文章编号：2095-2163(2022)07-0084-06 中图分类号：TP181；R917 文献标志码：A

基于改进的 PCA 和 ISSA-BPNN 的定量构效关系预测模型

陈 强[1]，王登文[1]，铁治欣[1,2]，洪 亮[3]
（1 浙江理工大学 信息学院，杭州 310018；2 浙江理工大学 科技艺术学院，浙江 绍兴 312369；
3 浙江传媒学院 媒体工程学院，杭州 310018）

摘 要：为提高药物研发的效率，通常使用定量构效关系（QSAR）模型来预测化合物的生物活性，从而进行筛选和优化。目前，基于统计分析的 QSAR 随着变量急剧增多变得束手无策，同时预测精度还有提高的空间。基于此，本文提出了一种基于改进的 PCA 算法对变量进行降维，并利用改进的麻雀搜索算法优化 BP 神经网络（ISSA-BPNN），以此提高预测的精度。改进的 PCA 算法先基于 Pearson、最大互信息系数（MIC）和随机森林（RF）的加权得分得到主要特征变量，再用 PCA 算法对原特征进行降维得到主要输入变量；ISSA-BPNN 算法优化 BPNN 的权值和阈值，达到输出稳定和保证全局收敛。以乳腺癌治疗时，化合物对 ERα 的生物活性数据为例进行了训练和预测。结果表明：本文所提算法预测精度更高，为药物研发提供了一种有效方法。

关键词：BP 神经网络；最大互信息系数；随机森林；SVR；XGBoost

Quantitative structure-activity relationship prediction model based on improved PCA and ISSA-BPNN

CHEN Qiang[1], WANG Dengwen[1], TIE Zhixin[1,2], HONG Liang[3]
(1 School of Information Science and Technology, Zhejiang Sci-Tech University, Hangzhou 310018, China;
2 Keyi College of Zhejiang Sci-Tech University, Shaoxing Zhejiang 312369, China)
3 College of Media Engineering, Communication University of Zhejiang, Hangzhou 310018, China)

【Abstract】 In order to improve the efficiency of drug research and development, quantitative structure - activity relationship (QSAR) models are usually used to predict and analyze the bioactivity of compounds to screen and optimize compounds. At present, QSAR based on statistical analysis has become ineffective with the rapid increase of variables, and there is still room for improvement in prediction accuracy. Based on this, we propose an improved PCA algorithm in this paper to reduce the dimensionality of variables and an improved sparrow search algorithm to optimize the back propagation neural network (ISSA-BPNN) to improve the accuracy of prediction. The improved PCA algorithm is based on weighted scoring algorithm of Pearson, maximum information coefficient (MIC) and random forest (RF) to obtain the main feature variables. Then the PCA algorithm is used to reduce the dimensionality of the original features to get the main input variables. The ISSA-BPNN algorithm optimize the weights and thresholds of BPNN, achieving output stability and ensuring global convergence. Taking the biological activity data of compounds on ERα during breast cancer treatment as an example for training and prediction, compared with several other algorithm, the results show that the algorithm proposed in this article has higher prediction accuracy. It provides an effective method for drug research and development.

【Key words】 BP neural network; maximum information coefficient; random forest; SVR; XGBoost

0 引 言

据近几年全球癌症统计数据表明，乳腺癌症发病率和死亡病例逐渐增加，其防治须引起人们高度重视[1]。临床、流行病学和生物学证据表明，雌激素参与了乳腺癌的发生和发展[2]。雌激素化合物的大多数生理功能，在基因调控水平上主要由雌激素受体（ER）调节，这些蛋白质在细胞核中发挥作用，控制着各种器官系统的关键生理功能，并通过与相关的 DNA 调控序列相结合，来调节特定靶基因的转录[3]。雌激素受体 α 亚型（Estrogen receptors alpha，ERα）在乳腺癌病中起着至关重要的作用[4]，但在正常乳腺上皮细胞中极少被表达。通过使用选择性雌激素受体调节剂（SERM）和雌激素受

基金项目：国家自然科学基金（61671407）。

作者简介：陈 强（1995-），男，硕士研究生，主要研究方向：数据挖掘和图像配准；王登文（1998-），男，硕士研究生，主要研究方向：机器学习与 3D 视觉领域；铁治欣（1972-），男，博士，副教授，硕士生导师，主要研究方向：数据挖掘、嵌入式系统、电力系统自动化等；洪亮（1973-），男，博士，讲师，主要研究方向：模式识别、光学测量。

通讯作者：铁治欣 Email：tiezx@ zstu.edu.cn

收稿日期：2022-01-06

体降解剂(SERD)，可用来降低 ERα 的稳定性[5]。

目前，在药物研发中，为了节约时间和成本，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。这种定量构效关系(Quantitative Structure Activity Relationship，QSAR)方法是一种预选工具，旨在减少化合物的数量，并增加选择候选药物的可能性。其以一系列分子结构描述符作为自变量，化合物的生物活性作为因变量建立模型，根据可测量的物理、化学参数，精确预测化合物的生物活性[6]，或者对已有活性化合物的结构进行优化，QSAR 本质上是数据驱动模型[7]。近年来，人工智能、机器学习、大数据等技术的发展，为 QSAR 带来了挑战和机遇，通过成千上万的化学结构数据集，为药物的生物活性和安全性进行更精确的回归和分类预测带来了可能，对推动中国化学品的管理有着重要的意义。

QSAR 预测模型主要分为基于统计分析方法的预测模型和基于机器学习算法的预测模型。例如：El Ghalia Hadaji[8]以多元线性回归构建 QSAR 预测模型；Afaf Zekri[3]以多元线性逐步回归构建 QSAR 预测模型；Lu Yang[9]基于遗传算法的多元线性回归构建 QSAR 预测模型；Svetnik Vladimir[10]以随机森林算法构建 QSAR 预测模型；代志军[11]以支持向量机回归构建 QSAR 预测模型；杨杰元[12]以 BP 神经网络算法构建 QSAR 预测模型；Li Jingshan[13]以梯度下降树决策树(GBDT)构建 QSAR 预测模型。虽然或多或少实现了预测，但是基于统计分析的方法随着变量急剧增多也变得束手无策。为了提高基于机器学习算法的预测精度，本文提出了基于改进的 PCA 和 ISSA-BPNN 的预测模型。

1　相关预测方法

1.1　BP 神经网络预测算法

BP 神经网络(BPNN)结构简单，使用方便，非循环多级网络训练算法，使其具有广泛的实用性[14]，能够实现输入到输出的非线性映射。BPNN 是单向传播的多层前向神经网络(结构如图 1 所示)，由输入层 x(m 个节点)、输出层 y(n 个节点)和多个隐含层组成。

1.2　SVR 预测算法

支持向量机回归(SVR)是将支持向量机分类(SVM)算法应用于回归预测中，两者不同的是：SVM 将间隔之内的空间样本算入损失函数中，以达到分类的目的；而 SVR 则是将间隔之外的空间样本算入损失函数中，以达到回归的目的。对于非线性 SVR 模型，使用核函数将数据映射到高维空间，而后进行回归预测。由于径向基核函数(RBF)应用广泛且具有较好的回归效果[15]，因此本文选择 RBF 作为 SVM 分析的核函数。

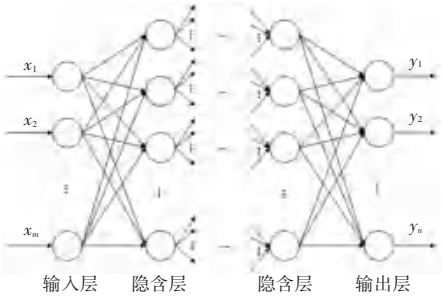


图 1　BP 神经网络结构图

Fig. 1　BP neural network structure

1.3　XGBoost 预测算法

XGBoost(Extreme Gradient Boosting)是在 Boosting 算法基础上进行改良的，在预测精度以及训练速度方面有较大的突破，属于 GBDT 的范畴[16]，并且也是一种前向特征的算法，本质上是由许多回归和分类的决策树组成[17]。XGBoost 相较于 GBDT 而言：前者加入正则项防止过拟合，对目标损失函数进行二阶泰勒展开，从而增加了精度，根据最佳切分点进行叶子节点分裂优化计算，从而优化结果。

2　QSAR 模型

本文实验数据集源自乳腺癌治疗靶标 ERα 时，得到的 1 974 个化合物作为 ERα 生物活性数据样本[18]。其中包括 729 个分子描述符信息和 pIC_{50}(实际 QSAR 建模中，一般采用 pIC_{50} 来表示生物活性值，即因变量)，pIC_{50} 值越大表明生物活性越高。

由于变量的数量比较多，本文首先提出基于改进的 PCA 特征选择算法，对模型的输入变量进行筛选，然后提出 ISSA-BPNN 算法对 BPNN 算法进行改进。

2.1　基于改进 PCA 的特征提取

改进的 PCA 算法流程如图 2 所示。首先对数据进行标准化，然后在 729 个分子描述符信息中，用基于 Pearson、MIC 和 RF 的加权得分算法得到前 20 个特征变量，最后基于 PCA 算法提取 4 个新特征代替原特征，作为模型的主要输入变量。

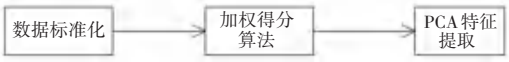


图 2 改进的 PCA 算法流程

Fig. 2 Improved PCA algorithm

2.1.1 最大互信息系数法(MIC)

MIC 是一种通过绘制变量散点图计算两个变量的互信息,来衡量变量间关联程度的算法[19]。其实现步骤如下:

(1)散点图网格化,计算互信息值。给定 n 个有序对数据集(x,y),将数据集划分为 $a \times b$ 的网格,x 方向和 y 方向的网格数分别为 a、b。互信息值的计算如式(1):

$$I(x;y)=\int p(x,y)\log_2\frac{p(x,y)}{p(x)p(y)}dxdy \quad (1)$$

式中,$p(x,y)$ 为 X 与 Y 之间的联合概率密度,$p(x)$ 和 $p(y)$ 分别为 X 和 Y 的边缘概率密度。

(2)互信息值归一化,如式(2):

$$I'(x;y)=\frac{I(x;y)}{\log_2\min(a,b)} \quad (2)$$

(3)变换网格划分情况,选择不同尺度下互信息的最大值作为 MIC 值,如式(3):

$$MIC(x;y)=\max_{a\times b<B(n)}\frac{I(x;y)}{\log_2\min(a,b)} \quad (3)$$

由文献[20]可知,当 $B(n)=n^{0.6}$ 时,效果最好。

2.1.2 随机森林(RF)特征选择法

随机森林(Random Forest, RF)[21]实质是一个包含多个决策树的组合分类器。其通过特征随机置换前后的误差分析,计算每个特征重要度得分,分值越高,特征越重要,从而进一步确定特征排序。随机森林结合 Bootstrap 重采样技术和决策树,构建一个包含多个基本分类器的树型分类器集合,采用简单多数投票的方法得到结果。

假设 RF 中决策树数目为 N_{tree},原始数据集有 d 个特征,单特征 $X^j(j=1,2,\cdots,d)$ 基于 OOB 误差分析的特征重要性度量,按以下步骤计算:

(1)计算第 i 棵决策树相应的袋外数据 OOB_i 的袋外错误样本数 $ErrOOB$;

(2)在保持其它特征不变的同时,对 OOB_i 中特征 X^j 进行随机序列改变得到 $\overline{OOB_i^j}$,重新计算袋外数据 $\overline{OOB_i^j}$ 的袋外错误样本数 $\overline{ErrOOB_i^j}$;

(3)重复步骤(1)、(2)得到:

$$\{ErrOOB_i^j \mid i=1,2,\cdots,N_{tree}\}$$

$$\{\overline{ErrOOB_i^j} \mid i=1,2,\cdots,N_{tree}\}$$

(4)由式(4)计算特征 X^j 的重要性得分。

$$VI(X^j)=\frac{1}{N_{tree}}\sum_i(\overline{ErrOOB_i^j}-ErrOOB_i^j) \quad (4)$$

2.1.3 基于 Pearson、MIC 和 RF 的加权得分算法

由于各变量的数值量纲之间存在较大差异,为了消除量纲的影响,需要对数据进行标准化处理[22]。本文采用 Z-score 标准化方法,对变量进行归一化处理,如式(5):

$$x^*=\frac{x-\mu}{\sigma} \quad (5)$$

Pearson 和 MIC 反映了自变量与因变量之间的线性和非线性关系,而 RF 是以特征重要度计算值来表示自变量与因变量的相关性。加权得分由式(6)计算得到:

$$grade_i=\alpha P_i+\beta MIC_i+(1-\alpha-\beta)\mathrm{RF} \quad (6)$$

其中,$grade_i$ 表示第 $i(i=1,2,3,\cdots,729)$ 个分子描述符的加权分;P_i 表示第 i 个自变量与因变量的 Pearson 系数绝对值;MIC_i 表示第 i 个自变量与因变量的最大互信息系数绝对值;RF_i 表示第 i 个自变量与因变量的特征重要度计算值,α 和 β 均应在 0 和 1 之间(本文取 $\alpha=\beta=0.25$)。

由式(6)计算得到 20 个主要特征变量见表 1。

表 1 加权得分分子描述符显著性排序

Tab. 1 Significance ranking of weighted score molecular descriptors

显著性排名	分子描述符	显著性排名	分子描述符
1	MDEC-23	11	BCUTc-1l
2	Lipoaffinity	12	MDEC-22
3	MLogP	13	AMR
4	minsOH	14	BCUTp-1h
5	maxHsOH	15	hmin
6	maxsOH	16	SaaCH
7	nC	17	SwHBa
8	minHsOH	18	SsOH
9	minsssN	19	maxsssNHp
10	C1SP2	20	SP-5

2.1.4 确定模型输入变量

PCA 算法的原理是以原始特征的线性组合方式,得到新特征来代替原特征,从而达到降维的效果[23]。根据方差越大新特征越重要的原则,对 p 个主成分按照贡献率进行排序,再从中提取 k 个主成分来代表全部数据,最后将新特征作为 QSAR 模型的输入值。算法流程如下:

(1)计算数据的协方差矩阵。假设原始数据集为 X,其协方差矩阵记为 $\boldsymbol{A}$;

（2）计算 A 的特征值 $\lambda_1,\cdots,\lambda_p$ 和对应的特征向量 $\xi_i=(\xi_{i1},\xi_{i2},\cdots,\xi_{ip})$；

（3）计算累计贡献率并确定主成分个数 k。

将表 1 中的 20 个特征变量由 PCA 算法特征提取后，得到新特征的贡献率见表 2。

表 2　新特征累计贡献率

Tab. 2　Cumulative variance contribution rate of new features

主成分	特征值	贡献率	累计贡献率
1	1 179.790 5	79.847 7	79.847 7
2	120.652 5	8.165 7	88.013 4
3	93.797 0	6.348 1	94.361 5
4	40.282 2	2.726 3	97.087 8
5	21.065 7	1.425 7	98.513 5
…	…	…	…
20	0.000 0	0.000 0	100

由表 2 可知，前 4 个新特征已包含原始特征 95%以上的信息。故本文取前 4 个新特征代替原特征作为模型的主要输入变量。

2.2　ISSA-BPNN

传统的 BPNN 对权值和阈值较敏感，存在收敛速度慢和极易陷入局部最优的问题[24]。因此，本文通过改进的麻雀搜索算法（improved sparrow search algorithm，ISSA）来优化 BPNN 的权值和阈值。

SSA 是根据麻雀觅食并逃避捕食者的行为而提出的群智能优化算法[25]，其模拟了麻雀群觅食的过程。在 SSA 中有 3 种状态，分别是发现者、加入者、侦察者。其中，适应度值较好的发现者是为了获得食物的同时，为所有加入者提供觅食的方向；侦察者选择安全第一为目标，在发现危险的情况下，提醒种群放弃食物。

由于 SSA 容易陷入局部最优，且全局搜索能力较弱，可将 SSA 中发现者和加入者位置更新公式分别改为式（7）、（8）。加入者以一定概率向发现者靠拢，保证了全局收敛[26]。同时，后加入的麻雀要尽快飞到其他区域觅食。

$$x_{ij}^{t+1}=\begin{cases}x_{ij}^{t}+x_{ij}^{t}\cdot randn(0,1), & R_2<ST\\ x_{ij}^{t}\cdot QD, & R_2\geqslant ST\end{cases}\tag{7}$$

$$x_{ij}^{t+1}=\begin{cases}Q\cdot\exp\left(\dfrac{i}{|x_{wj}^{t}-x_{ij}^{t}|+\varepsilon}\right), & i>\dfrac{N}{2}\\ x_{ij}^{t}+|x_{kj}^{t}-x_{ij}^{t}|\cdot FL\cdot randn(0,1), & other\end{cases}\tag{8}$$

其中，t 代表当前迭代次数；$randn(0,1)$ 和 Q 是服从标准正态分布的随机数；$\boldsymbol{D}$ 是 $1\times d$ 的矩阵，d 代表维度；x_{ij} 是第 i 个麻雀在第 j 维的位置；$R_2\in[0,1]$ 代表预警值；$ST\in[0.5,1]$ 代表安全值。

当 $R_2\geqslant ST$ 时，表示发现者已经发现捕食者，此时种群内其它麻雀尽可能飞到其它安全地方进行觅食；当 $R_2<ST$ 时，发现者可以广泛搜索。N 是种群规模，$x_{\omega j}$ 是当前全局最差的位置，x_{kj} 是当前发现者的位置，$FL\in[-1,1]$ 表示加入者跟随生产者寻找食物的概率。当 $i>N/2$ 时，表示适应度值较差的第 i 个加入者处于挨饿状态，需要尽快飞到其它区域继续寻找食物来获得能量。

侦察者的位置更新如式（9）：

$$x_{ij}^{t+1}=\begin{cases}x_{ij}^{t}+\beta\cdot|x_{ij}^{t}-x_{bj}^{t}|, & f_i\neq f_g\\ x_{ij}^{t}+K\cdot\dfrac{|x_{ij}^{t}-x_{\omega j}^{t}|}{(f_i-f_\omega)+\varepsilon}, & f_i=f_g\end{cases}\tag{9}$$

式中，K 是 $[-1,1]$ 范围内的一个随机数；β 是步长控制参数，其服从标准正态分布的随机数；x_{bj} 表示当前的全局最佳位置；f_ω、f_g 和 f_i 分别代表当前麻雀的全局最差、全局最优和个体适应度。分母加上一个常数量 ε，是为了防止分母出现 0 的情况。

本文提出的 ISSA-BPNN 流程如图 3 所示，其实现步骤为：

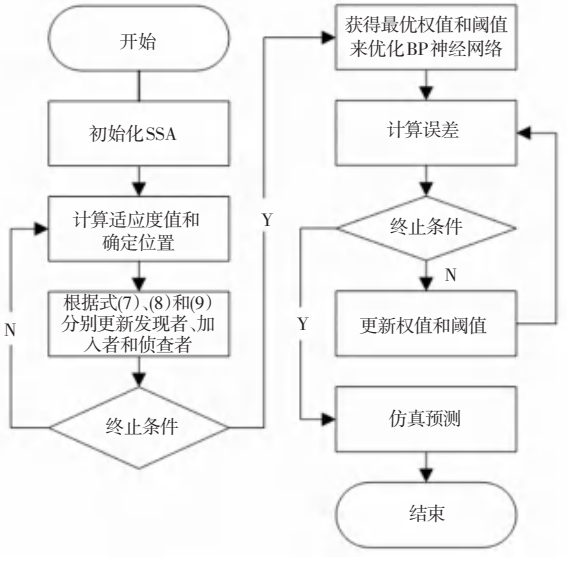


图 3　ISSA-BPNN 流程

Fig. 3　ISSA-BPNN flow chart

（1）初始化麻雀搜索算法；

（2）计算麻雀种群个体适应度，并得到最佳位置、最差位置和最佳适应度值、最差适应度值；

（3）根据式（7）~（9）分别更新发现者、加入者和侦查者的位置信息，并更新适应度值；

（4）若算法达到最大迭代次数或达到最初设定的收敛精度，则执行步骤（5），否则返回步骤（2）；

(5)将得到的最优值赋给 BPNN 的权值和阈值;

(6)使用 BPNN 进行学习,不断调整直至达到训练终止条件,最终实现预测输出。

3 实验结果与分析

依据上述方法对数据进行新特征选取后,将 1 974个样本按照 7:3 的比例划分训练集和测试集。训练集用来拟合模型,测试集用来对模型的性能进行评价。验证本文所提出模型的有效性,分别利用 SVR、XGBoost、BPNN 和 ISSA-BPNN 模型对前述数据集合进行预测。

3.1 预测模型的评价指标

本文采用平均绝对误差(*MAE*)、平均绝对百分比误差(*MAPE*)和均方根误差(*RMSE*)评价模型的预测精度。其计算公式分别为式(10) ~ (12):

$$MAE = \frac{1}{n}\sum_{i=1}^{n} | \hat{y}_i - y_i | \tag{10}$$

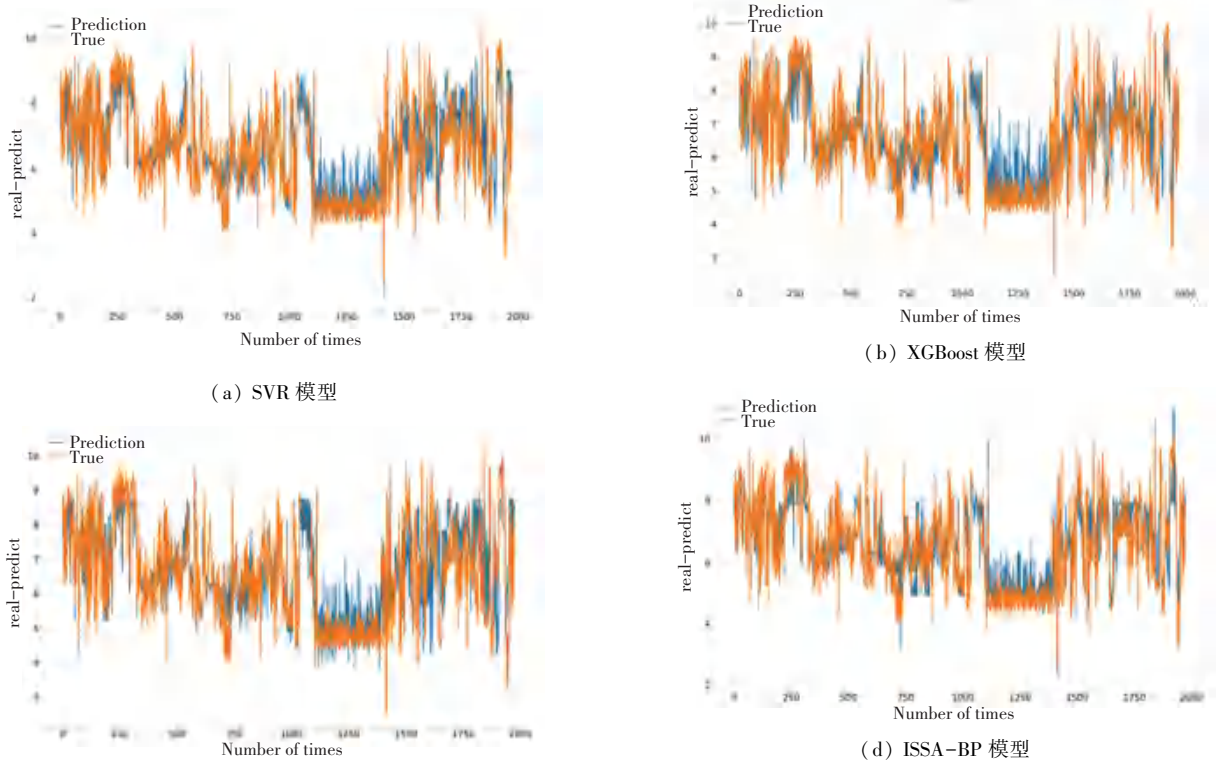
$$MAPE = \frac{1}{n}\sum_{i=1}^{n} | \hat{y}_i - y_i | \tag{11}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n}} \tag{12}$$

其中,y_i 表示测试集的真实值;n 为样本个数;$\hat{y}_i$ 表示模型预测值。

3.2 结果分析

4 种模型的 pIC_{50} 预测值与真实值曲线对比如图 4 所示,预测精度对比结果见表 3。



(a) SVR 模型 (b) XGBoost 模型

(c) BP 模型 (d) ISSA-BP 模型

图 4 4 种模型的 pIC_{50} 预测值与真实值对比

Fig. 4 Comparison of predicted pIC_{50} values and true values of four models

表 3 4 种模型预测精度对比

Tab. 3 Comparison of prediction accuracy of four models

模型	*MAE*	*MAPE*	*RMSE*
SVR	0.731 7	0.117 6	0.997 4
XGBoost	0.712 6	0.114 2	0.953 1
BPNN	0.847 4	0.128 1	0.968 6
ISSA-BPNN	0.655 6	0.102 2	0.808 8

由表 3 可知,ISSA-BPNN 模型的 *MAE*、*MAPE*、*RMSE* 均是最低的,表明 ISSA-BPNN 预测误差值最小、稳定性最高、效果最佳。其中,ISSA-BPNN 模型的 *MAPE* 值较 SVR 模型提高了 13.10%,较 XGBoost 模型提高了 10.53%,较 BPNN 模型提高了 20.22%。

4　结束语

为了更精确地预测化合物的生物活性，本文提出了一种基于改进的 PCA 和改进的麻雀搜索算法优化 BP 神经网络(ISSA-BPNN)预测模型，其具有良好的寻优能力。

算法中，利用改进的 PCA 算法提取模型的主要变量，再利用 ISSA 优化 BPNN 的权值和阈值，改善了 BPNN 易陷入局部极值的缺点。通过实验对比结果表明，基于 ISSA-BPNN 预测模型的预测精度最高，并具有较强的拟合能力和泛化能力。但是，由于训练的数据量较少，导致模型的预测精度不是太高，后期研究可增加训练数据来提高模型的预测精度。

参考文献

[1] 张雨迪. 小分子化合物 VP384 抗乳腺癌作用研究[D]. 武汉：武汉大学，2020.

[2] YAGER J D, DAVIDSON N E. Estrogen carcinogenesis in breast cancer[J]. New England Journal of Medicine, 2006, 354(3): 270-282.

[3] ZEKRI A, HARKATI D, KENOUCHE S, et al. QSAR modeling, docking, ADME and reactivity of indazole derivatives as antagonizes of estrogen receptor alpha (ER-α) positive in breast cancer[J]. Journal of Molecular Structure, 2020, 1217: 128442.

[4] MIYOSHI Y, MURASE K, SAITO M, et al. Mechanisms of estrogen receptor-α upregulation in breast cancers[J]. Medical molecular morphology, 2010, 43(4): 193-196.

[5] TECALCO - CRUZ A C, RAMÍREZ - JARQUÍN J O. Polyubiquitination inhibition of estrogen receptor alpha and its implications in breast cancer[J]. World journal of clinical oncology, 2018, 9(4): 60.

[6] KOVACEVIĆ S Z, PODUNAVAC-KUZMANOVIĆ S O, JEVRIĆ L R, et al. Preselection of A-and B-modified D-homo lactone and D-seco androstane derivatives as potent compounds with antiproliferative activity against breast and prostate cancer cells-QSAR approach and molecular docking analysis[J]. European Journal of Pharmaceutical Sciences, 2016, 93: 107-113.

[7] 王中钰，陈景文，傅志强，等. QSAR 模型应用域的表征方法[J]. 科学通报，2022，67(3)：12.

[8] HADAJI E G, BOURASS M, OUAMMOU A, et al. Organic Compounds Based on (E)-N-Aryl-2-ethene-sulfonamide as Microtubule Targeted Agents in Prostate Cancer: QSAR Study[J]. Advances in Physical Chemistry, 2017, 2017: 7629056.

[9] YANG L, SANG C, WANG Y, et al. Development of QSAR models for evaluating pesticide toxicity against Skeletonema costatum[J]. Chemosphere, 2021, 285: 131456.

[10] SVETNIK V, LIAW A, TONG C, et al. Random forest: a classification and regression tool for compound classification and QSAR modeling[J]. Journal of chemical information and computer sciences, 2003, 43(6): 1947-1958.

[11] 代志军. 特征选择与样本选择用于癌分类与药物构效关系研究[D]. 长沙：湖南农业大学，2014.

[12] 杨杰元，杨雪颖，杨沛艳，等. 基于人工神经网络研究芳胺喹唑啉衍生物的抗胃癌活性[J]. 化学通报，2021，84(8)：853-856，846.

[13] LI Jingshan, LUO Dehan, WEN Tengteng, et al. Representative feature selection of molecular descriptors in QSAR modeling[J]. Journal of Molecular Structure, 2021, 1244: 131249.

[14] 陈永当，曹坤煜. 基于 IIGA-BP 神经网络的钢材销售预测模型[J]. 计算机系统应用，2021，30(10)：138-147.

[15] 赵敬川，赵吉宾，李论，等. 基于支持向量机的复杂曲面磨削去除量预测[J]. 组合机床与自动化加工技术，2021(11)：58-61.

[16] XU Yonghui, ZHAO Xi, CHEN Yinsheng, et al. Research on a Mixed Gas Classification Algorithm Based on Extreme Random Tree[J]. Applied Sciences, 2019, 9(9).

[17] 陈延展，胡浩，任紫畅，等. 基于 XGBoost 和改进灰狼优化算法的催化裂化汽油精制装置的辛烷值损失模型分析[J]. 石油学报（石油加工），2022，38(1)：208.

[18] 中国学位与研究生教育学会“华为杯”第十八届中国研究生数学建模竞赛赛题 D 题 https://cpipc.acge.org.cn//cw/detail/4/2c9080147c73b890017c7779e57e07d2 2021.[2021-12-02].

[19] 甄成刚，张争鹏. 基于 VMD 分解与 MIC 特征分析的风电功率组合预测[J]. 郑州大学学报(理学版)：2022(3)：88-94.

[20] BATTITI R. Using mutual information for selecting features in supervised neural net learning[J]. IEEE Transaction on neural networks, 1994, 5(4): 537-550.

[21] BREIMAN L. Random forests[J]. Machine learning, 2001, 45(1): 5-32.

[22] 岳彪，闵永智，马宏锋，等. 钢轨表面缺陷检测系统中图像增强预处理方法研究[J]. 道科学与工程学报，2018，15(12)：3248-3256.

[23] 陈晋市，张淼淼，王普长，等. 基于 PCA 最优阈值选取的挖掘机主泵载荷谱外推研究[J]. 吉林大学学报(工学版)：1-8[2021-11-27].

[24] 张艺铭，陈明明，石磊，等. 基于 IGWO-BP 算法的轨道交通短时客流预测. 交通信息与安全，2021，39(3)：85-92.

[25] 吕鑫，慕晓冬，张钧. 基于改进麻雀搜索算法的多阈值图像分割[J]. 系统工程与电子技术，2021，43(2)：318-327.

[26] 许亮，张紫叶，陈曦，等. 基于改进麻雀搜索算法优化 BP 神经网络的气动光学成像偏移预测[J]. 电子·激光，2021，32(6)：653-658.

(上接第 83 页)

[7] 王邳翔，王余宽，许小卫，等. 基于随机森林算法的水陆联运枢纽客运量预测[J]. 水运管理，2021，43(9)：12-15.

[8] 彭丽洁，邵喜高，黄万明. 基于灰色马尔可夫模型的烟台市铁路客运量预测研究[J]. 鲁东大学学报(自然科学版)，2022，38(1)：50-56.

[9] 韦师，苏玉华. 基于组合预测模型的我国铁路客运量发展趋势分析[J]. 中国市场，2021(18)：162-164.

[10] 王国贤，范英兵，王凤玲，等. 时间序列和神经网络下我国铁路客运量的预测研究[J]. 黑河学院学报，2021，12(5)：182-185.