·实验研究·

融合堆叠自编码神经网络算法和全连接神经网络算法的化合物成药性预测模型

潘　蕾,倪冰苇,赵鸿萍
(中国药科大学理学院,南京 211198)

[摘要] **目的:**基于深度学习方法建立更加稳定、可靠、高实用性的化合物成药性预测模型。**方法:**通过 Integrity,Chembl 和 DrugBank 这 3 个数据库收集正、负样本数据,对正负样本大数据集进行数据清洗、解决数据不平衡问题之后,进一步对化合物的简化分子线性输入规范(SMILES)码进行标准化编码,在此基础上基于堆叠自编码神经网络算法(Stacked AutoEncoder,SAE)以及全连接神经网络算法(Fully Connected Neural Network,FCNN)构建并训练深度神经网络模型,对化合物进行特征提取,预测化合物的成药性。**结果:**模型最终稳定收敛,在验证集上准确率(ACC)和曲线下面积(AUC)分别达到 0.995 3 和 0.992 7,较之前文献报道的基于机器学习的模型提高了约 3% 的预测精度。**结论:**基于大数据集和深度神经网络技术构建的化合物成药性预测模型具备一定的实用性,可以提高化合物成药性预测的精准度。

[关键词] 堆叠自编码神经网络;全连接神经网络;深度学习;SMILES 码;成药性预测
[中图分类号] R95　　[文献标志码] A　　[文章编号] 1003 - 3734(2021)14 - 1309 - 07

Prediction model of the probability of a lead compound becoming a drug based on Stacked AutoEncoder and Fully Connected Neural Network

PAN Lei, NI Bing-wei, ZHAO Hong-ping
(*School of Science*, *China Pharmaceutical University*, *Nanjing* 211198, *China*)

[**Abstract**] **Objective**: To build a more stable, reliable and practical model for the probability prediction of a lead compound becoming a drug based on the deep learning method. **Methods**: The positive and negative sample data sets were collected from Iintegrity, Chembl and Drugbank databases firstly. After cleaning the large data set of positive and negative samples and solving the problem of data imbalance, the compounds' SMILES were further encoded. Then, Stacked AutoEncoder (SAE) and Fully Connected Neural Network (FCNN) were used to construct and train the deep neural network model to extract the features of the compounds and predict the probability of a lead compound becoming a drug. **Results**: The model finally converged stably, the ACC value and AUC value reached 0.995 3 and 0.992 9 respectively on the validation set, which improved the prediction accuracy by about 3% compared with the previously reported model based on machine learning. **Conclusion**: The prediction model based on large data set and deep neural network technology has certain practicability, and can improve the accuracy of the probability prediction of a lead compound becoming a drug.

[**Key words**] Stacked AutoEncoder; Fully Connected Neural Network; deep learning; SMILES; probability prediction of a lead compound becoming a drug

---

[基金项目] 国家自然科学基金面上项目(81973512);中国药科大学校级教学改革研究课题重点项目(3050050188)
[作者简介] 潘蕾,女,副教授,主要从事深度学习与生物医药数据挖掘研究。E-mail: yaya5185@ sina. com。
[通讯作者] 赵鸿萍,女,教授,主要从事深度学习与大数据挖掘研究。E-mail: zhaohongping@ cpu. edu. cn。

药物化学家 Lipinski 及其同事们分析了待上市的候选药物,发现由于较差的理化性质和生物学特性致使约 40% 的候选药物未能成功上市。然而这些候选药物的理化性质特别是生物学特性往往需要历经大量实验,到药物研发后期才能明确,这一现象导致药物研发的试错成本极高[1]。

为了解决这类问题,人们开始尝试从多角度出发,探索类药性规则。从现有研究成果来看,基于分子特性的类药性评判规则存在预判使用的特征值少、假阳性高的问题[2];基于类药官能团或者骨架结构的药物设计虽然在一定程度上提高了药物研发的成功率,但不利于全新结构的药物设计与研发[3]。近年来兴起的利用机器学习方法构建的成药性预测模型,主要基于贝叶斯神经网络、支持向量机、朴素贝叶斯和分类决策树的算法构建,将成药性预测的精准度提升到了 96.9%,但仍然没有根本解决假阳性高的问题[4-6]。

鉴于以上背景,本研究尝试通过多个数据库收集正、负样本数据,从而增加训练集数据量,提高数据多样性。在对正负样本进行数据清洗处理后,完成对样本化合物 SMILES 码的编码,实验基于堆叠自编码神经网络算法(SAE)对编码进行特征提取,结合全连接神经网络算法(FCNN)构建并训练深度神经网络模型,以期实现一个稳定、可靠、实用性较高的化合物成药性预测模型。

数据与方法

1 数据

1.1 数据采集

通过 Integrity,Chembl 和 DrugBank 数据库收集所有上市药物和达到Ⅲ期临床试验阶段的化学药物,按照 InChIKey 去重后得到共计 3 351 个药物作为正样本;同时,收集 Integrity 和 DrugBank 数据库中研发失败的药物,合并 Chembl 数据库中的所有活性化合物作为负样本候选数据集。

1.2 数据清洗

在解决实际问题时,所采集的数据可能包含大量的缺失值,也可能存在大量的特异值[7]。数据清洗的目标是使数据的意义明确、内容正确、格式规范,从而提高数据的可用性。在本课题中,由于化合物来源于业界认同度很高的三大数据库,数据整体质量较高,但仍然存在某些数据信息不一致的问题,主要表现在以下 3 个方面:① 化合物结构描述不统一。② 小分子化合物中混杂着大分子化合物。③ 负样本候选数据集中有疑似正样本的化合物。

针对以上 3 个问题,首先利用 Openbabel 将采集到的化合物的结构信息统一转换成 canonical SMILES 码;其次为了剔除非小分子化合物,利用 Openbabel 对化合物进行去盐加氢处理,并计算分子量,剔除掉分子量 >900 的化合物,得到仅含小分子化合物的数据集;针对负样本候选数据集中的疑似正样本的化合物,首先利用 chemfp 计算所有化合物的分子指纹,并基于分子指纹,计算所有正样本和负样本化合物的结构相似度,然后从负样本候选数据集中剔除与任意正样本化合物的结构相似度 >0.85 的类药化合物,最终得到共计 1 048 512 个非药物化合物作为负样本数据集。

1.3 化合物编码

本模型的化合物编码采用了基于化合物 SMILES 码的自定义字符编码。

1.3.1 SMILES 码 利用计算机处理和存储化学信息时,首先依赖于对化学结构的描述,尤其在构建化合物相关的预测模型时,能够采用精确而有意义的编码表示化学实体至关重要。现有的化学结构的描述方法主要包括:系统命名法、碎片码、线性码、拓扑指数码、连接表等。SMILES 全称为简化分子线性输入规范,属于线性码,是一种用 ASCII 字符串明确描述分子结构的规范。一个化合物结构往往可以书写为多个 SMILES,但是一个 SMILES 只能表示一个确定的化合物结构。使用 canonicalization 算法生成 canonical SMILES 可以保证一个化合物结构只能书写为一个 canonical SMILES。canonical SMILES 因具有唯一性、可读性高、存储空间小等优点,被广泛应用于数据库搜索和化合物模型构建中。SMILES 码结构紧凑,一个典型的 SMILES 码所占用的存储空间比一个等效的连接表(甚至是二进制连接表)下降 50% ~70%。例如,一个包含 23 137 个结构(平均每个结构有 20 个原子)的数据库,使用 SMILES 表示时每个原子仅占用 1.6 字节。此外,SMILES 的常规压缩也效果显著,通过 Ziv-Lempel 压缩可以将每个原子所占空间压缩至 0.42 字节,从而将以上数据库的规模再降低至原来的 27%[8]。

1.3.2 自定义字符编码　本课题统计了所有药物和非药物化合物的 canonical SMILES 码中出现的不同化学元素的个数，共计 61 个，因此每个原子可采用 6 位的二进制编码表示。同时，设定每个化合物的 canonical SMILES 码字符串长度不超过 120（SMILES 码字符串长度大于 120 的化合物在数据清理阶段已剔除）。具体编码过程包括 3 步：① 统计得到 canonical SMILES 字符串中不重复字符的个数为 61。② 对样本中字符数量小于 120 的编码，使用空格将字符串向右补足至 120 个字符。③ 用 6 位的二进制数表示化合物 canonical SMILES 编码中的每一个字符（包括空格）。

综上，每个化合物就表示为 720 位的 0 或 1，以这样的方式得到的 720 位数据作为 SAE + FCNN 模型的化合物编码，即模型输入向量。

同时，设置模型正样本的输出向量为[1,0]，负样本的输出向量为[0,1]。

2　方法

2.1　FCNN

FCNN 是全连接神经网络，结构见图 1。

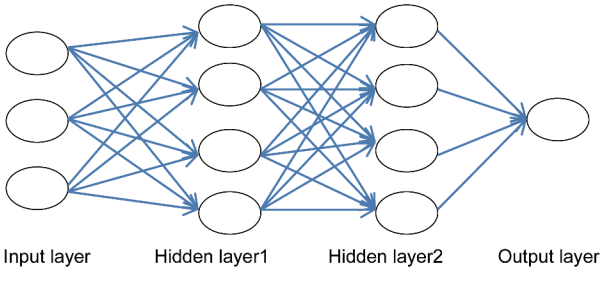


图 1　FCNN 示意图

FCNN 的结构不固定，通常包含：输入层、隐藏层和输出层。一般来说第一层是输入层，最后一层是输出层，而中间的层数都是隐藏层。神经元是构成神经网络的基本单位，层内神经元互不相连，层与层之间的神经元是全连接的，即第 n 层的任意一个神经元一定与第 $n+1$ 层的任意一个神经元相连，神经元之间的每个连接都可以向另一个神经元传递信号。所有神经元都接受上一层网络所有神经元的输出作为输入，并处理这些信号，然后向与之相连的下游神经元发出信号。每个输入都与一个随着学习过程而变化的权重相关，该权重可以增加或减少向下游发送的信号强度。神经元在计算时由输入向量 x、线性加权、激活函数、输出向量这 4 个部分组成，其中激活函数要求非线性，以方便求导数。在训练前事先设置好神经网络的连接方式、网络层数、每层的节点数等超参数，训练时通过算法不断优化每个神经网络的每个连接上的权值与偏差，使输出层的值与实际输出向量最为接近，最终得到的权值和偏差即为训练所需的结果。FCNN 的优点是其具有非常强的非线性拟合能力，几乎可以拟合任何函数，但深层的 FCNN 网络由于参数基数大，要想获得好的训练效果则需要大量数据和训练技巧作为支撑[9-11]。

2.2　SAE

AE（autoencoder）是一种典型单隐层全连接神经网络，其包含 1 个输入层，1 个隐藏层和 1 个输出层，是一种全连接前馈神经网络，其结构见图 2。从输入到隐藏层的转换称为编码，从隐藏层到输出层的转换称为解码。AE 尝试学习到一种非线性映射，由于限制 AE 隐藏层的节点数必须比输入层小，这种设定会迫使模型通过学习数据中的相关性从而在隐藏层中创建输入数据的压缩表示，再由隐藏层的激活向量重构出输入向量，从而使模型能学习输入数据隐含的结构和相关性特征[12]。

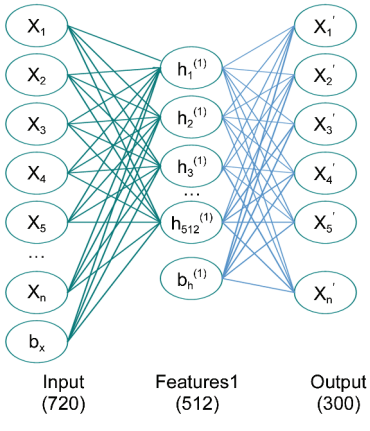


图 2　AE 示意图

SAE 网络是由多层 AE 堆叠构成的层级深度神经网络，是一种基于逐层贪婪的无监督学习方法。SAE 网络在预训练时第 1 层 AE 通过最小化重构误差将 X 维的原始输入数据映射到第 1 个含有 n_1 个神经元的隐藏层，得到 n_1 维的 X'，即训练 X-X'-X 这个自编码器，训练结束后保留编码层的权重和偏差；以编码层得到的 n_1 维的特征向量 X' 作为输入向量，再训练一个 X'-X''-X' 的自编码器，然后保留第 2

层 AE 的网络参数;以第 2 层 AE 编码得到的特征向量 X" 再作为下一层的输入向量,训练第 3 层的 AE,……,以此类推,通过最小化输入和输出信号的重构误差来完成对输入数据的层层压缩,实现对输入数据的特征提取。通过这种方式,对整个 SAE 网络进行逐层预训练直到得到最后一个 AE 层,保存每一个 AE 的编码器参数,并将每个编码层串联起来形成预训练过的 SAE 模型,该 SAE 模型可以对化合物原始结构信息进行表征学习和降维[13-14]。

3 模型构建与训练

3.1 模型构建

由于 FCNN 参数多、不易训练,尝试采用堆叠自编码器 SAE 算法,首先构建了特征提取网络;之后将具有 2 个神经元的分类器连接到 SAE 网络上,最终构成预测化合物成药性的二分类网络,网络结构见图 3。其中,隐藏层共使用 2 层,第 1 层节点数为 512,第 2 层节点数为 300。

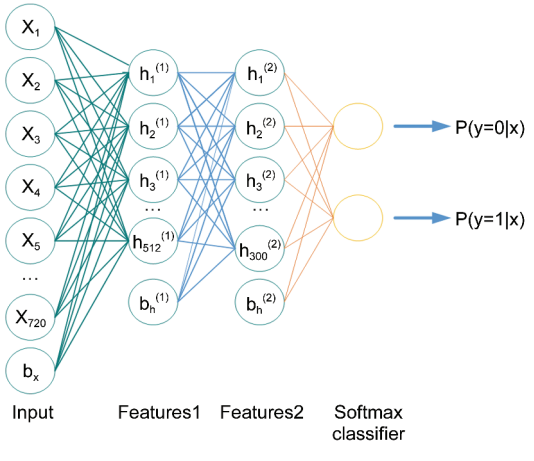


图 3 SAE + FCNN 预测化合物成药性模型架构

3.2 模型训练

模型的训练过程包括 2 个步骤:首先,利用 SAE 算法对特征提取网络进行逐层训练;之后接入分类层,将 SAE 训练得到的权重和偏差作为网络的初始化参数,并通过反向传播对整个网络进行微调,通过最小化目标变量的预测误差来获得改进的权重,直至模型稳定收敛。

模型训练过程主要调节了激活函数、层数和节点数、损失函数、优化器、batch_size 和学习率等超参数,并进行了过拟合处理。SAE 训练特征提取网络的超参数设置见表 1。

表 1 SAE 训练特征提取网络的超参数设置

Hyperparameter	Setting
Initializer	Truncated Normal
Number of hidden layers nodes	512, 300
Dropout	0.2
Regulaization	L2
Activation	sigmoid
Batch size	600
Optimizer	Adam
Learning rate	0.000 1
Loss function	mse

其中,初始化函数使用 Truncated Normal 产生截断正态分布随机数;特征提取网络共有 2 个隐藏层,节点数分别是 512 和 300;为了防止过拟合,在 SAE 训练时使用了 L2 正则化并将输入层的 Dropout 设为 0.2;自编码器使用 sigmoid 作为激活函数;Batch size 设为 600;优化器采用自适应矩估计(adaptive moment estimation,Adam)对模型进行训练, Adam 初始学习率(learning rate, Lr)设置为 0.000 1,之后 Adam 利用梯度的一阶矩估计和二阶矩估计动态调整 Lr,使得参数平稳变化,直至模型收敛。

训练化合物成药性预测的分类模型时,特征提取层超参数保留 SAE 训练时的参数值不变,权重和偏差的初始值设置为 SAE 训练结果值;分类层节点数设为 2;激活函数采用分类函数 softmax;此外,分类模型的损失函数设为 crossentropy。

结　果

在 SAE + FCNN 化合物成药性预测模型构建并训练完成后,对模型进行了全面评估,主要包括收敛性评估和性能评估两方面。其中,性能评估采用了 9 折交叉验证法,主要评价指标包括均方误差(mean-square error, MSE)、二元交叉熵(binary cross-sentropy)、准确率(accuracy, ACC)和 ROC 曲线下的面积(area under curve, AUC)。

1 收敛性评估

化合物成药性预测的分类模型在第 20 个 Epoch 之前,模型训练集上的 ACC 和 AUC 稳定上升,验证集上的 ACC 和 AUC 在开始时下降,后逐渐上升,在第 20 个 Epoch 值之后,模型训练集和验证集上的 ACC 和 AUC 趋于稳定,保持在约 0.99 基本不变;同时,模型在训练集和验证集上的二元交叉熵损失函数从第 1 个 Epoch 到第 40 个 Epoch 之间平稳下降,

40 个 Epoch 值之后达到了稳定,见图 4,这说明模型已完全收敛。

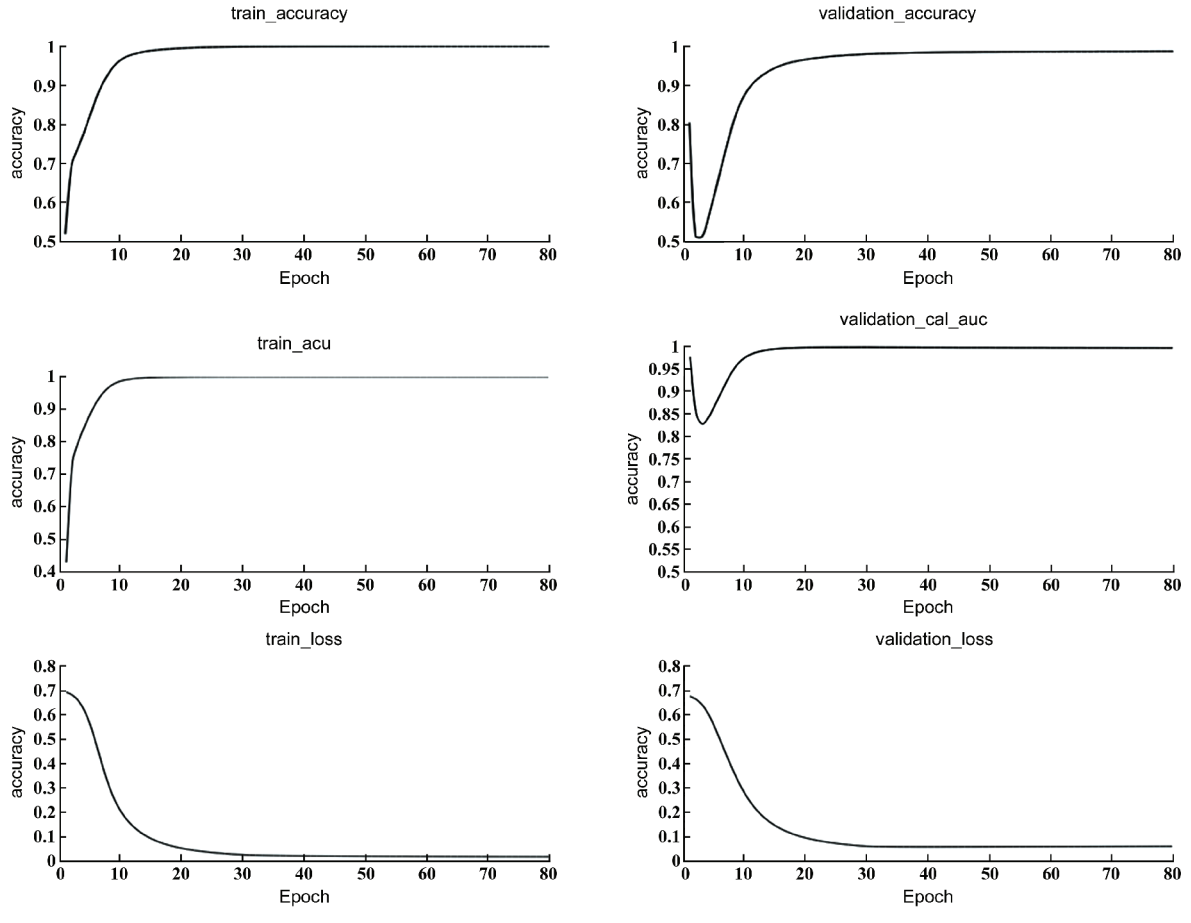


图 4 SAE + FCNN 分类模型在训练集和验证集上的 ACC,AUC,Loss 趋势图

2 9 折交叉验证法

k 折交叉验证(*k*-fold cross-validation)是将训练集分割成 *k* 个子样本,一个单独的子样本被保留作为验证模型的数据,其他 *k*-1 个样本用来训练。交叉验证重复 *k* 次,每个子样本验证 1 次,平均 *k* 次的结果或者使用其他结合方式,最终得到一个单一估测。该方法的优势在于重复运用随机产生的子样本进行训练和验证,从而能更合理地评估模型的泛化能力。

本课题选用 9 折交叉验证法来评估模型,即将正样本集和负样本集的数据都随机等分成 9 份,轮流将其中正样本和负样本的 8 份整合在一起作为训练数据集,1 份作为验证数据集,以验证数据集上的评价指标作为评价模型质量的依据。

3 评价指标

3.1 均方误差(MSE)

$$\mathrm{MSE} = \frac{1}{N}\sum_{t=1}^{N}(\mathrm{observed}_t - \mathrm{predicted}_t)^2$$

(公式 1)

SAE 训练特征提取网络使用的评估指标为 MSE。如公式 1 所示,MSE 是指参数估计值与参数真值之差平方的期望值,是回归模型中常用的损失函数。MSE 可以评价数据的变化程度,MSE 的值越小,说明预测模型描述实验数据具有更好的精确度,在本课题中,使用 MSE 均方误差作为损失函数调整每一层自编码器的参数[15]。

3.2 二元交叉熵

$$L = -[y \times \log(p) + (1 - y) \times \log(1 - p)]$$

(公式 2)

如公式 2 所示,本课题采用二元交叉熵作为接入分类器之后的 SAE 化合物成药性预测模型的损失函数,其中 y 表示样本的类别,正样本为 1,负样本为 0;p 表示样本预测为正样本的概率。在采用二元交叉熵作为损失函数时,必须对输出向量进行 one-hot 编码,正样本的输出向量为[1,0],负样本的输出向量为[0,1],此时激活函数需要使用 softmax 函数。

3.3 ACC

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \quad (公式\ 3)$$

如公式 3 所示,ACC 表示预测正确的样本数量占总样本数量的比例。评估一个分类模型表现的指标有很多种,其中混淆矩阵应用最广。混淆矩阵通常被用于衡量机器学习中分类模型的预测结果表现。有 4 个一级指标,在本课题中,其中真阳性 TP 表示:预测值 =[x,y],其中 $x > y$,并且真实值 =[1,0];假阳性 FP 表示:预测值 =[x,y],其中 $x > y$,并且真实值 =[0,1];假阴性 FN 表示:预测值 =[x,y],其中 $x < y$,并且真实值 =[1,0];真阴性 TN 表示:预测值 =[x,y],其中 $x < y$,并且真实值 =[0,1]。

3.4 AUC

AUC 即 ROC 曲线下的面积。ROC 曲线包括横纵坐标两部分,纵坐标即真正类率(true postive rate, TPR),如公式 4 所示,代表分类器预测的正类中实际正样本占所有正样本的比例。横坐标即负正类率(false postive rate, FPR),如公式 5 所示,代表分类器预测的正类中实际负样本占所有负样本的比例。理想目标为 TPR = 1,FPR = 0,即图中(0,1)点,故 ROC 曲线越靠拢(0,1)点,越偏离 45 度对角线越好,灵敏性、特异性越大效果越好。AUC 值介于 0.1 和 1 之间,可以直观地评价分类器的好坏,值越大分类器效果越好。

$$TPR = \frac{TP}{TP + FN} \quad (公式\ 4)$$

$$FPR = \frac{FP}{FP + TN} \quad (公式\ 5)$$

4 性能评价结果

模型精度评测指标值见表 2。

表 2 SAE + FCNN 模型性能评测结果

评测指标	数值
mse(layer 1)	0.049 4
val_mse(layer 1)	0.046 5
mse(layer 2)	0.000 014
val_mse(layer 2)	0.000 011 3
crossentropy	0.016 8
accuracy	0.999 1
cal_auc	0.999 8
val_loss	0.059 4
val_ accuracy	0.995 3
val_cal_auc	0.992 7

其中,SAE 训练模型第一层自编码器在训练集上的均方误差值为 0.049 4,在验证集上的均方误差值为 0.046 5;第二层自编码器在训练集上的均方误差值为 0.000 014,在验证集上的均方误差值为 0.000 011 3。化合物成药性预测的分类模型在训练集上的二元交叉熵损失函数 crossentropy 达到了 0.016 8,在验证集上的二元交叉熵损失函数 val_loss 达到了 0.059 4。同时,模型无论是在训练集上还是验证集上,分类的 ACC 和 AUC 这 2 个重要指标都超过了 0.99。

结果数据显示,本课题构建的 SAE + FCNN 化合物成药性预测的分类模型分类效果较好,模型具备一定的实用性。

讨　论

是否能建成一个稳定、可靠、实用性高的成药性预测模型,不仅和模型所选择的算法和结构高度相关,训练集数据量的大小与多样性、特征值的选择也直接影响模型的预测效果。研究中首次使用 SAE + FCNN 构建并训练深度神经网络模型,采用了对包括全部结构信息的 SMILES 码进行标准化编码来对化合物进行描述,目前模型训练集数据量约 1 050 000,远大于之前文献报道的模型样本数据量 300 000。训练集正样本覆盖了全部已上市药物和达到Ⅲ期临床试验的药物,负样本涵盖了所有非药的活性小分子,训练集数据具有采集全面、多样性好、数据量大的优点,能在一定程度上保障模型的普适性和实用价值,有效解决假阳性高的问题。最终模型达到验证集上 ACC > 0.99,AUC > 0.99,较之前文献报道的基于机器学习的模型(ACC:~0.969 9)提高了约 3% 的预测准确率,有效提高了化合物成药性预测的精准度。

本研究中还发现:当采用未经处理的全新测试集进行测试时,发现负样本预测精度较高(AUC > 0.998 5),而正样本预测精度偏低(AUC = 0.446 6)。分析造成问题的原因主要是正、负样本数据不平衡,加上正样本数据多样性大、数据量小,后期将着力于解决这 2 个问题,进一步提高正样本数据预测的准确率。随着人工智能、深度神经网络技术的发展,未来有望通过优化网络结构、提高模型训练技能,最终解决正样本分类数据量少及正、负样本数据不平衡的问题,进一步提高模型精准度,以期开发出一个实用性更强的化合物成药性预测平台。

[参 考 文 献]

[1] LIPINSKI CA, LOMBARDO F, DOMINY BW, *et al*. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings[J]. *Adv Drug Deliv Rev*, 2001, 46(1-3): 3-26.

[2] LU JJ, CRIMIN K, GOODWIN JT, *et al*. Influence of molecular flexibility and polar surface area metrics on oral bioavailability in the rat[J]. *J Med Chem*, 2004, 47(24): 6104-6107.

[3] WANG JM, HOU TJ. Drug and drug candidate building block analysis[J]. *J Chem Inf Model*, 2010, 50(1): 55-67.

[4] LI Q, BENDER A, PEI J, *et al*. A large descriptor set and a probabilistic kernel-based classifier significantly improve drug-likeness classification[J]. *J Chem Inf Model*, 2007, 47(5): 1776-1786.

[5] TIAN S, WANG JM, LI YY, *et al*. Drug-likeness analysis of traditional Chinese medicines: prediction of drug-likeness using machine learning approaches[J]. *Mol Pharm*, 2012, 9(10): 2875-2886.

[6] CHEN H, ENGKVIST O, WANG Y, *et al*. The rise of deep learning in drug discovery[J]. *Drug Discov Today*, 2018, 23(6): 1241-1250.

[7] 戴维.萨洛蒙著, 吴乐南. 数据压缩原理与应用[M]. 第2版. 北京: 电子工业出版社, 2003.

[8] SMILES:一种简化的分子语言[EB/OL]. (2018-07-03)[2020-05-01]. https://www.jianshu.com/p/8c915de5ad4d.

[9] RINIKER S, WANG Y, JENKINS JL, *et al*. Using information from historical high-throughput screens to predict active compounds[J]. *J Chem Inf Model*, 2014, 54(7): 1880-1891.

[10] 许煌, 许晓双, 高英, 等. 以PDE2为靶标的抑制剂药物高通量筛选模型的建立[J]. 中国新药杂志, 2020, 29(10): 1175-1180.

[11] FERRERO E, DUNHAM I, SANSEAU P. In silico prediction of novel therapeutic targets using gene-disease association data[J]. *J Transl Med*, 2017, 15(1): 182.

[12] 赵广军, 王旭初, 牛彦敏, 等. 基于SAE深度特征学习的数字人脑切片图像分割[J]. 计算机辅助设计与图形学学报, 2016, 28(8): 1297-1305.

[13] GODINEZ WJ, HOSSAIN I, LAZIC SE, *et al*. A multi-scale convolutional neural network for phenotyping high-content cellular images[J]. *Bioinformatics*, 2017, 33(13): 2010-2019.

[14] MAMOSHINA P, VOLOSNIKOVA M, OZEROV IV, *et al*. Machine learning on human muscle transcriptomic data for biomarker discovery and tissue-specific drug target identification[J]. *Front Genet*, 2018, 9(7): 242.

[15] DENG F. Diagnostic performance of deep learning-augmented radiology residents[J]. *Radiology*, 2020, 295(2): 200227.

编辑:杨青/接受日期:2021-02-08