计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200700066

基于融合神经网络模型的药物分子性质预测

谢良旭[1,2]　李　峰[3]　谢建平[4]　许晓军[1]
1 江苏理工学院电气信息工程学院生物信息与医药工程研究所　江苏 常州 213001
2 江苏省中以产业技术研究院　江苏 常州 213100
3 江苏理工学院电气信息工程学院　江苏 常州 213001
4 湖州师范学院理学院　浙江 湖州 313000
(xieliangxu@jsut.edu.cn)

摘　要　在生物信息学领域,人工智能方法在预测药物分子的物理化学性质和生物活性中获得了重大成功,特别是神经网络已被广泛应用到药物研发中。但是浅层神经网络的预测精度低,深度神经网络又容易出现过拟合的问题,而模型融合策略有望提升机器学习中弱学习器的预测能力。据此,文中将模型融合方法首次应用到药物分子性质的预测中,通过对药物分子的化学结构进行信息化编码,采用平均法、堆叠法融合浅层神经网络,提高对药物分子 pKa 预测的能力。与深度学习方法相比,堆叠法(Stacking)融合的模型具有更高的预测准确性,其预测结果的相关系数达到 0.86。通过将多个弱学习器的神经网络有机组合可使其达到深度神经网络的预测精度,同时保留更好的模型泛化能力。研究结果表明,模型融合方法可提高神经网络对药物分子 pKa 预测结果的准确性和可靠性。

关键词:计算机辅助药物设计;生物信息学;模型融合;深度学习;机器学习
中图法分类号　TP183

Predicting Drug Molecular Properties Based on Ensembling Neural Networks Models

XIE Liang-xu[1,2], LI Feng[3], XIE Jian-ping[4] and XU Xiao-jun[1]
1 Institute of Bioinformatics and Medical Engineering, School of Electrical and Information Engineering, Jiangsu University of Technology, Changzhou, Jiangsu 213001, China
2 Jiangsu Sino-Israel Industrial Technology Research Institute, Changzhou, Jiangsu 213100, China
3 School of Electrical and Information Engineering, Jiangsu University of Technology, Changzhou, Jiangsu 213001, China
4 School of Science, Huzhou University, Huzhou, Zhejiang 313000, China

Abstract　Artificial intelligence (AI) methods have made great success in predicting chemical properties and bioactivity of drug molecules in the Bioinformatics field. Neural network gains wide applications in the process of drug discovery. However, the shallow neural network (SNN) gives lower accuracy while deep neural networks (DNN) are easy to be overfitting. Model ensembling is expected to further improve the predictive performance of weak learners in traditional machine learning methods. Therefore, it is the first time to apply model ensembling strategy to predict the properties of drug molecules. By encoding molecular structures, the combination strategies, averaging, and stacking methods are adopted to increase predicting accuracy of pKa of drug molecules. Compared with DNN, the stacking strategy presents the best predictive accuracy and the Pearson coefficient reaches to 0.86. Ensembling weak learners of the neural networks can reproduce the accuracy of DNN while keeping the satisfied generalization ability. The results show that ensembling method can increase the predictive accuracy and reliability.

Keywords　Computer aided drug discovery, Bioinformatics, Model ensembling, Deep learning, Machine learning

1　引言

药物分子的酸性解离系数 pKa 反映了分子的离子化状态,直接关系到药物分子的溶解性、在生物体内的穿膜性、油水之间的分配比例等。因此,pKa 作为药物分子的重要物理化学性质,也是评价其生物活性的重要指标,在药物发现和设计领域经常被用来衡量药物分子在体内的吸收、分布、代谢、毒性(ADMET)等药物代谢的重要性质。通过生物实验测定

到稿日期:2020-07-10　返修日期:2020-10-20　本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家自然科学基金(12074151,22003020);江苏省自然科学基金(BK20191032);常州市重点研发项目(CJ20200045);江苏省中以产业技术研究院开放课题(JSIITRI202009)
This work was supported by the National Natural Science Foundation of China (12074151, 22003020), Natural Science Foundation of Jiangsu Province, China (BK20191032), Changzhou Sci & Tech Program (CJ20200045) and Funding from Jiangsu Sino-Israel Industrial Technology Research Institute (JSIITRI202009).
通信作者:许晓军(xuxiaojun@jsut.edu.cn)

分子的 pKa 既费时又费力。为此,在药物研发过程中准确预测药物的 pKa 可以有效地降低药物研发的风险,控制药物研发成本。对于已知药物分子的结构,借助高性能计算机进行基于分子结构的药物分子性质预测,是近几十年来生物信息学领域研究的热点。目前,已经发展了多种数学模型计算药物分子的 pKa,其中定量构效关系(QSAR)是 20 世纪 60 年代提出的,是人类在药物发现领域使用最早的合理药物设计方法[1]。在计算机技术快速发展之前,QSAR 是使用最为广泛的药物设计方法[2]。

人工智能方法被认为是 21 世纪的三大尖端技术之一[3-5],近几年来,在药物设计领域获得了巨大的成功[6]。通过对药物分子编码,利用人工智能方法助力药物发现成为当前研究的热点[7]。诺华、辉瑞等国际药企纷纷加入到人工智能辅助药物发现的研究中。人工智能方法被用来提升传统虚拟药物筛选的打分函数的精度[8]、生成可能的药物分子[9],以及计算靶向药物和小分子结合自由能[10]。随着计算机性能的提升和计算方法的改进,深度神经网络在近几年获得了巨大的发展[11]。利用深度神经网络等深度学习方法预测药物分子的理化性质获得了广泛的关注。深度神经网络的应用显著提升了对分子 ADMET 性质的预测性能[12-13]。国内外学者也借助多种人工智能方法预测分子的性质,如使用传统神经网络、随机森林、支持向量机等方法预测药物分子的重要性质。Hou 课题组评测了多个机器学习方法在药物分子性质预测中的性能,利用机器学习方法预测了 logD 等重要的药物分子性质[14-15]。目前,人们已经可以利用经验公式或者机器学习方法计算药物分子的 pKa,并且已有较为成熟的计算软件,如 ACD/Labs、ChemAxon 和薛定谔。Liao 等对 9 种软件进行了对比,发现现有软件在计算未见过的新分子时的准确性仍有待提高[16]。另外,预测 pKa 的软件都是商业软件,目前还未有一款开源的软件可以预测分子的 pKa 等性质。Mansouri 等也指出,如何借助已知药物分子的数据,通过开源的方法实现对分子 pKa 的预测是亟待解决的问题[17]。而机器学习具有广泛的应用性和较高的可移植性,通过机器学习预测药物分子性质是目前最具有潜力的解决方案。

利用机器学习预测药物分子的准确性依赖于有效的机器学习方法和充足的测试数据集。传统的机器学习方法采用较少的参数,具有较高的计算效率,然而对数据的拟合效果难以达到深度学习算法的精度。而深度学习方法虽然在训练过程中通常能取得较好的预测结果,但在测试集中却存在着泛化效果不佳的问题。深度神经网络计算的准确性受到模型初始条件的影响,如初始随机权重、训练过程中的统计噪声等,即使经过大量耗时费力的训练,仍不能保证深度神经网络模型的预测效果的准确性。因此,单一机器学习方法或者深度学习方法在实际应用中的预测效果并不理想。为结合机器学习方法和深度学习方法的优势来提高传统机器学习方法的预测性能,模型融合的策略应运而生。由于神经网络算法的随机性,每次对神经网络进行训练时可得到一个稍微不同的映射函数。在同一数据集上训练多个神经网络,将多个神经网络的预测结果组合在一起,可抵消单个神经网络模型的误差。因此模型融合方法可避免训练方案、模型参数选择和单次偶然结果等因素的影响,提高预测结果的准确性。通过组合多个弱学习器可以达到更好的预测结果。Tang 等提出了集成神经网络,通过对多个网络赋予权重,提升了模型的泛化能力[18]。Fu 等利用平均法集成了机器学习方法,并将该集成方法用于分子的性质预测中[15]。集成学习采用的一般策略是先分别产生多个弱学习器,再采用合理的方法将它们组合起来,通过"博采众长"的策略使其达到强学习器的预测能力。可采用的模型融合策略主要有平均法、投票法和堆叠法(Stacking)等。不同的模型融合策略各有优劣,然而目前对于组合策略的性能也未有相应的评测,在实际应用中该采取何种模型融合方法还未有定论。虽然神经网络已获得了广泛的应用,但是由于学科交叉存在难度,在生物信息领域的研究仍使用的是单一的神经网络学习器,模型融合方法还未能应用到生物信息学的研究中[19]。

将融合神经网络模型应用到生物信息学的研究中,将有望提升原有计算方案的准确性,具有重要的应用前景。本工作首次将模型融合策略应用到药物分子 pKa 的预测中,通过组合方法将多个浅层神经网络模型整合,并与深度神经网络方法进行对比,探讨该策略在预测药物分子 pKa 中的表现,定量表征组合策略相比浅层神经的性能提升。

2 材料和方法

2.1 数据集整理

首先构建一个包含药物分子 pKa 的本地数据库,选取药物分子数据库 DrugBank 进行数据整理与挖掘。DrugBank 数据库是目前对药物数据收集比较全面的数据库,其中收集了 11 895 种可能的药物分子,包含 1 184 种已批准的药物分子[20]。数据库中的药物分子性质以 SDF 文件格式保存,利用脚本语言从数据库中抽提出药物分子的结构和每个分子所对应的 pKa 数值。通过数据整理发现有 8656 个药物分子包含 pKa 等数值。将输入数据按照 8∶2 的比例分为训练集和测试集。训练集中的数据用于超参数的优化,测试集中的分子是神经网络训练过程中未见过的分子,用于表征所训练的神经网络的泛化性能。

2.2 分子描述符

通过人工智能技术预测药物分子性质所面临的关键问题是如何将药物分子的分子结构转变为机器学习和深度学习可以直接识读和处理的文件格式。在化学信息学和生物信息学研究领域,一般采用分子描述符将分子结构编码为有用的数字化信息。Keiser 等强调了有效的表征分子将直接影响机器学习算法的准确性[21]。目前应用较广泛的分子描述符是 MACCS 密钥[22],MACCS 密钥通过检索分子中是否存在字典中规定的子结构,将整个分子转变为二进制的化学信息。采用 RDKit 软件将所选取的分子结构进行编码。MACCS 密钥由 166 个描述符组成,每个描述符采用 0 或 1 来表示分子中是否包含相应的原子种类、成键信息、原子周围的环境等。MACCS 密钥中不包含冗余的信息,在之前的药物定量构效关系和机器学习中获得了广泛的应用,在不进行特征工程的情况下,成功地用于药物分子相似性寻找、药物构效关系预测、对蛋白结合口袋的编码[23]等。因此本文选取 MACCS 密钥对药物分子进行编码,并将其应用于药物分子的 pKa 的预测中。MACCS 密钥具有独一性,每个分子可编码为独特的数字串,因此 MACCS 密钥也被称为 MACCS 指纹。如图 1 所示,以 SIRT1 的抑制剂分子为例,利用 RDKit 将该分子结构转变为神经网络算法易于处理的二进制数。

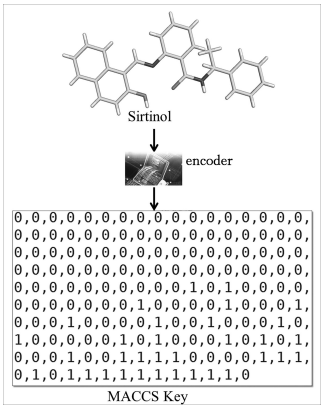
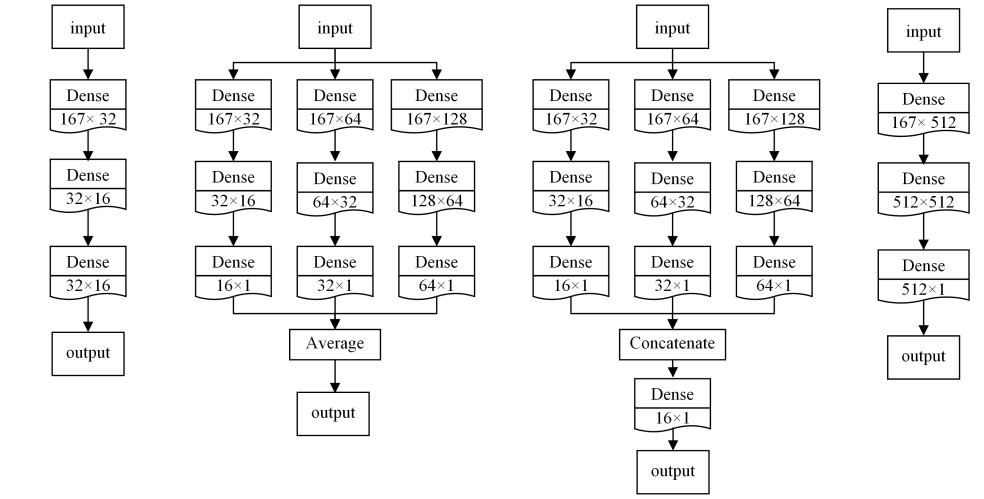


图 1 将药物分子的结构转变为 MACCS 密钥的示意图

Fig. 1 Schematic figure of converting drug molecules to MACCS keys

2.3 神经网络

神经网络具有广泛的应用性，尤其是随着计算机软硬件的提升，神经网络进入了高速发展的时期。针对本次研究的数据预测和回归问题，本文通过神经网络预测药物分子的 pKa。为验证组合计算策略，选取浅层神经网络作为模型融合所需要的个体学习器。神经网络包含输入层、隐藏层和输出层。输入层的神经元数目与输入数据的维度一致。如图 2 所示，本文采用 3 个浅层神经网络作为独立的弱学习器，通过组合策略将 3 个浅层神经网络整合。浅层神经网络的输入层的输出维度分别设置为 32，64，128，隐藏层的神经元数目分别为 16，32，64。深度学习器的隐藏层的神经元数目设置为 512。神经网络的其他参数通过超参数优化步骤获得。



(a)浅层神经网络 (b)平均法融合的神经网络 (c)堆叠法融合的神经网络 (d)深度神经网络

图 2 所评测的模型

Fig. 2 Evaluated models

2.4 超参数优化

对神经网络的调参采用 GridSearchCV 方法结合 K 折(K-fold)验证方法。在 K 折交叉验证过程中选取 20% 的数据作为验证集，验证所选出的超参数是否合理。超参数的优化范围设置如下：对优化器分别采用 RMSprop，Adagrad，Adadelta，Adam，Adamax，Nadam。对激活函数的设置分别为 softmax，softplus，softsign，relu，tanh，sigmoid，hard_sigmoid，linear。在训练过程中，批量尺寸设置为 10，20，40，60，80，100。在隐藏层和输出层之间添加 dropout 层，随机地将该层的一些输出特征舍弃，筛选的 dropout 的比例范围为 0～0.6，间隔为 0.1。

2.5 组合策略

组合策略的一般结构是先产生一组个体学习器，个体学习器通常是由一个比较弱的机器学习算法开始进行初步计算，通过个体学习器分别对学习任务进行学习，然后采用特定的策略将它们结合起来。个体学习器的相关性越低，预测结果就越好。通过组合多个学习器，可以获得比单一的学习器更好的泛化性能。目前可使用的组合策略包括平均法、投票法和堆叠法。其中最直接的方法为平均法，如图 2 所示。在平均法的计算过程中，首先通过几个个体学习模型预测分子的性质，然后对预测结果直接平均得到预测数值，其公式为：

$$H(x)=\frac{1}{T}\sum_{i=1}^{T}h_i(x) \tag{1}$$

其中，T 表示基学习器的个数，$h(x)$ 表示基学习器的输出，平均法直接对弱学习器做平均。

Stacking 模型融合方法分两步，首先从初始数据集训练出初级学习器，然后把初级学习器的输出组合成新的数据集，用于训练次级学习器。其与平均法的不同之处在于额外使用一层神经网络进行数据的学习。在图 2(c)中，初级学习器的输出拼接为一个输入的数据集，然后输入到后续神经网络的紧密层中。通过不同学习器的互相补充可使神经网络获得更好的预测性能。在实际应用中可使用不同的学习模型的组合，为方便对比，选取浅层神经网络和深度神经网络分别作为弱学习器和强学习器，模型融合方法直接对不同的浅层神经网络进行组合。本文对组合的浅层神经网络和深度神经网络分别进行对比。

3 结果与讨论

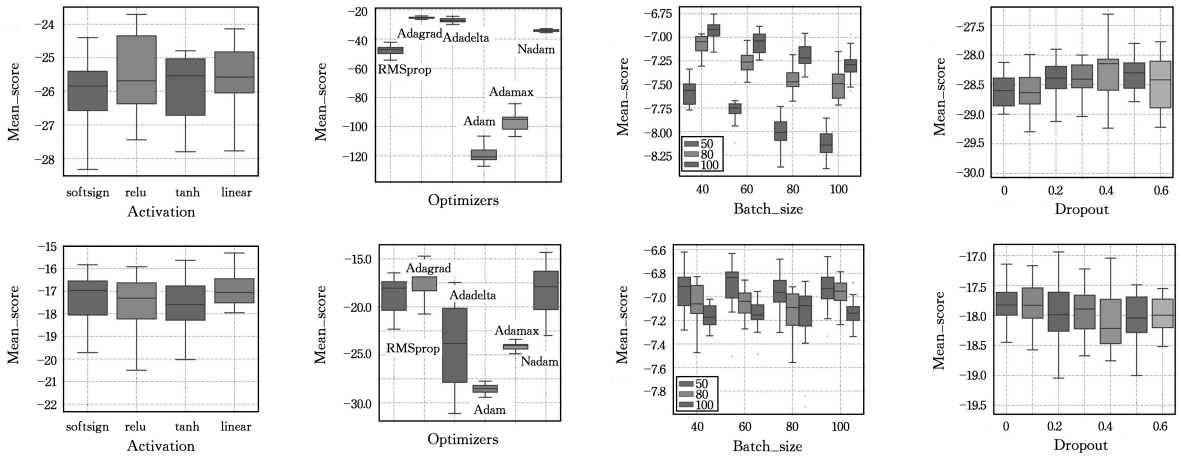
3.1 实验环境

实验使用 Ubuntu 7.4.0 的操作系统，内存为 32GB，采用 48 核的分布式计算机。计算环境需安装生物信息学软件 RDKit-2017，深度学习库 Keras-2.2.2 和 Tensorflow-1.3.0，

数据的分析和存储使用 Scikit-learn-0.19.2 和 Numpy-1.18.1。使用 Python-3.5.6 编码进行药物分子的编码、模型训练和数据分析。

3.2 超参数优化

为得到更好的预测结果，首先对神经网络的超参数进行优化，在通过 GridSearchCV 优化超参数的过程中发现，单次优化过程中存在随机误差。因此，本文通过 20 次重复计算，利用统计分析获得最可能的超参数，分别对优化器、激活函数、批尺寸等进行了优化。如图 3 所示，激活函数 linear 和优化器 Adagrad 在浅层和深度神经网络中都表现较好。在浅层神网络中，批尺寸为 40，*Epoch* 为 100 时，有较低的损失。在深层神经网络中，批尺寸为 60，*Epoch* 为 50 时，有较低的损失。dropout 的比例选择 0.5。因此在浅层神经网络和深度神经网络中，选取以上超参数组合。



(a)对浅层神经网络优化的超参数和对应的打分 (b)对深度神经网络优化的参数和对应的打分

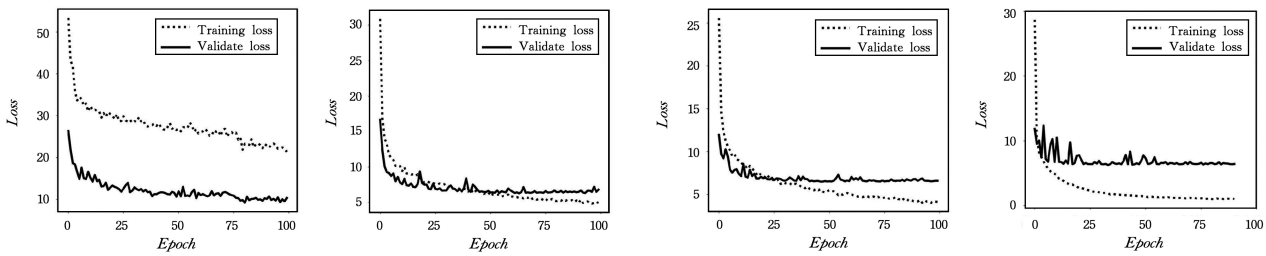
图 3 超参数优化的结果

Fig. 3 Results of hyperparameter optimization

3.3 组合神经网络与深度神经网络的泛化能力对比

利用挑选出的超参数组合，分别测试平均法与 Stacking 法组合的神经网络的性能，并选择单个浅层神经网络和深度神经网络作为对比。通过计算均方差比较 4 种方法在训练集和测试集上的表现。

首先对比 4 种方法在拟合过程中在训练集和验证集上的损失函数。通过对比模型在训练集和测试集中的表现，选出具有较好泛化能力的模型。如图 4 所示，浅层神经网络在训练集和验证集上都有较大的损失，未得到理想的预测结果。各个模型的损失结果如表 1 所列，可以发现深度神经网络的表现最好，在对训练集和测试集预测时，预测药物分子的 pKa 的损失分别为 1.04 和 6.32，但是深度神经网络容易出现过拟合，即在训练集和验证集上的损失有明显的偏差(损失差别为 5.38)。Stacking 法的融合神经网络虽然在训练集中的损失为 4.12，但是在测试集中的损失为 6.60，接近深度神经网络的 6.42。平均法和 Stacking 方法融合的模型达到了深度神经网络的预测性能，并且二者都表现出了更好的泛化能力。



(a)SNN (b)Averaging (c)Stacking (d)DNN

图 4 不同神经网络在训练集和验证集上的损失

Fig. 4 Loss of training and validation subset of each evaluated model

表 1 不同机器学习方法的 Pearson 系数和损失结果

Table 1 Pearson coefficients and the loss for different evaluated methods

Methods	Train		Test	
	P	*Loss*	*P*	*Loss*
SNN	0.82	21.95	0.79	10.43
Averaging	0.92	7.93	0.84	6.80
Stacking	0.94	4.12	0.85	6.60
DNN	0.97	1.04	0.85	6.42

为进一步对比泛化能力，对模型在训练集和测试集上的结果计算 Pearson 系数。Pearson 系数表征了模型的预测值与真实值之间的符合程度。

如表 1 所列，DNN 在训练集中的 Pearson 系数达到了 0.97，而在测试集中的 Pearson 系数为 0.85。平均法和 Stacking 法融合的神经网络在测试集上的 Pearson 系数为0.85，达到了较好的预测性能。结果表明，平均法和 Stacking 法的组合神经网络的预测性能可以达到深度神经网络的预测性能，其中 Stacking 法获得了最好的 Pearson 系数和最小的损

失。因此采用 Stacking 方法对模型进行融合，可有效降低过拟合，并提升模型的泛化能力。

3.4 组合神经网络与深度神经网络的计算精度对比

为进一步评测上述 4 种方法的计算精度，本文计算了预测 pKa 数值与数据库中存储数值的均方根误差 *RMSE* 及其在不同误差区间的分布。从表 2 中的 *RMSE* 数值可以看到，在训练集中，深度神经网络方法的 *RMSE* 最小，为 1.23。然而，在测试集中，Stacking 融合的神经网络的 *RMSE* 最小，为 2.32。Stacking 融合的神经网络在未训练过的子集上的效果甚至优于深度神经网络，这进一步说明模型泛化能力的重要性。

表 2 不同模型的均方根误差和分布

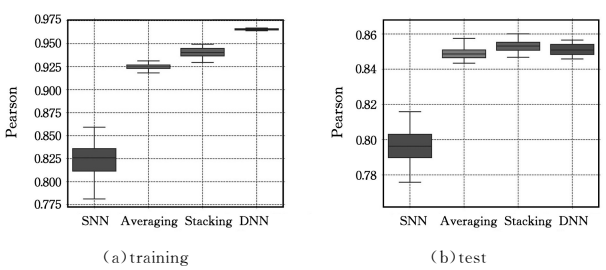
Table 2 Root mean square error and classification distribution for validated methods

Methods	Training				test			
	Population of molecules within error range/%				Population of molecules within error range%			
	RMSE	<0.5	0.5～1.0	>1.0	*RMSE*	<0.5	0.5～1.0	>1.0
SNN	3.03	13.3	11.4	75.2	3.06	12.5	10.6	76.9
Averaging	1.81	27.8	23.5	48.7	2.39	20.3	19.3	60.4
Stacking	1.51	36.0	27.4	36.6	2.32	24.2	20.3	55.5
DNN	1.23	68.0	16.8	15.2	2.42	27.2	17.7	55.1

对预测值和真实值之间的偏差大小进行分层统计，另外表 2 也统计了预测的 pKa 处于不同误差范围内的药物分子数目占总体药物分子数目的百分比，平均法、Stacking 法和深度神经网络的结果显著优于浅层神经网络。预测值与真实值之间的偏差小于 0.5 的范围内，深度神经网络在该区间内的药物分子数目百分比最高，为 27.2%；在 0.5～1.0 区间内，Stacking 融合的神经网络在该区间内的药物分子数目的百分比最高，为 20.3%。统计分布的结果表明，组合计算策略可以有效提升弱学习器的计算准确性，降低预测的误差。

3.5 组合神经网络与深度神经网络的可靠性比较

为进一步评测结果的稳定性和可靠性，对模型重复进行 20 次计算。以模型计算出的 Pearson 相关系数为指标，分析模型在不同计算过程中的稳定性。在多次计算中的 Pearson 系数分布能更好地反映模型整体预测结果的稳定性。如图 5 所示，在多次计算中，Stacking 方法使浅层神经网络的预测准确性得到有效提升。构建的组合计算方法在测试集中具有更好的可靠性。其中，Stacking 法给出了最窄的数据分布和相对较高的 Pearson 系数，甚至略优于深度神经网络的结果。



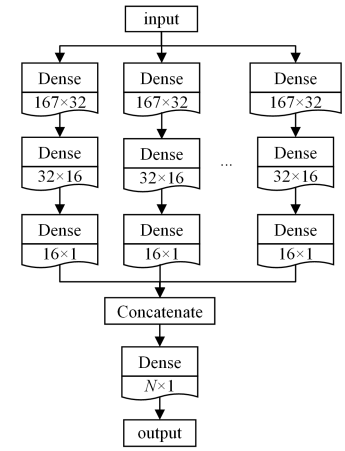
(a)training (b)test

图 5 Pearson 系数在训练集和测试集中的分布

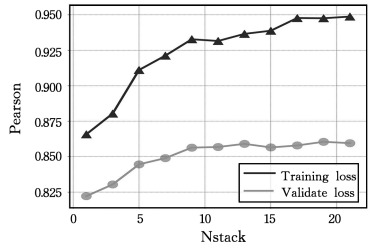
Fig. 5 Pearson coefficients distribution on training and test subsets

平均法是对几个弱学习器直接求平均值，因此其预测结果的准确性受限于预测效果最差的个体学习器，单个弱学习器的预测性能直接影响到组合后的预测结果。堆叠法对弱学习器的初级输出数据进行再学习，评估初级学习器中结果的准确率，并根据结果的准确率为不同的初级学习器中的结果赋予不同的权重。由于堆叠法使用了不同的组合方式，其对各个学习器预测结果的准确性没有特定要求。在堆叠法中使用多个学习器是为了能够涵盖更多的差异化组合。

融合神经网络模型通过组合多学习器，其预测的性能甚至可以超过深度神经网络。为验证该猜测，对浅层神经网络进行多个堆叠。如图 6 所示，可以看出当有 9 个并列的浅层神经网络时，其预测性能达到 0.86 的平台峰值。组合模型获得了更好的预测精度，同时保留了更好的泛化性能。可能的原因是模型融合中使用了多个有差异的学习器，不同学习器之间实现了模型的互补，在合理范围内增加个体学习器的数目，有助于增加网络的复杂性，从而得到更好的预测结果。而深度神经网络虽然能在输入与输出之间建立起更多的通路，但在处理复杂的数据时，其泛化能力可能达不到组合模型中多学习器的泛化能力。



(a)本文使用的堆叠神经网络的结构



(b)堆叠法中预测的系数与堆叠网络个数的关系

图 6 堆叠法所使用网络个数与预测系数的关系

Fig. 6 Relationship between the number of stacked neural network and the predicted Pearson coefficients

结束语 针对浅层神经网络在预测药物分子 pKa 时精度不高以及深度神经网络泛化能力不佳的问题，本文首次在药物分子 pKa 预测中验证了模型的组合策略的适用性，结果表明，平均法和 Stacking 法将传统的神经网络有机组合在一起可进一步提高浅层神经网络的泛化能力。值得注意的是，Stacking 法的模型组合策略有效提高了模型的计算精度、计算的稳定性和泛化能力。本文不仅验证了集合学习在药物分子 pKa 预测问题中的适用性，也验证了合理的组合计算策略可以达到深度神经网络的预测精度，同时能保持更好的泛化

能力，为高效准确预测药物分子性质提供了新的开源计算方法。在下一步的研究中，可将该方法拓展到对其他机器学习方法的融合中，以及利用融合模型策略方法预测药物分子的其他性质。

参 考 文 献

[1] DANISHUDDI N, KHAN A U. Descriptors and their selection methods in QSAR analysis: paradigm for drug design[J]. Drug Discovery Today, 2016, 21(8): 1291-1302.

[2] CHERKASOV A, MURATOV E N, FOURCHES D, et al. QSAR modeling: Where have you been? Where are you going to? [J]. Journal of Medicinal Chemistry, 2014, 57(12): 4977-5010.

[3] SUN Z, LU C, SHI Z, et al. Reasearch and advances on deep learning[J]. Computer Science, 2016, 43(2): 1-8.

[4] TIAN Q, WANG M. Research progress on deep learning algorithms. Computer Engineering and Applications [J]. 2019, 55(22): 25-33.

[5] LECUN Y, BENGIO Y, HINTON G. Deep learning[J]. Nature, 2015, 521(7553): 436-444.

[6] CHAN H C S, SHAN H, DAHOUN T, et al. Advancing drug discovery via artificial intelligence[J]. Trends in Pharmacological Sciences, 2019, 40(8): 592-604.

[7] SHI X Y, YU L, TIAN S. et al. Research on calssification of oral bioavailability based on deep learning [J]. Computer Science, 2016, 43(4): 260-263.

[8] SHEN C, DING J, WANG Z, et al. From machine learning to deep learning: Advances in scoring functions for protein-ligand docking[J]. WIREs Computational Molecular Science, 2020, 10(1): e1429.

[9] SEGLER M H S, KOGEJ T, TYRCHAN C, et al. Generating focused molecule libraries for drug discovery with recurrent neural networks[J]. ACS Central Science, 2018, 4(1): 120-131.

[10] SMITH J S, ROITBERG A E, ISAYEV O. Transforming computational drug discovery with machine learning and AI[J]. ACS Medicinal Chemistry Letters, 2018, 9(11): 1065-1069.

[11] XU Y, YAO H, LIN K. An overview of neural networks for drug discovery and the inputs used[J]. Expert Opinion on Drug Discovery, 2018, 13(12): 1091-1102.

[12] FEINBERG E N, JOSHI E, PANDE V S, et al. Improvement in ADMET prediction with multitask deep featurization[J]. Journal of Medicinal Chemistry, 2020, 63(16): 8835-8848.

[13] 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.

[14] LEI T, SUN H, 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.

[15] FU L, LIU L, YANG Z J, et al. Systematic modeling of log D7.4 based on ensemble machine learning, group contribution, and matched molecular pair analysis[J]. Journal of Chemical Information and Modeling, 2020, 60(1): 63-76.

[16] LIAO C, NICKLAUS M C. Comparison of nine programs predicting pKa values of pharmaceutical substances[J]. Journal of Chemical Information and Modeling, 2009, 49(12): 2801-2812.

[17] 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.

[18] ZHOU Z H, WU J, TANG W. Ensembling neural networks: Many could be better than all[J]. Artificial Intelligence, 2002, 137(1): 239-263.

[19] MIN S, LEE B, YOON S. Deep learning in bioinformatics[J]. Briefings in Bioinformatics, 2016, 18(5): 851-869.

[20] WISHART D S, FEUNANG Y D, GUO A C, et al. DrugBank 5.0: a major update to the DrugBank database for 2018[J]. Nucleic Acids Research, 2018, 46(D1): D1074-D1082.

[21] CHUANG K V, GUNSALUS L M, KEISER M J. Learning molecular representations for medicinal chemistry[J]. Journal of Medicinal Chemistry, 2020, 63(16): 8705-8722.

[22] DUAN J, DIXON S L, LOWRIE J F, et al. Analysis and comparison of 2D fingerprints: Insights into database screening performance using eight fingerprint methods[J]. Journal of Molecular Graphics and Modelling, 2010, 29(2): 157-170.

[23] LI L, KOH C C, REKER D, et al. Predicting protein-ligand interactions based on bow-pharmacological space and Bayesian additive regression trees[J]. Scientific Reports, 2019, 9(1): 7703.

XIE Liang-xu, born in 1987, postgraduate, associate professor, is a member of China Computer Federation. His main research interest includes AI aided drug design and data mining.

XU Xiao-jun, born in 1979, professor, Jiangsu distinguished professor. His main research interest includes computational biophysics and AI aided biomolecules structure prediction.