基于改进的降噪自编码药物透血脑屏障预测*

周兴发[1], 禹 龙[2†], 田生伟[1], 李 莉[3], 王 梅[4]
(1. 新疆大学 软件学院, 乌鲁木齐 830008; 2. 新疆大学 网络中心, 乌鲁木齐 830046; 3. 新疆医科大学 医学工程技术学院, 乌鲁木齐 830011; 4. 新疆医科大学 制药部门, 乌鲁木齐 830054)

摘 要: 药物透血脑屏障是新药研发的一个重要因素。在传统栈式降噪自编码(stacked denoising autoencoder, SDAE)基础上,提出一种改进的 SDAE 药物透血脑屏障预测方法。利用主成分分析(principal components analysis, PCA)无监督训练一组权值初始化 SDAE,避免随机初始化权值造成模型收敛速度较慢的问题;然后为降噪自编码(denoising autoencoder,DAE)增加一层隐藏层,构造双隐层 DAE,提高单个 DAE 提取药物分子抽象特征的能力;融合 SDAE 最后两个 DAE 的第一层隐藏层输出作为 softmax 分类器的输入,最终实现药物透血脑屏障预测。实验表明,与传统的 SDAE 及浅层机器学习模型 SVM 相比,改进后的模型对药物透血脑屏障具有更好的预测效果。

关键词: 血脑屏障; 栈式降噪自编码; PCA; 双隐层; 融合输出
中图分类号: TP391 **文献标志码:** A **文章编号:** 1001-3695(2018)05-1355-05
doi:10.3969/j.issn.1001-3695.2018.05.016

Predicting drugs penetration across blood-brain barrier based on improved stacked denoising autoencoder

Zhou Xingfa[1], Yu Long[2†], Tian Shengwei[1], Li Li[3], Wang Mei[4]
(1. *College of Software, Xinjiang University, Urumqi* 830008, *China*; 2. *Network Center, Xinjiang University, Urumqi* 830046, *China*; 3. *College of Medical Engineering Technology, Xinjiang Medical University, Urumqi* 830011, *China*; 4. *Dept. of Pharmaceutical, Xinjiang Medical University, Urumqi* 830054, *China*)

Abstract: The ability to cross the blood-brain barrier (BBB) is a very important property in drug design. This paper proposed an improved method based on the traditional SDAE to predict the ability or inability of a drug to penetrate the BBB. Firstly, in order to accelerate the training speed, it extracted a set of features from the training data by unsupervised training based on principal components analysis as the initial weights for SDAE. Then, it added a hidden layer to DAE to improve the individual DAE's ability of extracting the characteristics of the drug molecules. Finally, the model concatenated the first hidden layer's outputs of the last two DAE as the inputs of the softmax classifier. Experimental results show that the improved model has better prediction effect on the drug through blood-brain barrier compared with the traditional SDAE and shallow machine learning model support vector machine (SVM).

Key words: blood-brain barrier; stacked denoising autoencoder(SDAE); principal components analysis(PCA); double hidden layer; concatenate outputs

0 引言

候选药物的药代动力学是一个重要考虑因素,好的药代动力学能满足到达靶标所需的浓度,同时又能限制它向其他地方分布从而降低其副作用。药代动力学中一个重要性质是看它能否穿透血脑屏障(blood-brain barrier,BBB)[1]。BBB 受多种因素的影响,如溶解性、亲脂性、胃肠道吸收、代谢与外排等。此外,脑组织结构复杂且具有特异性,特别是血脑屏障的存在使得治疗脑病药物难以进入脑组织发挥作用,导致采用传统的通过临床实验等方法来确定大量候选药物的 BBB 通透性耗时长和代价昂贵[2]。因此,为了节约资源,降低成本,加快新药的研发速度,通过计算机辅助药物设计的方法进行药物透血脑屏障预测是至关重要的。

近年来,基于回归法和机器学习的计算方法已广泛用于药物透血脑屏障的研究。黄斌[3]结合遗传算法和支持向量机(support vector machine,SVM)构建了药物透血脑屏障预测模型,总体预测精度达到 85.6%。Yan 等人[4]用二维自相关和 RDF 方法来描述分子结构并结合 SVM 及 ANN 取得了较好的结果。然而与黄斌一样,实验所用数据集较小,不能充分验证模型的泛化能力。Zhao 等人[5]结合 binomial-PLS、递归划分方法和决策树思想在 1 593 个化合物上取得了 95% 的测精度,但仅在单一测试集上验证模型,未作交叉测试,不能充分验证模型的泛化能力。

研究表明,相对 SVM、逻辑回归(logistic regression,LR)、最大熵方法等浅层学习方法深度神经网络中,非线性操作层级数

收稿日期: 2017-01-05; **修回日期:** 2017-02-28 **基金项目:** 国家自然科学基金资助项目(31160341)

作者简介: 周兴发(1990-),男,重庆人,硕士研究生,主要研究方向为计算机智能技术、生物信息学;禹龙(1974-),女(通信作者),新疆乌鲁木齐人,教授,硕士,主要研究方向为计算机智能技术、计算机网络(yul_xju@163.com);田生伟(1973-),男,新疆乌鲁木齐人,教授,博士,主要研究方向为计算机智能技术、自然语言处理;李莉(1978-),女,新疆乌鲁木齐人,副教授,硕士,主要研究方向为医学信息处理;王梅(1977-),女,新疆乌鲁木齐人,博士,主要研究方向为药物传输系统.

更多,能够拟合更加复杂的函数,学习到数据更抽象的特征。同时浅层网络模型学习依靠人工经验抽取样本特征,学习得到的是没有层次结构的单层特征;而深度学习通过对原始输入信号进行逐层抽象特征抽取,将样本在原空间的特征表示变换到新的特征空间,自动学习得到层次化的特征表示。

DAE 是深度学习中自动编码器的一种变形结构,它与卷积神经网络(convolutional neural network,CNN)、受限玻尔兹曼机(restricted Boltzmann machine,RBM)一样作为深度学习架构中的训练模块,具有良好的学习数据抽象特征的能力。SDAE[6]是由多个 DAE 堆叠形成的深层神经网络,通过无监督学习过程和对数据的"破坏"过程,相对于浅层学习方法,SDAE 能进一步学习到数据集中更本质的特征和数据结构。

本文在传统的 SDAE 基础上,通过改进其结构,并将它应用到药物透血脑屏障预测,在公开数据集上,使用准确率、召回率和马修斯相关系数作为衡量标准。同时与传统的 SDAE 和 SVM 等算法进行比较,结果表明,相对这些算法,改进的 SDAE 具有更好的鲁棒性和较好的预测能力。

1 数据源与特征提取

1.1 数据来源

为成功地建立药物透血脑屏障预测模型,收集大量可靠和一致的实验数据至关重要。本文采用 Adenot 等人[7]收集整理的实验数据,包含 1 336 个可透血脑屏障化合物(BBB +)和 360 个不可透血脑屏障的化合物(BBB -)。为避免重复性和不确定性,去除数据中 8 个既被标记为 BBB + 又被标记为 BBB-的化合物,以及缺失结构的化合物,最后筛选出了 1 283 个 BBB + 化合物和 310 个 BBB - 化合物作为最终实验数据。

1.2 药物分子特征生成与筛选

描述符是药物设计中描述化合物分子物化特征的基本要素。从结构特征上看,目前描述符主要分为以下三类:

a)1D 描述符。由化合物自身属性描述其分子特征,如脂水分配系数(logP)、溶解度(logS)、疏水原子数(a_hyd)、疏水表面积(vsa_hyd)等。

b)2D 描述符。主要有 2D 分子指纹、拓扑指数、子结构描述符等。

c)3D 描述符。包括 MQS(molecular quantum similarity)、QSCD(quantized surface complementarity diversity)、分子形状、分子总表面积等。

在这些描述符中,相较于 1D 和 3D 描述符,分子指纹具有特征性强、计算速度快等特点,使得分子指纹成功应用于分子相似性筛选等领域[8]。它通过检测分子结构中某些特定的分子结构片段是否存在,从而把分子结构转换为一系列二进制指纹序列。

在本文实验中,使用 paDEL-Descriptor[9]描述符计算软件生成常用的 Pubchem 分子指纹作为化合物的分子特征,最终生成了 613 维的分子特征。同时基于以下规则对分子特征进行预处理:a)去除标准差为 0 的分子特征;b)去除彼此之间相同的分子特征。

该过程的主要代码如下:

代码 1 分子指纹特征预处理代码

```
datas = get_experiment_data( ) ;remove = [ ]
for i in datas. num:
    if datas[i]. std( ) = = 0: remove. append(i)
endfor
datas. drop(remove) ;remove = [ ]
c = datas. columns
for i in range(datas. feature. num) :
    v = datas[c[i]]. values
    for j in range(i + 1, len(c)) :
     if equal(v,datas[c[j]]. values) :
            remove. append(c[j])
     endfor
endfor
datas. drop(remove)
```

通过上述筛选方法,最终选取化合物分子的特征表示共 546 维。

2 构建药物透血脑屏障预测模型

2.1 PCA 非监督训练初始化权值

PCA 是一种常用的数据分析方法。它对原先提出的所有变量,删除多余重复的,建立尽可能少的新变量,使得这些新变量间两两不相关,且使这些新变量在反映数据信息方面尽可能保持原有的信息。其可以视为一种简单的自动编码器。运用到神经网络中,PCA 能够极大提升无监督特征学习速度,降低数据维度。基于以上优点,PCA 成功应用在图像处理、特征选择等领域[10~13]。本文通过 PCA 非监督训练获得训练数据的主成分,使这些主成分能够最大程度地表示数据的特征,并作为 DAE 的初始化权值来提高模型的收敛速度。

假设有 N 个训练数据 $\boldsymbol{X}=(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N)^{\mathrm{T}}$,其中 x_i 表示第 i 个化合物的分子指纹特征。首先将 $\boldsymbol{X}$ 去均值,计算协方差矩阵,然后利用 PCA 方法最小化重构误差求特征向量。$\boldsymbol{X}$ 的均值为

$$\overline{\boldsymbol{X}}=\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{x}_i \tag{1}$$

每一个输入与均值的差值为

$$\widetilde{\boldsymbol{x}_i}=\boldsymbol{x}_i-\overline{\boldsymbol{X}} \tag{2}$$

则训练数据去均值后的输入为

$$\boldsymbol{X}=(\widetilde{\boldsymbol{x}}_1,\widetilde{\boldsymbol{x}}_2,\cdots,\widetilde{\boldsymbol{x}}_N)^{\mathrm{T}} \tag{3}$$

那么训练数据的协方差矩阵为

$$\boldsymbol{C}=\frac{1}{N}\sum_{i=1}^{N}\widetilde{\boldsymbol{x}}_i\,\widetilde{\boldsymbol{x}}_i^{\mathrm{T}} \tag{4}$$

利用 PCA 方法最小化重构误差求特征向量:

$$\min\|\boldsymbol{X}_i-\boldsymbol{U}_i\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{X}\|^2,\text{s.t. }\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{U}_i=\boldsymbol{E}_k\quad k=1,2,3\cdots \tag{5}$$

$$\boldsymbol{X}_i=\boldsymbol{U}_{i-1}^{\mathrm{T}}\boldsymbol{X}_{i-1}\quad i=2,3,4\cdots \tag{6}$$

其中:$\boldsymbol{X}_1=\boldsymbol{X}$;$\boldsymbol{E}_k$ 是 $k\times k$ 的单位矩阵;$\boldsymbol{U}_i\in\mathrm{R}^{n\times k}$ 是协方差矩阵 $\boldsymbol{X}_i\boldsymbol{X}_i^{\mathrm{T}}$ 的前 k 个特征向量。则 PCA 学习到初始化 SDAE 的一组初始化权值可表示为

$$\boldsymbol{W}_i^1=\mathrm{map}(\boldsymbol{U}_i)\quad i=1,2,3\cdots \tag{7}$$

其中:$\boldsymbol{W}_i^1\in\mathrm{R}^{n\times k}$ 表示 SDAE 的第 i 个 DAE 输入层到隐藏层的初始权值,n 为 DAE 输入层神经元个数,k 为隐藏层神经元个数;map 是将矩阵 $\boldsymbol{U}_i$ 一一映射到矩阵 $\boldsymbol{W}_i^1$ 的映射函数。

2.2 栈式降噪自编码

2.2.1 栈式自动编码器

自动编码器(autoencoder,AE)[14]是一种无监督学习网络。AE 共分为三层,即输入层、隐藏层和输出层。其输入层与输出层的维度相同。AE 尝试学习 $h_{w,b}(x)\approx x$ 这样一个函数,从而

使得输出 $h_{w,b}(x)$ 接近于输入 x ,将输入信号从目标中重构出来,实现样本特征的重新表述。从结构上划分,AE 可分为编码器和解码器两部分。

编码器完成从一个输入 $x \in \mathrm{R}^n$ 到输出 y 的映射转换,如式(8)所示。

$$y = f_\theta(x) = s(Wx + b) \tag{8}$$

其中:s 为 sigmoid(·)激活函数;$\theta = \{W, b\}$ 是编码参数集合。其表达式为

$$f(z) = \frac{1}{1 + \exp(-z)} \tag{9}$$

然后通过解码器将激活值 y 反向变换为对原始输入 x 的重构表示 z,y 与 z 满足式(10)。

$$z = g_{\theta'}(y) = s(W'y + b') \tag{10}$$

其中:$\theta' = \{W', b'\}$ 为解码参数。

最后通过不断地调整 θ 和 θ' 的值,得到最小化重构误差 J。重构误差的表达式为

$$J = \sum_{X \in D} L(x, z) \tag{11}$$

其中:D 为训练样本集;L 为重构误差函数。本文选用交叉熵误差函数,公式如下:

$$L(x, z) = -\sum_{k=1}^{n} [x_k \log z_k + (1 - x_k) \log(1 - z_k)] \tag{12}$$

为得到更抽象的特征表达,提高网络学习抽象特征的能力,把 AE 逐层堆叠起来,构成一个由 AE 上下连接而成的模型,即栈式自编码(stacked autoencoder,SAE)。相较于 SVM、决策树和逻辑回归等浅层机器学习方法,深层 SAE 神经网络模型能从无标签数据集中提取高维复杂输入数据更抽象的分层特征,得到原始数据的分布式特征表示[15]。

2.2.2　降噪自动编码器

AE 在学习过程中只是简单地保留原始输入样本的信息,并不能确保对输入样本提取出一种有用的抽象特征表示。因为 AE 可能仅仅简单地拷贝原始输入或者简单地选取能够稍微改变重构误差却不包含特别有用的信息。为避免这种情况,学习更好的特征表示,需要给数据表示一定的约束。DAE 通过重构含有噪声的输入数据来解决该问题,使之成功应用在特征提取、语音识别、动作识别等领域[16~18]。

DAE 与传统 AE 具有相同的结构,但在样本输入时按照某种方法加入了噪声。DAE 所实现的功能是学习加入了噪声的原始数据,虽然学到的特征与未叠加噪声的数据学到的特征几乎一样,但 DAE 从添加了噪声的输入中学习得到的特征更具鲁棒性,且可避免 AE 简单地学习相同特征值的问题。其计算过程如下:

假设输入样本为 X。DAE 通过一个随机的映射变换 $X \sim q_D(X' \mid X)$,对原始输入数据 X 进行“破坏”,从而得到一个含有噪声的数据 X',其中 D 表示数据集。则 DAE 的编码器输出为

$$Y' = f_\theta(X) = s(WX' + b) \tag{13}$$

然后通过解码器将激活值 Y'反向变换为对原始输入 X 的重构表示 Z,Y'与 Z 满足式(14)。

$$Z = g_{\theta'}(Y') = s(W'Y' + b') \tag{14}$$

2.3　Softmax 回归

Softmax 回归是逻辑回归在多分类问题上的扩展形式,通过估算样本所属类别的概率对多个类别的样本进行分类。假设训练样本及标签数据为$\{(x^1, y^1), \cdots, (x^n, y^n)\}$,$x^i$ 为训练样本,$y^i \in \{1, \cdots, k\}$ 为 x^i 对应的标签,则估计样本 x^i 在 softmax 函数中所对应的类别标签 k 的概率为

$$h_\theta(x^i) = \begin{bmatrix} p(y^i = 1 \mid x^i) \\ \vdots \\ p(y^i = k \mid x^i; \theta) \end{bmatrix} = \frac{1}{\sum_{j=1}^{k} \mathrm{e}^{\theta_j^{\mathrm{T}} x^{(i)}}} \begin{bmatrix} \mathrm{e}^{\theta_1^{\mathrm{T}} x^{(i)}} \\ \vdots \\ \mathrm{e}^{\theta_k^{\mathrm{T}} x^{(i)}} \end{bmatrix} \tag{15}$$

其中:$\theta_1, \theta_2, \cdots, \theta_k \in \mathrm{R}^{n+1}$ 为 softmax 的模型参数;$\frac{1}{\sum_{j=1}^{k} \mathrm{e}^{\theta_j^{\mathrm{T}} x}}$是对概率分布进行归一化,使得所有概率和为 1。本文利用 SDAE 提取出的化合物药物分子抽象特征作为 softmax 的输入,解决药物透血脑屏障预测问题。

2.4　优化的栈式降噪自编码

深层次的神经网络能获得数据更抽象层次的特征,然而随着神经网络层数的增加,会出现梯度消失问题[15]。传统的 SDAE 采用堆叠单隐层的 DAE 构造深度网络结构,然后逐层训练 SDAE 获得各层参数。为提高单个 DAE 提取抽象特征的能力,本文为除最后一个 DAE 外的每一个 DAE 增加一层隐藏层,构造双隐层 DAE,其神经元个数为下一层 DAE 的第一隐层的神经元个数,然后采用传统 SDAE 逐层预训练 DAE 方法获得模型的权值。同时为充分利用提取的化合物分子的抽象特征表示,融合 SDAE 最后两个 DAE 的第一隐藏层的输出作为 softmax 分类器的输入,如图 1 所示。

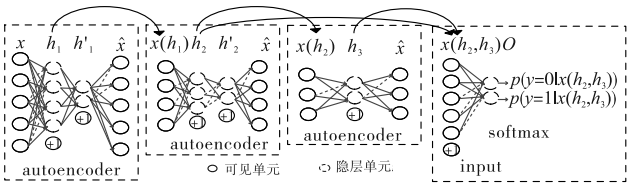

图1　双隐层SDAE体系结构

基于以上内容和 2.1 节提出的 PCA 非监督预训练初始化权值方法,本文算法的具体步骤如下:

a)利用 PaDEL-Descriptor 描述符计算软件生成 pubChem 分子指纹特征,构造数据集。

b)按照 1.2 节的预处理代码对分子指纹特征进行预处理。

c)PCA 预训练初始化权值。

(a)用原始训练数据集初始化 X。

(b)按照式(1)~(3)的原理构造初始数据集 X_i,并按照式(4)计算对应的协方差矩阵 C_i。

(c)按照式(5)训练得到最优 U_i,然后按照式(7)初始化第 i 个 DAE 输入层到第一隐层的权值。

(d)使用满足正态分布的随机化初始化方法初始其他各层间的权值。

(e)判断是否初始化完所有 DAE 输入层到第一隐层间的权值,若是结束本过程,否则跳到步骤 f)。

(f)递增 i,然后按照式(6)计算 X_i,并用 X_i 对 X 重新赋值,执行步骤 b)。

d)按照 2.4 节方法构造双隐层 DAE,逐层训练。

e)合并 SDAE 最后两个 DAE 的第一隐层输出,作为全连接输入,利用 softmax 分类器进行预测。

f)通过反向传播算法从后往前微调更新 SDAE 的权值参数。

3 实验与分析

3.1 评估标准

为衡量预测性能,使用 BBB + 的准确率(BBB + precision, apr)、BBB + 的召回率(BBB + recall, are)、BBB - 的准确率(BBB - precision, dpr)、BBB - 的召回率(BBB - recall, dre)以及准确率(accuracy, acc)作为预测指标。分类的混淆矩阵如表1所示,其中真正例(true positive, TP)、假负例(false negative, FN)、假正例(false positive, FP)、真负例(true negative, TN)。

$$apr = TP/(TP + FP) \tag{16}$$

$$are = TP/(TP + FN) \tag{17}$$

$$dpr = TN/(TN + FN) \tag{18}$$

$$dre = TN/(TN + FP) \tag{19}$$

$$acc = (TP + FP)/(TP + FP + TN + FN) \tag{20}$$

表1 分类混淆矩阵

总样本	分类为 BBB +	分类为 BBB -
真实正例	*TP*	*FN*
真实负例	*FP*	*TN*

文献[19]的研究表明,在机器学习的分类问题中,马修斯相关系数(Matthews correlation coefficient, MCC)是最好的性能测试指标之一。相比其他指标,MCC 对于准确率提供了更为平衡的评估。MCC 的返回值为 -1 ~ +1: -1 表示是相反的预测;0 表示随机猜测; +1 表示的则是完美的猜测。MCC 的计算公式如式(21)所示。

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{21}$$

3.2 实验结果与分析

Theano[20]是目前常用的深度学习框架,本文所有模型基于 Theano 实现。同时采用批量梯度下降法更新模型参数。为保证实验结果的稳定性,所有实验均首先将 BBB + 、BBB - 数据随机打乱,然后采用5折交叉验证,取五次结果的平均值作为最终结果。最终通过网格搜索算法反复实验尝试网络模型不同的参数组合,确定了本实验数据量下的最优参数,如表2所示。

表2 模型最优参数

参数	值
α	0.1
β	0.3
λ	0.01
numepochs	200
batchsize	10

表2中,α 表示模型训练过程中的学习率;β 表示对输入进行降噪处理的概率参数;λ 表示权重衰减项;numepochs 表示训练迭代次数;batchsize 表示每一次迭代训练,批量处理样本个数。

3.2.1 降噪自编码层数选择对分类结果的影响

SDAE 的 DAE 层数选择对实验结果有一定影响,选取合适的 DAE 层数非常关键。经过数据预处理后,实验选择化合物分子特征维度为 546 维,采用 DAE 层数从 3 ~ 7 层递增的 SDAE 进行药物透血脑屏障的预测对比分析实验。实验结果如表3所示。

表3 DAE 层数的有效性验证

模型	are/%	apr/%	dre/%	dpr/%	acc/%	MCC
$SDAE^3$	96.5	98.6	93.2	84.0	95.9	0.863
$SDAE^4$	97.2	99.0	94.4	88.6	96.8	0.891
$SDAE^5$	97.6	99.2	96.3	89.7	97.3	0.913
$SDAE^6$	97.2	99.0	94.6	87.8	96.7	0.891
$SDAE^7$	96.9	98.2	92.4	86.2	96.0	0.868

由表3实验结果可知,MCC 值都在 0.86 以上,说明改进的 SDAE 对药物透血脑屏障具有较好的预测效果。同时,随着 DAE 层数的增加,are 值变化不大,在 1.1 个百分点内波动;而 dre 值则在 3.9 个百分点内变化,且对应 are 值高于 dre 值。这是由于数据集中 BBB + 和 BBB - 所占比例不均衡,BBB + 所占比例较大(占 80%),模型能够学习到更多的 BBB + 的特征,所以 are 值较高。可以看到,acc 值在 1.4 个百分点内变化;随着层数的增加,acc 值逐渐上升,当模型层数增加到5层时,药物透血脑屏障预测的 acc 值最高,达到 97.3%。因此,在后续的实验中,均采用5层改进后的栈式降噪自编码器对药物透血脑屏障预测进行对比实验。

3.2.2 模型对比实验

为验证优化后的 SDAE 对药物透血脑屏障预测的有效性,实验采用在相同数据集上比较改进后的 SDAE(MSDAE)与传统的 SDAE、双隐层 SDAE(DSDAE)、融合了 SDAE 最后两个 DAE 的第一隐藏层输出的 SDAE(CSDAE)及浅层机器学习方法 SVM 的预测准确率。实验结果如表4所示。

表4 不同模型预测结果

模型	BBB +			BBB -			acc/%
	FN	*TP*	are/%	*FP*	*TN*	dre/%	
SDAE	6	248	97.6	7	40	85.1	95.7
DSDAE	8	240	96.8	3	50	94.3	96.4
CSDAE	10	243	96.0	1	47	97.9	96.3
MSDAE	6	241	97.6	2	52	96.3	97.3
SVM	12	243	95.3	4	42	91.3	94.7

从表4可以看出,改进后的 SDAE 与传统的 SDAE 药物透血脑屏障预测召回率比 SVM 分别提高了 2.6% 和 1.6%。这是因为 SDAE 不同于传统的浅层机器学习模型,其利用深度学习思想,对分子特征有更本质的学习,可以自动学习分子特征之间的内在表示,从而提升模型分类的准确性;同时 DSDAE 相较于 SDAE,acc 值提高了 0.7%,验证了双隐藏 DAE 能够提高 DAE 提取化合物分子抽象特征的能力;同时,CSDAE 和 DSDAE 组合构造的 MSDAE 相较于 SDAE,acc 值提高了 1.6%,尤其是 dre 值提高了 11.2%,说明在小数据量下 MSDAE 比 SDAE 更利于药物透血脑预测。表4实验结果表明,改进的 SDAE 的药物透血脑屏障预测模型要优于传统的 SDAE 和 SVM 模型。

3.2.3 不同输入特征对模型预测结果的影响

为验证不同输入特征对模型预测性能的影响,本文分别使用分子模拟软件 Molecular Operating Environment(MOE)和 Accelrys Discovery Studio 生成化合物的 2D 和 3D 分子特征,使用仅含 2D 分子特征、仅含 3D 分子特征,以及融合 2D 和 3D 分子特征作为模型的输入,比较不同输入特征下改进的 SDAE 药物透血脑屏障预测性能。实验结果如表5所示。

表5　不同特征下的预测结果

分子特征	BBB +			BBB −			acc/%
	FN	*TP*	are/%	*FP*	*TN*	dre/%	
2D	8	238	96.7	5	49	90.7	95.7
3D	15	234	94.0	7	45	86.5	92.7
2D + 3D	7	240	97.2	4	50	92.6	96.3

从表5的实验结果可以看出,三种输入的 acc 值均高于92%,表明了改进的模型对不同输入特征都有较好的预测性能。当输入特征为2D和3D时,模型的 acc 值最高,比仅含输入为2D分子特征时高0.6%,比仅含输入为3D分子特征时高3.6%;同时 are 值和 dre 值均高于输入仅含2D分子特征和仅含3D分子特征的预测值,这表明2D和3D特征组合给模型带来了新的信息,提高了模型的分类性能。另外2D加3D的最好结果低于分子指纹的最好结果,验证了分子指纹适用于药物透血脑屏障的预测。

3.2.4　PCA 初始化对模型收敛速度的影响

为验证权值初始化方法对模型收敛速度的影响,本文分别对 MSDAE 模型采用 PCA 权值初始化和随机权值初始化方法。在表2的参数设置下,对两种初始化方法进行对比实验,记录每20次的迭代结果,如图2所示。

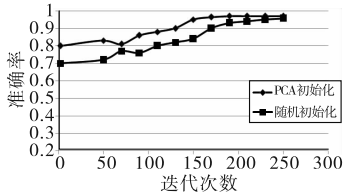

图2　权值初始化方法对比研究

从图2结果可以看出,采用 PCA 方法对模型权值进行初始化,相对于随机初始化方法,在迭代次数0~50时,准确率明显高于随机初始化方法;随着迭代次数的增加,PCA 方法在迭代到200次时收敛,而随机初始化方法则在250次时开始收敛。图2的结果表明,采用 PCA 初始化模型的权值,能有效促进模型的收敛。

4　结束语

药物透血脑屏障的穿透性对新药的研发至关重要,现有的方法主要基于临床实验、逻辑回归、SVM 或决策树等浅层的机器学习方法。这些方法存在计算量、只能提取化合物的浅层分子特征等问题,使得总体预测精度较差。

针对以上不足,本文提出一种改进的栈式降噪自编码器药物透血脑屏障预测方法。与以往研究方法相比,本文利用深度学习方法无监督提取药物化合物分子中的深层结构信息;通过引入 PCA 非监督训练预初始化权值,促进模型快速收敛;同时构造两层隐藏层的 DAE,提高模型提取抽象特征的泛化能力;最后融合两个隐藏层的化合物分子特征,从而更加准确地反映化合物分子信息。实验结果表明,改进后的模型预测准确率高于传统的栈式降噪自编码器以及 SVM 等浅层的机器学习方法。

参考文献:

[1] Hardman J G, Limbird L E, Gliman A G. Goodman and Gilman's the pharmacological basis of therapeutics[J]. Journal of Medicinal Chemistry,2002,45(6):1392-1393.

[2] Raevsky O A, Solodova S L, Lagunin A A, *et al*. Computer modeling of blood brain barrier permeability for physiologically active compounds[J]. Biochemistry Supplement,2014,7(2):95-107.

[3] 黄斌. 基于支持向量学习机预测药物透血脑屏障的活性[J]. 计算机与应用化学,2009,26(2):188-190.

[4] Yan A, Liang H, Chong Y, *et al*. In silico prediction of blood-brain barrier permeability[J]. SAR and QSAR in Environmental Research,2013,24(1):61-74.

[5] Zhao Y H, Abraham M H, Ibrahim A, *et al*. Predicting penetration across the blood-brain barrier from simple descriptors and fragmentation schemes[J]. Journal of Chemical Information & Modeling, 2007,47(1):170-175.

[6] Vincent P, Larochelle H, Lajoie I, *et al*. Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion[J]. Journal of Machine Learning Research, 2010,11(6):3371-3408.

[7] Adenot M, Lahana R. Blood-brain barrier permeation models: discriminating between potential CNS and non-CNS drugs including P-glycoprotein substrates[J]. Journal of Chemical Information and Computer Sciences,2004,44(1):239-248.

[8] Franco P, Porta N, Holliday J D, *et al*. The use of 2D fingerprint methods to support the assessment of structural similarity in orphan drug legislation[J]. Journal of Cheminformatics,2014,6(1):1-10.

[9] Yap C. PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints[J]. Journal of Computational Chemistry,2011,32(7):1466-1474.

[10] Borade N S, Deshmukh R R, Ramu S. Face recognition using fusion of PCA and LDA: borda count approach[C]//Proc of the 24th Mediterranean Conference on Control and Automation. Piscataway, NJ: IEEE Press,2016:1164-1167.

[11] Shreyamsha-Kumar B K, Swamy M N S, Omair-Ahmad M. Weighted residual minimization in PCA subspace for visual tracking[C]//Proc of IEEE International Symposium on Circuits and Systems. Piscataway, NJ:IEEE Press,2016:986-989.

[12] Shaw L, Routray A. Statistical features extraction for multivariate pattern analysis in meditation EEG using PCA[C]//Proc of IEEE EMBS International Student Conference. Piscataway, NJ: IEEE Press, 2016: 1-4.

[13] Sridharamurthy S K, Sudarshana-Reddy H R. PCA based feature vector for handwritten Kannada characters recognition[C]//Proc of International Conference on Emerging Research in Electronics, Computer Science and Technology. Piscataway, NJ:IEEE Press,2015:423-428.

[14] Bengio Y. Learning deep architectures for AI[J]. Foundations and Trends in Machine Learning,2009,2(1):1-127.

[15] 刘建伟,刘媛,罗雄麟. 深度学习研究与进展[J]. 计算机应用研究,2014,31(7):1921-1930,1942.

[16] Xing Chen, Ma Li, Yang Xiaoquan. Stacked denoise autoencoder based feature extraction and classification for hyperspectral image[J]. Journal of Sensors,2016,2016:Article ID 3632943.

[17] Ueda Y, Wang Longbiao, Kai A, *et al*. Environment-dependent denoising autoencoder for distant-talking speech recognition[J]. EURASIP Journal on Applied Signal Processing,2015,2015(1):1-11.

[18] Gu Feng, Flórez-Revuelta F, Monekosso D, *et al*. Marginalised stacked denoising autoencoders for robust representation of real-time multi-view action recognition[J]. Sensors, 2015, 15(7): 17209-17231.

[19] Baldi P, Brunak S, Chauvin Y, *et al*. Assessing the accuracy of prediction algorithms for classification: an overview[J]. Bioinformatics, 2000,16(5):412-424.

[20] Team T T D, Ai-Rfou R, Rami A, *et al*. Theano: a Python framework for fast computation of mathematical expression[EB/OL]. (2016-05-09). https://arxiv.org/abs/1605.02688.