

基于深度学习的口服生物利用度分类研究

史新宇¹ 禹 龙^{2,3,4} 田生伟^{1,3,4} 叶飞跃^{3,4} 钱 进^{3,4} 高双印¹

(新疆大学软件学院 乌鲁木齐 830008)¹ (新疆大学网络中心 乌鲁木齐 830046)²

(江苏理工学院计算机工程学院 常州 213000)³

(江苏理工学院云计算与智能信息处理重点实验室 常州 213000)⁴

摘 要 针对采用传统方法测量口服生物利用度(OB)代价昂贵、花费周期长,而现有的一些机器学习方法对其预测精度较低的问题,提出了一种基于栈式自编码(SAE)神经网络的口服生物利用度分类方法,利用经筛选过的分子特征结合栈式自编码模型对生物利用度进行分类。实验表明,与浅层机器学习模型支持向量机(SVM)以及人工神经网络(ANN)相比,深度网络对化合物分子的特征有更本质的学习,采用经筛选过的 2D 和 3D 分子特征组合对人体口服生物利用度的分类效果较好,其平均预测精度为 83%,灵敏度(SE)为 94%,特异性(SP)为 49%。

关键词 口服生物利用度,深度学习,分子描述符,栈式自编码,softmax 回归

中图分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.4.053

Research on Classification of Oral Bioavailability Based on Deep Learning

SHI Xin-yu¹ YU Long^{2,3,4} TIAN Sheng-wei^{1,3,4} YE Fei-yue^{3,4} QIAN Jin^{3,4} GAO Shuang-yin¹

(School of Software, Xinjiang University, Urumqi 830008, China)¹ (Network Center, Xinjiang University, Urumqi 830046, China)²

(College of Computer Engineering, Jiangsu University of Technology, Changzhou 213000, China)³

(Key Laboratory of Cloud Computing & Intelligent Information Processing of Changzhou,

Jiangsu University of Technology, Changzhou 213000, China)⁴

Abstract It is expensive and time-consuming to measure oral bioavailability using traditional methods, and existing machine learning methods show lower accuracy. To solve the problems, a classification method of human oral bioavailability based on stacked autoencoder(SAE) was presented. Filtered features of molecular are combined with the model of SAE to classify human oral bioavailability. Experimental results show that the deep network can study more essential features of molecules comparing with other shallow learning models like support vector machine and artificial neural network, and the combination of screened 2D and 3D molecular features achieves better classification effect of oral bioavailability, with a average accuracy value of 83%, a sensitivity(SE) value of 94% and a specificity(SP) value of 49%.

Keywords Oral bioavailability, Deep learning, Molecular descriptors, Stacked autoencoder, softmax regression

1 引言

口服给药方式简便,不直接损伤皮肤或粘膜,是患者最理想的选择之一^[1]。然而在口服药物的发现和开发过程中,约有 30% 的口服药物由于不良的药代动力学性质而失败,在这些药代动力学属性中,较低的生物利用度是阻碍药物进一步发展的主要原因^[2]。生物利用度指制剂中药物被吸收入人体循环的速度与程度,包括生物利用程度与生物利用速度^[3]。生物利用度受多种因素的影响,例如溶解性、亲脂性、胃肠道吸收、代谢与外排等,难以预测。采用传统的方法在人体内和体外测量药物的生物利用度是耗时和昂贵的^[4]。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口服生物利用度的预测是至关重要的。

近年来,基于浅层机器学习的口服生物利用度(Oral Bioavailability, OB)预测方法已得到了广泛应用。2008 年, Ma 等人采用结合遗传算法和共轭梯度以及支持向量机(Support Vector Machine, SVM)的算法构建了口服生物利用度预测模型,其总体预测精度可达到 80%,但对于生物利用度较低分子,预测精度仅为 25% 左右^[5]。2011 年,田盛等人用遗传函数逼近算法成功地建立了基于分子性质参数和结构片段指纹的多重线性回归模型,但生成的结构片段指纹并不能应用于所有的药物种类,这限制了模型对新的化合物的预测能力^[6]。2012 年, Ahmed 等人采用逻辑回归分类模型对口服生物利用度进行了预测,总的预测精度为 71%^[7]。

由于浅层的机器学习算法对复杂函数的表示能力不足,泛化能力受限,本文利用深度学习思想,提出了基于栈式自编

到稿日期:2015-05-19 返修日期:2015-09-26 本文受国家自然科学基金(31160341),自治区研究生科研创新项目(XJGRI2015034)资助。

史新宇(1990—),男,硕士生,主要研究方向为数据挖掘, E-mail: 1442714574@qq.com;禹 龙(1974—),女,硕士,教授,主要研究方向为计算机智能技术、计算机网络;田生伟(1973—),男,博士,教授,主要研究方向为计算机智能技术、自然语言处理;叶飞跃(1960—),男,博士,教授,主要研究方向为数据挖掘、智能信息系统;钱 进(1975—),男,博士,主要研究方向为粗糙集理论、数据挖掘、云计算;高双印(1990—),男,硕士生,主要研究方向为数据挖掘。

码(Stacked Autoencoder, SAE)神经网络的口服生物利用度分类模型,利用栈式自编码算法学习经筛选的 2D 和 3D 分子特征,通过 softmax 完成分类。通过与浅层机器学习模型 SVM 和人工神经网络(Artificial Neural Network, ANN)进行对比,验证了栈式自编码模型对口服生物利用度预测的有效性。

2 药物分子特征的生成与筛选

2.1 数据来源

为了成功地建立口服生物利用度预测模型,收集大量可靠和一致的实验数据至关重要。本文实验采用 ADME 团队收集整理的 1013 种药物和类药化合物的口服生物利用度数据。这些数据包含药物的名称、口服生物利用度以及部分分子特征等信息。数据库中所有药物分子全部以 SDF 和 SMILES 文件格式保存以便做进一步分析研究。

2.2 药物分子特征生成

药物分子的特征由分子模拟软件 Accelrys Discovery

Studio 生成。共生成了 355 个分子特征,这些分子特征覆盖了 12 种不同的类别,每个类别又包含多个分子特征。这些分子特征包括脂水分配系数、分子性质、分子表面积和体积、偶极、分子拓扑结构等。

2.3 药物分子特征的筛选

分子特征的选取是口服生物利用度预测的关键步骤。一些无意义的分子特征不仅与生物利用度无关,而且会增加模型的计算量和模型的运行时间。因此,为了避免分子特征的冗余,减少无关分子特征,需要对生成的分子特征进行筛选。

本文按以下规则对分子特征进行筛选:

- (1) 去除分子特征值大多数为零或者相同的分子特征;
- (2) 去除标准差值相对较小(标准差值 $<0.5\%$)的分子特征;
- (3) 去除彼此之间相关系数大于 90% 的分子特征^[8]。

通过上述筛选规则,最终选取了 168 个 2D 和 78 个 3D 分子特征。部分分子特征描述如图 1 所示。

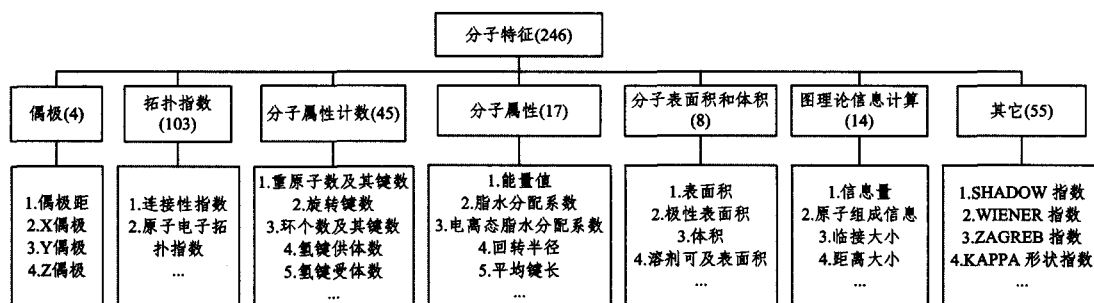


图 1 部分分子特征描述

3 构建口服生物利用度预测模型

3.1 栈式自编码神经网络

栈式自编码神经网络是深度学习中广泛采用的一种包含多个隐层的非线性无监督逐层贪心算法,从无标签数据集中提取高维复杂输入数据的分层特征,得到原始数据的分布式特征表示^[9]。假设一个无类别标签的训练样本集为 $x(x \in [0,1]^d)$,通过线性变换将输入 x 转换为激活值 $y(y \in [0,1]^d)$ 。 x 和 y 的关系如式(1)所示:

$$y = f_{\theta}(x) = s(Wx + b) \quad (1)$$

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

$$f(z) = \frac{1}{1 + \exp(-z)} \quad (2)$$

然后通过解码器变换式(2)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(3):

$$z = f_{\theta'}(y) = s(W'y + b') \quad (3)$$

其中, $\theta' = \{W', b'\}$ 为解码参数。式(1)、式(3)两个参数集合中的参数 W, W' 被限制满足关系 $W = W'$ 。最后通过不断修改 θ 和 θ' 的值得到最小化重构误差 J 。重构误差的表达式为:

$$J = \sum_{x \in D} L(x, z) \quad (4)$$

其中, L 为重构误差函数,选用平方误差函数公式:

$$L(x, z) = \|x - z\|^2 \quad (5)$$

3.2 softmax 回归

softmax 回归是逻辑回归的扩展形式,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^[10]。其基本思想是设训练样本及标签数据为 $\{(x^1, y^1), \dots, (x^n, y^n)\}$, x^i 为训练样本, $y^i \in \{1, \dots, n\}$ 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:

$$h_{\theta}(x^i) = \begin{bmatrix} p(y^i=1|x^i; \theta) \\ \vdots \\ p(y^i=k|x^i; \theta) \end{bmatrix} = \frac{1}{\sum_{j=1}^k e^{\theta_j^T x^i}} = \begin{bmatrix} e^{\theta_1^T x^i} \\ \vdots \\ e^{\theta_k^T x^i} \end{bmatrix} \quad (6)$$

其中, $\theta_1, \theta_2, \dots, \theta_k \in \mathbb{R}^{n+1}$ 为 $h_{\theta}(x^i)$ 的参数, $\frac{1}{\sum_{j=1}^k e^{\theta_j^T x^i}}$ 是对概率分布进行归一化,使得所有概率和为 1。

3.3 口服生物利用度的分类

首先,使用已获取的口服生物利用度数据计算其 2D 和 3D 分子特征的值,对其进行预处理。其次,根据其值的范围将数据分为阴性和阳性。然后,把数据集随机分为训练集和测试集,输入到栈式自编码模型。最后得出口服生物利用度的分类结果。具体分类过程如图 2 所示。

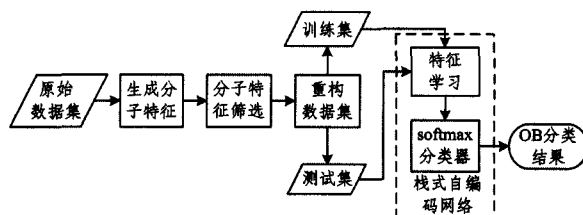


图 2 口服生物利用度分类流程

4 实验与分析

本文实验选取了 900 条数据用于训练和测试,其中随机选取了 720 条数据作为训练集,180 条数据作为测试集。这些数据根据以下规则分为阳性和阴性两种类别:如果化合物的生物利用度值大于等于 20%,则为阳性;否则为阴性^[5]。它们的分布比例如表 1 所列。

表 1 OB 类别分布

OB 类别	数据量	所占比例/%
阳性	243	27
阴性	657	73
总计	900	100

栈式自编码神经网络中可变参数对实验结果有一定的影响,例如迭代次数、权重衰减系数、稀疏惩罚项权重等。本文通过反复实验调整了不同的网络模型参数,得出了最优的参数组合。因此,本文后续实验结果均在参数组合最优的条件下得出。为了避免实验的随机性和偶然性对分类结果的影响,实验均采用 5 折交叉验证。

4.1 评估标准

由于在口服生物利用度模型构建过程中各个模型使用的方法和选取的化合物分子特征有所不同,需要通用的统计指标来评价模型的性能^[11]。

本文使用敏感度(Sensitivity,SE)即式(7)、特异性(Specificity,SP)即式(8)以及总的预测精度(Accuracy,Q)即式(9)来衡量基于栈式自编码神经网络的生物利用度预测的性能。

$$SE=TP/(TP+FN)$$
 (7)

$$SP=TN/(TN+FP)$$
 (8)

$$Q=(TP+TN)/(TP+TN+FP+FN)$$
 (9)

口服药物生物利用度高的预测正确率用敏感度表示,反之用特异性表示。通过统计分子的真阳性(TP)、假阳性(FP)、真阴性(TN)、假阴性(FN)的个数,利用式(7)和式(8)分别计算出敏感度和特异性。

4.2 自编码层数选择对分类结果的影响

栈式自编码神经网络的自编码(Autoencoder,AE)层数选择对实验结果有一定影响,选取合适的 AE 层数非常关键。实验选择分子特征维度为 246,采用 AE 层数从 1 层到 5 层递增的 SAE 进行口服生物利用度分类的对比分析实验。实验结果如表 2 所列。

表 2 不同 AE 层数的分类结果

AE 层数	阳性			阴性			Q/%
	TP	FN	SE/%	TN	FP	SP/%	
AE ¹	130	5	96.3	9	36	20	77.2
AE ²	129	6	96.0	12	33	27	78.0
AE ³	127	8	94.0	22	23	49	83.0
AE ⁴	128	7	94.8	8	37	17.8	75.6
AE ⁵	133	2	98.5	7	38	15.6	77.8

由表 2 实验结果可知,随着 AE 层数的增加,Q 值逐渐上升,当模型层数增加到 3 时,口服生物利用度的 Q 值最高,虽然 SE 值有所降低,比其它几层平均降低了 3.4%,但模型的 SP 值平均提升了 28.9%。这表明当自编码层数为 3 时,与其它几层相比,模型在保证阳性识别准确率的同时,对阴性的识别正确率有一定的提升。因此,在后续的实验中,均采用 3 层

的栈式自编码器对口服生物利用度进行分类。同时,从表 2 还可看出,由于数据集的阳性和阴性所占比例不均衡,阳性所占比例较大(占 73%),模型能够学习到更多的阳性特征,所以 SE 值较高。

4.3 不同输入特征对分类结果的影响

本文实验分别使用仅含 2D 分子特征、仅含 3D 分子特征以及分子的 2D 和 3D 特征作为输入,比较不同输入特征下栈式自编码神经网络对口服生物利用度的分类性能。实验结果如表 3 所列。

表 3 不同特征组合的分类结果

分子特征	阳性			阴性			Q/%
	TP	FN	SE/%	TN	FP	SP/%	
2D	128	7	94.8	9	36	20	76
3D	88	47	65.2	28	17	62	64
2D+3D	127	8	94.0	22	23	49	83

从表 3 的实验结果可以看出,当栈式自编码神经网络的输入特征为 2D+3D 时,模型的 Q 值最高,比仅含输入 2D 分子特征时模型分类的准确率高 7%,比仅含输入 3D 分子特征时的分类准确率高 19%。同时当仅含 3D 分子特征时,SP 值最高,且比仅含 2D 分子特征高出 42%。这表明仅含 3D 分子特征的模型对阴性的识别能力强于仅含 2D 分子特征的模型。因此融合 2D 和 3D 分子特征作为模型的输入对口服生物利用度分类效果较好,对 SP 值的提升有较大的影响。

4.4 对比实验

为验证栈式自编码神经网络对口服生物利用度分类的有效性,本实验在相同数据集上比较栈式自编码神经网络与浅层机器学习模型 ANN 和 SVM 的分类准确率。实验结果如图 3 所示。

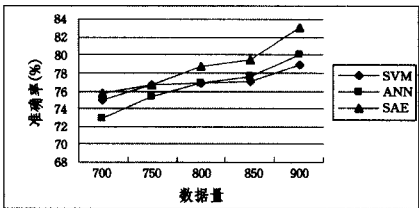


图 3 SAE 与 SVM 和 ANN 的分类准确率

从图 3 可以看出,随着数据量的增加,SAE 与 ANN 和 SVM 对口服生物利用度的分类准确率均有所提升。基于 SAE 的口服生物利用度分类模型的准确率比 ANN 平均提高了 2.2%,比 SVM 平均提高了 1.8%。这是因为 SAE 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。因此,实验验证了基于栈式自编码神经网络的口服生物利用度分类模型要优于传统的 ANN 和 SVM 模型。

结束语 分子特征的选取是口服生物利用度预测的关键步骤,经筛选后的分子特征作为栈式自编码神经网络的输入,较浅层机器学习对分子特征有更本质的学习能力;同时融合 2D 和 3D 分子特征后模型对口服生物利用度的分类精度较高。未来的工作是收集更多的可靠的分子数据,获取更精确的分子特征,以进一步提高口服生物利用度的分类精度。

参考文献

- [1] Tian Sheng, Li You-yong, Wang Jun-mei, et al. ADME Evaluation in Drug Discovery. 9. Prediction of Oral Bioavailability in Humans Based on Molecular Properties and Structural Fingerprints[J]. Molecular Pharmaceutics, 2011, 8(3): 841-851
 - [2] Hou Ting-jun, Wang Jun-mei, Zhang Wei, et al. ADME Evaluation in Drug Discovery. 6. Can Oral Bioavailability in Humans Be Effectively Predicted by Simple Molecular Property-Based Rules? [J]. Journal of Chemical Information and Modeling, 2007, 47(2): 460-463
 - [3] Lai Jun, Liao Zheng-gen, Yang Ming-fu, et al. Research and progress of bioavailability[J]. Chinese Journal of Experimental Traditional Medical Formulae, 2010, 18: 226-229 (in Chinese)
赖珺, 廖正根, 杨明福, 等. 生物利用度研究进展[J]. 中国实验方剂学杂志, 2010, 18: 226-229
 - [4] Wang Jun-mei, Hou Ting-jun. Advances in computationally modeling human oral bioavailability[J]. Advanced Drug Delivery Reviews, 2015, 86(6): 11-16
 - [5] Ma Chang-ying, Yang Sheng-yong, Zhang Hui, et al. Prediction models of human plasma protein binding rate and oral bioavailability derived using GA-CG-SVM method[J]. Journal of Pharmaceutical and Biomedical Analysis, 2008, 47: 677-682
 - [6] Tian Sheng. Theoretical studies on the prediction of drug-likeness and oral bioavailability[D]. Suzhou: Soochow University, 2011 (in Chinese)
田盛. 类药性和生物利用度的理论预测研究[D]. 苏州: 苏州大学, 2011
 - [7] Ahmed S S S J, Ramakrishnan V. Systems Biological Approach of Molecular Descriptors Connectivity: Optimal Descriptors for Oral Bioavailability Prediction[J]. Plos One, 2012, 7(7): e40654
 - [8] Zhang Hui, Chen Qing-yi, Xiang Ming-li, et al. In silico prediction of mitochondrial toxicity by using GA-CG-SVM approach [J]. Toxicology in Vitro, 2009, 23(1): 134-140
 - [9] Liu Jian-wei, Liu Yuan, Luo Xiong-lin. Research and development on deep learning[J]. Application Research of Computers, 2014, 31(7): 128-134 (in Chinese)
刘建伟, 刘媛, 罗雄麟. 深度学习研究进展[J]. 计算机应用研究, 2014, 31(7): 128-134
 - [10] Wang Yong, Zhao Jian-hui, Zhang Deng-yi, et al. Forest fire image classification based on deep neural network of sparse autoencoder [J]. Computer Engineering and Applications, 2014, 50(24): 173-177 (in Chinese)
王勇, 赵俭辉, 章登义, 等. 基于稀疏自编码深度神经网络的林火图像分类[J]. 计算机工程与应用, 2014, 50(24): 173-177
 - [11] Kim M T, Sedykhy A, Suman K, et al. Critical Evaluation of Human Oral Bioavailability for Pharmaceutical Drugs by Using Various Cheminformatics Approaches[J]. Pharm Res, 2014, 31: 1002-1014
-
- (上接第 255 页)
- 申国伟, 杨武, 王巍, 等. 面向大规模微博消息流的突发话题检测 [J]. 计算机研究与发展, 2015, 52(2): 512-521
 - [17] Holzinger A. Social Media Mining and Social Network Analysis: Emerging Research [J]. Online Information Review, 2014, 38 (1): 157-158
 - [18] Wang Li, Cheng Xue-qi. Dynamic Community in Online Social Networks[J]. Chinese Journal of Computers, 2015, 38(2): 219-237 (in Chinese)
王莉, 程学旗. 在线社会网络的动态社区发现及演化[J]. 计算机学报, 2015, 38(2): 219-237
 - [19] Zhou D, Huang J, Schölkopf B. Learning with hypergraphs: Clustering, classification, and embedding [M] // Advances in Neural Information Processing Systems, 2006: 1601-1608
 - [20] Sang J, Xu C, Liu J. User-aware image tag refinement via ternary semantic analysis [J]. IEEE Transactions on Multimedia, 2012, 14(3): 883-895
 - [21] Bi J, Huang J, Qin Z. A Relationship Strength-Aware Topic Model for Communities Discovery in Online Social Networks [M] // Advances in Computer Science and its Applications. Springer Berlin Heidelberg, 2014: 709-715
 - [22] Diakopoulos N, De Choudhury M, Naaman M. Finding and assessing social media information sources in the context of journalism[C] // Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, 2012: 2451-2460
 - [23] Li Huan-yu, Bi Du-yan, Yang Yuan, et al. Research on Visual Tracking Algorithm Based on Deep Feature Expression and Learning[J]. Journal of Electronics & Information Technology, 2015, 37(9): 2033-2039 (in Chinese)
李寰宇, 毕笃彦, 杨源, 等. 基于深度特征表达与学习的视觉跟踪算法研究[J]. 电子与信息学报, 2015, 37(9): 2033-2039
 - [24] Chen Shuo. Application research of deep learning neural networks in speech recognition[D]. Guangzhou: South China University of Technology, 2013
陈硕. 深度学习神经网络在语音识别中的应用研究[D]. 广州: 华南理工大学, 2013
 - [25] Deng L, Li J, Huang J T, et al. Recent advances in deep learning for speech research at Microsoft[C] // 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2013: 8604-8608
 - [26] He X, Gao J, Deng L. Deep learning for natural language processing and related applications (Tutorial at ICASSP) [C] // IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). 2014
 - [27] Huiskes M J, Lew M S. The MIR Flickr retrieval evaluation [C] // Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval. ACM, 2008: 39-43
 - [28] Ermiş B, Acar E, Cemgil A T. Link prediction in heterogeneous data via generalized coupled tensor factorization[J]. Data Mining and Knowledge Discovery, 2015, 29(1): 203-236
 - [29] Popescu A, Ungar L H. Structural logistic regression for link analysis[J/OL]. http://repository.upenn.edu/cgi/viewcontent.cgi?article=1139&context=cis_papers
 - [30] Lü L, Zhou T. Link prediction in complex networks: A survey [J]. Physica A: Statistical Mechanics and its Applications, 2011, 390(6): 1150-1170