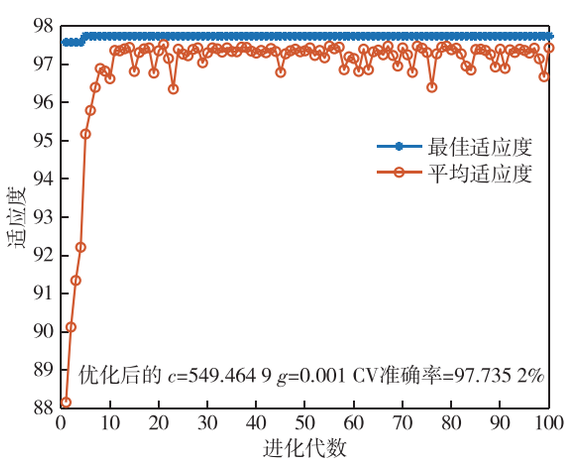

图 16　PSO 优化 SVM 的 CYP3A4 迭代过程

Fig. 16　PSO optimizing the CYP3A4 iterative process of SVM

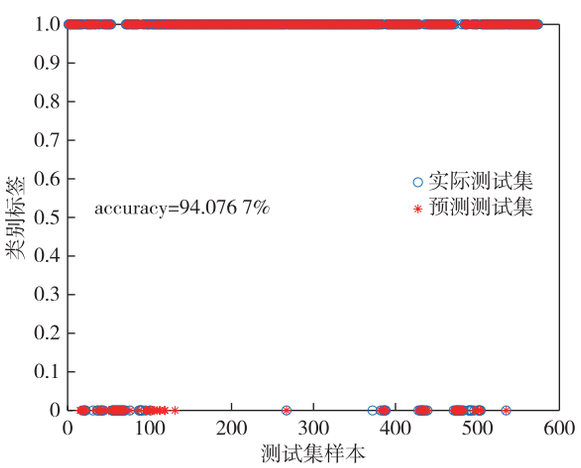

图 18　预测 CYP3A4 指标时测试集的实际分类和预测分类

Fig. 18　The actual classification and predicted classification for the test set when predicting CYP3A4 indicators

优化后的惩罚系数 c = 119.618 4 和核参数 g = 0.001,交叉验证迭代过程中的 CV 准确率为 87.971 9%,精度一般.图 23 表示 574 个测试数据的混淆矩阵,其中有 394 个化合物的实际样本分类值和模型预测分类值均为“0”,60 个化合物的实际样本分类值和模型预测分类值均为“1”,混淆矩阵的精确度为 50%,召回率为 50%,特异度为 86.8%.图 24 为预测 HOB 指标时测试集的实际分类和预测分类结果[23],测试集的实际分类和预测分类也相对较高,其预测准确度为 79.094 1%.

4.2.5　化合物的遗传毒性 MN 预测分析

对于指标 MN 的预测,图 25 表示其迭代过程,优化后的惩罚系数 c=63.284 6 和核参数 g=0.001,交叉验证迭代过程中的 CV 准确率为 92.508 7%,精度一般.图 26 表示 574 个测试数据的混淆矩阵,其中有 104 个化合物的实际样本分类值和模型预测分类值均为“0”,381 个化合物的实际样本分类值和模型预测分类值均为“1”,混淆矩阵的精确度为 86.4%,召回率为 92.9%,特异度为 63.4%.图 27 为预测 MN 指标时测试集的实际分类和预测分类结果[24].测试集的实际分类和预测分类也相对较高,其预测准确度为 84.494 8%.

根据前文所构建的 Caco-2、CYP3A4、hERG、HOB、MN 的分类预测模型,由于所建模型预测准确度都相对较高,即可由化合物分子的结构式对 50 个新化合物的 ADMET 性质进行相应预测,从而判断新化合物的性质好坏,对药物性质判断提供一定的参考价值,预测结果如表 3 所示.

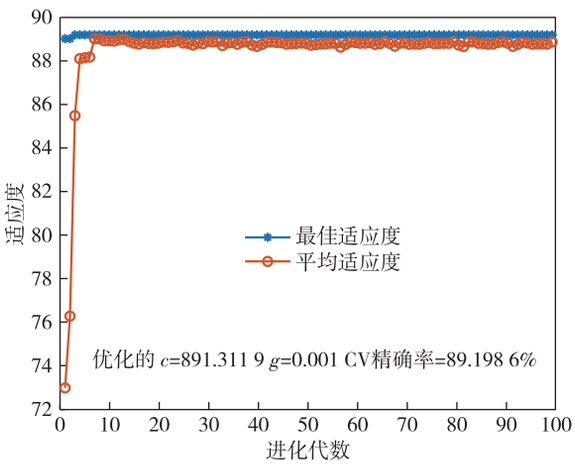

图 19　PSO 优化 SVM 的 hERG 迭代过程

Fig. 19　PSO optimizing the hERG iterative process of SVM

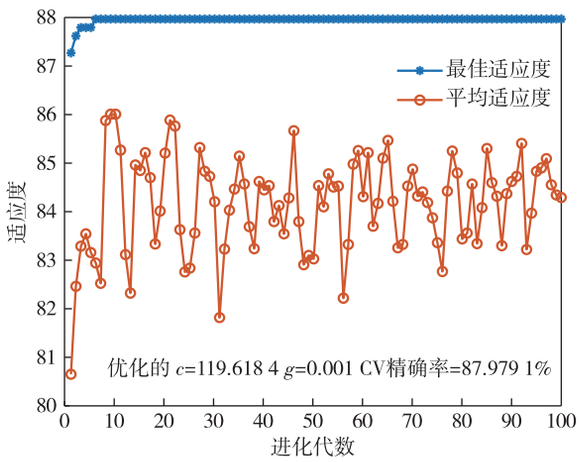

图 22　PSO 优化 SVM 的 HOB 迭代过程

Fig. 22　PSO optimizing the HOB iterative process of SVM

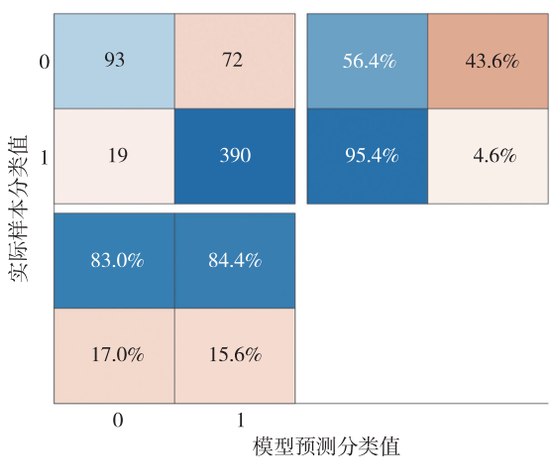

图 20　hERG 测试组数据的混淆矩阵

Fig. 20　Confusion matrix of hERG test group data

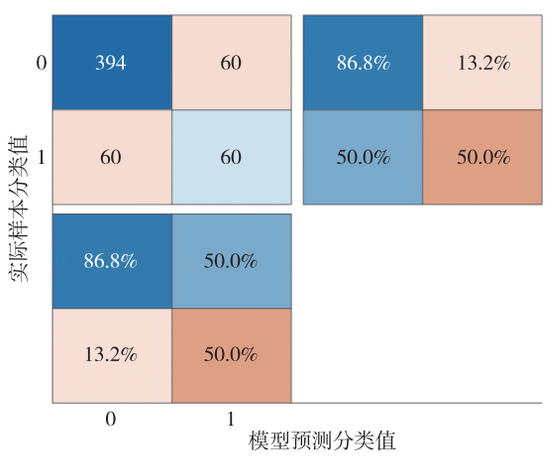

图 23　HOB 测试组数据的混淆矩阵

Fig. 23　Confusion matrix of HOB test group data

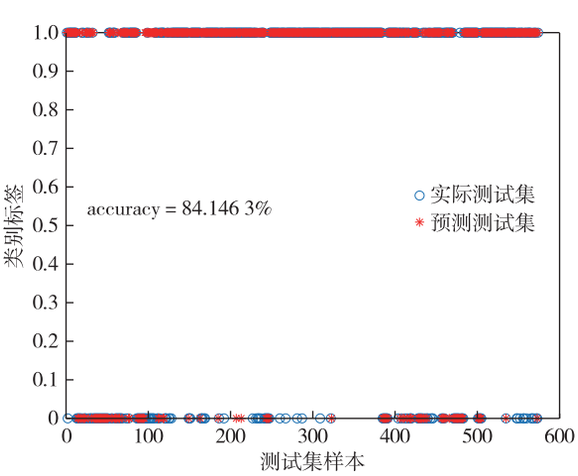

图 21　预测 hERG 指标时测试集的实际分类和预测分类

Fig. 21　The actual classification and predicted classification for the test set when predicting hERG indicators

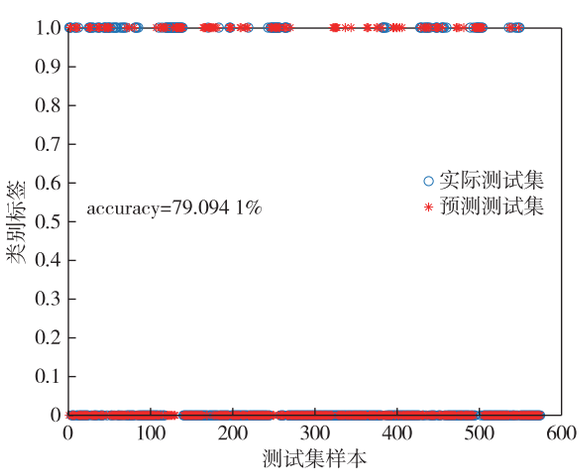

图 24　预测 HOB 指标时测试集的实际分类和预测分类

Fig. 24　The actual classification and predicted classification for the test set when predicting the HOB index

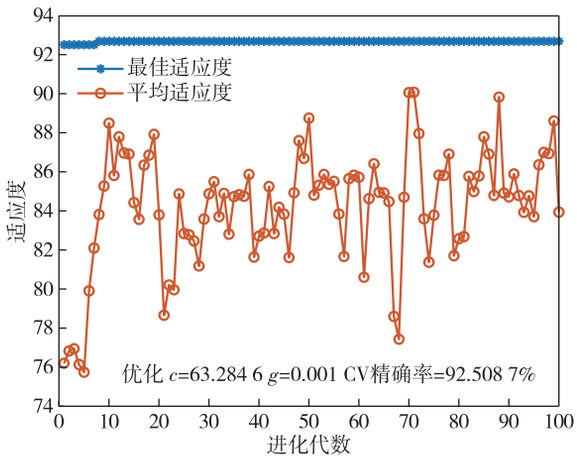

图 25　PSO 优化 SVM 的 MN 迭代过程

Fig. 25　PSO optimizing the MN iterative process of SVM

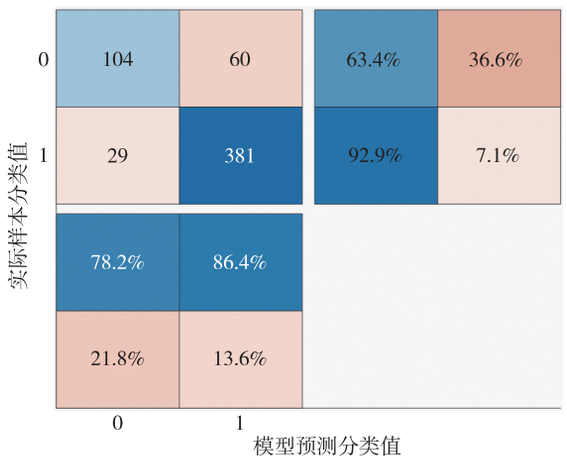

图 26　MN 测试组数据的混淆矩阵

Fig. 26　Confusion matrix of MN test group data

表 3　50 个化合物的 ADMET 性质预测结果

Table 3　ADMET property prediction results of 50 compounds

序号	Caco-2	CYP3A4	hERG	HOB	MN	序号	Caco-2	CYP3A4	hERG	HOB	MN
1	0	1	1	0	1	26	1	1	1	0	0
2	0	1	1	0	0	27	0	1	1	0	0
3	0	1	1	0	0	28	0	1	1	0	1
4	0	1	1	0	0	29	0	1	1	0	1
5	0	0	1	0	0	30	0	1	1	1	1
6	0	1	1	0	0	31	1	1	1	1	1
7	0	1	1	0	0	32	1	1	1	1	1
8	0	1	1	0	0	33	1	1	1	1	1
9	0	1	1	0	0	34	1	1	1	1	1
10	0	1	1	0	1	35	0	1	1	1	1
11	0	1	1	0	1	36	0	1	1	0	0
12	0	1	1	0	1	37	0	1	1	0	0
13	0	1	1	0	1	38	0	1	1	0	0
14	0	1	1	0	1	39	0	1	0	1	1
15	0	1	1	0	1	40	0	1	1	1	1
16	0	1	1	0	1	41	0	1	1	1	1
17	0	1	1	0	0	42	0	1	1	1	1
18	0	1	0	0	0	43	0	1	0	1	1
19	1	1	1	0	0	44	0	1	1	1	1
20	0	0	1	0	0	45	0	1	1	1	1
21	0	1	1	0	0	46	0	1	1	0	1
22	0	1	1	0	0	47	0	1	1	0	1
23	1	0	1	0	0	48	0	1	1	0	1
24	1	0	1	0	0	49	0	1	1	0	1
25	1	1	1	0	1	50	0	1	1	0	0

5　结论

针对抗乳腺癌候选药物研发过程中的生物活性和 ADMET 性质预测问题,本文选择利用计算机辅助方法.从化合物的“特征重要性分析”角度出发,首先采用随机森林分类器对 1 974 种化合物进行特征重

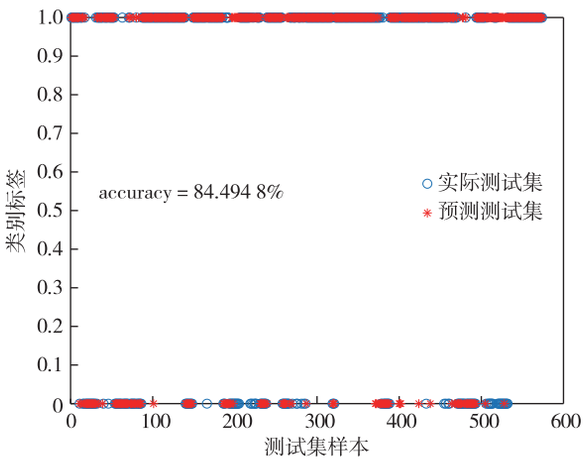

图 27 预测 MN 指标时测试集的实际分类和预测分类

Fig. 27 The actual classification and predicted classification for the test set when predicting the MN index

要性评估,从而将分子描述符对生物活性影响的重要性进行重新排序,筛选出对生物活性最具显著影响的前 20 个分子描述符.其次利用粒子群优化 BP 神经网络构建定量预测模型求取 50 个化合物的 IC_{50}和 pIC_{50}值,模型拟合度为 0.833 7,对比优化前的 BP 神经网络,其 RMSE 值降低且 R^2 有所提高,优化后的生物活性预测值更贴近真实值.再者结合粒子群优化支持向量机算法构建化合物 ADMET 性质 5 个指标 Caco-2、CYP3A4、hERG、HOB、MN 的分类预测模型,进行训练和测试得到交叉验证 CV 准确率达到 94.076 7%,准确性较好.5 个指标的模型预测准确率分别为 89.372 8%、94.067 7%、84.146 3%、79.094 1%、84.494 8%,求得 50 个化合物的 ADMET 二分类法的取值.

研究表明文中所构建的预测模型比基准模型的预测效果更好,验证了模型的适用性.通过对化合物分子描述符的预测分析能够在抗乳腺癌候选药物研制方面提供有效的借鉴作用,所建立的模型还可以拓宽到求解其他关于数据分析预测和多目标优化等实际问题中,在防治抗击乳腺癌、白血病、宫颈癌或其他肿瘤疾病等人体生命健康的研究具有一定的指导作用[25].

参考文献

References

[1] Chan H C S,Shan H B,Dahoun T,et al.Advancing drug discovery via artificial intelligence[J].Trends in Pharmacological Sciences,2019,40(8):592-604

[2] Shen C,Ding J J,Wang Z,et al.From machine learning to deep learning:advances in scoring functions for protein-ligand docking[J].Wiley Interdisciplinary Reviews:Computational Molecular Science,2020,10(1):e1429

[3] 顾耀文,张博文,郑思,等.基于图注意力网络的药物 ADMET 分类预测模型构建方法[J].数据分析与知识发现,2021,5(8):76-85
GU Yaowen,ZHANG Bowen,ZHENG Si,et al.Predicting drug ADMET properties based on graph attention network[J].Data Analysis and Knowledge Discovery,2021,5(8):76-85

[4] 谢良旭,李峰,谢建平,等.基于融合神经网络模型的药物分子性质预测[J].计算机科学,2021,48(9):251-256
XIE Liangxu,LI Feng,XIE Jianping,et al.Predicting drug molecular properties based on ensembling neural networks models[J].Computer Science,2021,48(9):251-256

[5] 秦洁.基于矩阵补全的药物前体分子生物活性预测方法研究[D].南京:南京邮电大学,2020
QIN Jie.Research on matrix completion with side information for better modeling bioactivates of drug leads[D].Nanjing:Nanjing University of Posts and Telecommunications,2020

[6] 贾聪敏.基于分子振动特征的药物靶点识别及活性预测模型研究[D].北京:北京中医药大学,2019
JIA Congmin.Study on drug target recognition and activity prediction model based on molecular vibration characteristics[D].Beijing:Beijing University of Chinese Medicine,2019

[7] 沈杰.药物 ADMET 理论预测方法开发和靶向雌激素受体的药物设计研究[D].上海:华东理工大学,2011
SHEN Jie.Development of drug ADMET theory prediction method and drug design research targeting estrogen receptor[D].Shanghai:East China University of Science and Technology,2011

[8] Wenzel J,Matter H,Schmidt F.Predictive multitask deep neural network models for ADME-tox properties:learning from large data sets[J].Journal of Chemical Information and Modeling,2019,59(3):1253-1268

[9] Lei T L,Sun H Y,Kang Y,et al.ADMET evaluation in drug discovery.18.reliable prediction of chemical-induced urinary tract toxicity by boosting machine learning approaches[J].Molecular Pharmaceutics,2017,14(11):3935-3953

[10] 路珩,张一奇.雄激素受体在雌激素受体阳性乳腺癌患者中的表达及其临床意义[J].中国现代医学杂志,2021,31(18):55-59
LU Heng,ZHANG Yiqi.Expression and significance of androgen receptor in estrogen receptor-positive breast cancer[J].China Journal of Modern Medicine,2021,31(18):55-59

[11] 丛斌斌,王永胜.激素受体阳性早期乳腺癌治疗现状与挑战[J].中国癌症杂志,2021,31(8):689-696
CONG Binbin,WANG Yongsheng.Treatment landscape and challenges of managing the hormone receptor-positive early breast cancer[J].China Oncology,2021,31(8):689-696

[12] Wu Z Q,Ramsundar B,Feinberg E N,et al.MoleculeNet:

a benchmark for molecular machine learning [J]. Chemical Science,2017,9(2):513-530
[13] 杨德俊,姚香草,许重远,等.红茴香小分子化合物降尿酸活性及 ADMET 性质的分子对接[J].中国临床药理学杂志,2018,34(23):2750-2752,2777
YANG Dejun,YAO Xiangcao,XU Zhongyuan,et al.Molecular docking of the chemicals of *Illicium lanceolatum* in lowering uric acid and ADMET properties [J]. The Chinese Journal of Clinical Pharmacology, 2018, 34(23):2750-2752,2777
[14] 张翠锋,谢海棠,潘国宇.大分子药物的吸收、分布、代谢、排泄和毒性特征及药代模型的应用[J].药学学报,2016,51(8):1202-1208
ZHANG Cuifeng,XIE Haitang,PAN Guoyu.Absorption, distribution, metabolism, excretion and toxicity of biologics and its application in pharmacokinetic modeling [J].Acta Pharmaceutica Sinica,2016,51(8):1202-1208
[15] Mansouri K,Cariello N F,Korotcov A,et al.Open-source QSAR models for pKa prediction using multiple machine learning approaches [J]. Journal of Cheminformatics, 2019,11(1):60
[16] 陈宪.基于 OECD 准则对 QSAR/QSPR 模型几个重要问题的研究[D].长沙:中南大学,2013
CHEN Xian.Studies on a few key problems of QSAR/QSPR modeling based on the OECD principles [D]. Changsha:Central South University,2013
[17] Shar P A,Tao W Y,Gao S,et al.Pred-binding:large-scale protein-ligand binding affinity prediction [J]. Journal of Enzyme Inhibition and Medicinal Chemistry, 2016,31(6):1443-1450
[18] 苏敏仪,刘慧思,林海霞,等.应用机器学习方法构建药物分子解离速率常数的预测模型[J].物理化学学报,2020,36(1):179-187
SU Minyi,LIU Huisi,LIN Haixia,et al.Machine-learning model for predicting the rate constant of proteinligand dissociation[J].Acta Physico-Chimica Sinica,2020,36(1):179-187
[19] 刘光徽,胡俊,於东军.基于多视角特征组合与随机森林的 G 蛋白偶联受体与药物相互作用预测[J].南京理工大学学报,2016,40(1):1-9
LIU Guanghui,HU Jun,YU Dongjun.Predicting GPCR-drug interactions with multi-view feature combination and random forest [J]. Journal of Nanjing University of Science and Technology,2016,40(1):1-9
[20] 李小强,莫淼,吴菲,等.基于问卷调查的上海女性乳腺癌人工神经网络预测模型[J].肿瘤,2018,38(9):883-893
LI Xiaoqiang,MO Miao,WU Fei,et al.Artificial neural network models based on questionnaire survey for prediction of breast cancer risk among Chinese women in Shanghai[J].Tumor,2018,38(9):883-893
[21] 刘雅琴,王成,章鲁.基于神经网络的乳腺癌生存预测模型[J].中国生物医学工程学报,2009,28(2):221-225
LIU Yaqin,WANG Cheng,ZHANG Lu.Neural network based models for predicting breast cancer survivability [J].Chinese Journal of Biomedical Engineering,2009,28(2):221-225
[22] 闵倩,廖俊,陆涛.基于大型药物数据库的药物相互作用预测模型[J].中国临床药理学杂志,2016,32(11):1034-1036
MIN Qian,LIAO Jun,LU Tao.Drug-drug interaction predicting model based on large scale drug databases[J]. The Chinese Journal of Clinical Pharmacology,2016,32(11):1034-1036
[23] 汤井田,曹扬,肖嘉莹,等.基于粒子群优化支持向量机的瑞芬太尼血药浓度预测模型[J].中国药学杂志,2013,48(16):1394-1399
TANG Jingtian,CAO Yang,XIAO Jiaying,et al.Remifentanil blood concentration forecast model based on support vector machine with particle swarm optimization[J].Chinese Pharmaceutical Journal,2013,48(16):1394-1399
[24] 白茹,滕奇志,杨晓敏,等.基于 SVM 和 GA 的药物与人血清白蛋白结合的预测[J].计算机工程与应用,2009,45(12):226-228,248
BAI Ru,TENG Qizhi,YANG Xiaomin,et al.Prediction of combinative activity of drugs and human serum albumin by using SVM and GA [J]. Computer Engineering and Applications,2009,45(12):226-228,248
[25] 袁仙琴.基于基因表达数据的化合物肝毒性 SVM 预测模型研究[D].镇江:江苏大学,2018
YUAN Xianqin.Study on SVM prediction model of compound hepatotoxicity based on gene expression data [D]. Zhenjiang:Jiangsu University,2018

Prediction of properties of anti-breast cancer drugs based on PSO-BP neural network and PSO-SVM

XU Meixian[1] ZHENG Yan[1] LI Yanju[1] WU Weihao[1]
1 College of Automobile and Traffic Engineering,Nanjing Forestry University,Nanjing 210037

Abstract The process of screening and developing new drugs through experiments is very slow and requires a lot of manpower and material resources,and the use of computer-aided prediction of the molecular properties of drugs can greatly save time and cost of drug development.Therefore,in order to enable anti-breast cancer candidate drugs to have good biological activity and ADMET properties for inhibiting ERα,the random forest classifier was first used

for the collected 1 974 compounds to screen the top 20 molecular descriptors with the most significant effects on biological activity.Then a QSAR model was established using this and pIC50 value as characteristic data.The biological activity values of 50 new compounds were predicted via the PSO optimized BP neural network,with the model fit of 0.833 7 and the root mean square error of 0.731 5,which were more consistent with the actual values than the predicted results of the BP neural network.Subsequently,in order to improve the success rate of drug development,the ADMET classification prediction model was constructed using PSO to optimize the SVM based on the existing ADMET property data. The algorithm cross-validation CV accuracy rate reached 94.076 7%, and the prediction accuracy rates of the five index models were all above 79%.The results show that the proposed model has better prediction performance than the benchmark model,and the adopted prediction strategy is effective,which can provide reference for the discovery and development of anti-breast cancer drugs.

Key words anti breast cancer drugs;biological activity;ADMET properties;particle swarm optimization (PSO);BP neural network;support vector machines (SVM)