计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200700068

基于图卷积神经网络的药物靶标作用关系预测方法

高 创[1] 李建华[1,2] 季秀怡[1] 朱程龙[1] 李诗良[2] 李洪林[2]
1 华东理工大学信息科学与工程学院 上海 200237
2 上海市新药设计重点实验室 上海 200237
(gaochuang0814@163.com)

摘 要 药物-靶标作用关系预测在药物研发以及药物重定位中扮演着重要角色,但现有的机器学习方法在正负样本高度不平衡的数据上仍存在预测能力不足的问题。为此,提出一种基于图卷积神经网络的药物靶标作用关系预测方法。该方法首先构造一个结合多种药物(靶标)相关信息的异质信息网络,然后采用图卷积神经网络在此异质信息网络上学习得到能精确表达每个节点拓扑特征及邻居特征信息的低维向量表征,最后利用这些向量信息通过向量空间投影预测节点间概率的评分。在DrugBank_FDA和Yammanishi_08数据集上进行的药物-靶标作用关系预测的对比实验中,所提方法的AUPR(Area Under the Precision-Recall Curve)值都优于其他4种方法,并且在较大型数据集上也有较好的表现。实验结果表明,所提方法提高了样本高度不平衡时的药物-靶标作用关系预测性能;同时在生物药物数据库上的实验也验证了所提方法所发现的未知药物-靶标作用关系的有效性。

关键词:图卷积神经网络;药物-靶标作用关系;异质信息网络;机器学习;向量表征
中图法分类号 TP391

Drug Target Interaction Prediction Method Based on Graph Convolutional Neural Network

GAO Chuang[1],LI Jian-hua[1,2],JI Xiu-yi[1],ZHU Cheng-long[1],LI Shi-liang[2] and LI Hong-lin[2]
1 College of Information Science and Engineering,East China University of Science and Technology,Shanghai 200237,China
2 Shanghai Key Laboratory of New Drug Design,Shanghai 200237,China

Abstract Drug-target interaction prediction plays an important role in drug discovery and repositioning. However,existing prediction methods have the problem of insufficient predictive performance while processing data with highly unbalance positive and negative samples. Therefore,a novel computational method based on graph convolutional neural network(GCN) for predicting drug-target interactions is proposed. In this method,a heterogeneous information network is constructed,which integrates diverse drug-related information and target-related information. From the heterogeneous information network,low-dimensional vector representation of features,which accurately explains the topological properties of individual and neighborhood feature information,is learned by using GCN and then prediction is made based on these representations via a vector space projection scheme. The AUPR(Area Under the Precision-Recall Curve) values of the proposed method outperforms other four existing methods in the prediction of drug-target interaction on both DrugBank_FDA and Yammanishi_08 datasets,and it preforms well on bigger datasets. The experimental results indicate that the proposed method improves the prediction performance of drug-target interaction on datasets with highly unbalanced samples. Furthermore,we validate novel(unknown) drug-target interactions which are predicted by GCN in biomedical databases.

Keywords Graph convolutional neural networks,Drug-target interactions,Heterogeneous information network,Machine learning,Vector representation

1 引言

预测药物和靶标的作用关系(Drug-Target Interactions,DTI)是新药研发与药物重定位中非常重要的一步。新药研发是为某种蛋白靶标寻找合适的新药物;药物重定位是根据现有的药物进行重新定位分析,以寻找药物的新用途。相比新药研发,药物重定位有降低成本以及加快发现新药的优点,并逐渐成为药物发现中一种比较流行的策略[1]。但新药研发

到稿日期:2020-07-10 返修日期:2020-08-16 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:国家重点研发计划项目(2016YFA0502304);国家重大新药创制项目(2018ZX09735002)
This work was supported by the National Key R&D Program of China(2016YFA0502304) and National Major Scientific and Technological Special Project for "Significant New Drugs Development"(2018ZX09735002).
通信作者:李建华(jhli@ecust.edu.cn)

与药物重定位都要建立在 DTI 预测的基础上，即针对新药物发现其对应的靶标，针对新靶标发现其对应的药物[2]。

通过计算机进行 DTI 预测避免了烦琐的药物实验，现已成为药物靶标预测的主流方法。目前 DTI 预测的方法可分为传统模拟方法与机器学习方法。传统预测 DTI 的方法主要分为两类：1)基于分子对接模拟[3-4]的方法；2)基于配体的方法[5]。但是，对接模拟方法特别耗时，且在部分蛋白靶标的结构信息未知时无法进行对接，而靶标已知配体较少时，基于配体方法的预测能力也随之下降[6]。现有的机器学习方法大多采用支持向量机、贝叶斯、矩阵分解以及随机森林等预测 DTI，但是上述机器学习方法在正样本非常稀疏以及较大型的数据集上仍存在表现能力不足的问题。

尽管卷积神经网络与循环神经网络在图像与文本方面带来了性能提升，但是它们无法解决机器学习方法预测 DTI 时存在的问题，因为平移不变性使得传统神经网络无法处理图结构的数据[7]。新出现的图卷积神经网络[8]扩展了神经网络在图结构数据表征方面的方法，在图结构上具有根据自身特征、邻居特征信息以及拓扑信息更新节点的特点。这种在图结构上降维的计算能力恰好满足药物靶标预测中根据相似靶标(药物)节点预测相似药物(靶标)的计算需求。因此，本文用图卷积神经网络的方法预测 DTI。

针对目前机器学习方法存在的问题以及图卷积神经网络在图结构上优势，本文提出了一种基于图卷积神经网络[9]的方法(Graph Convolutional Neural Networks Drug-Target Interactions，GCNDTI)来预测 DTI。本文的主要贡献如下：

(1)首次在异质信息网络上提出使用图卷积神经网络来预测 DTI，GCNDTI 所使用的信息不仅包括当前节点与其邻居特征信息，还包括其他模型较少使用的节点拓扑结构信息。

(2)现有基于异质网络预测 DTI 的机器学习方法一般是根据单一权重接收邻居特征信息，而本文方法则根据节点和节点间的关系类型来分配不同权重参数以进行特征信息提取，并通过迭代训练来优化作用关系边的权值。

(3)大量的实验证明，在 AUPR 标准下，该方法在高度不平衡数据集中的表现力优秀，在较大型的数据集上与预测未知 DTI 时，皆有良好的表现力。

2 相关工作

2.1 药物靶标关系预测研究现状

随着机器学习的快速发展，机器学习的方法避免了烦琐的药物实验，并已成为药物靶标关系预测的主流方法。现有的 DTI 预测大都是基于“相似的药物有着相似的靶标、相似的靶标有着相似的药物”的理念[2]。例如，Bleakley 等[10]把药物靶标预测转变为一个二分类任务，首次提出基于二分局部模型使用支持向量机从药物和靶标两个方面预测 DTI。Mei 等[11]进一步扩展了二分局部模型，利用药物与靶标的相似性信息对药物与靶标进行特征填充，使其不仅可以在二分局部网络上预测 DTI，对于新药物、新靶标也可以预测其相对应的作用关系。Xia 等[12]提出通过一种半监督方法(NetLapRLs)预测 DTI，该方法利用从药物-靶标关系网络中构建的核来提高预测性能。

随着药物基因表达相似性、药物-疾病关系、药物化合物结构、蛋白基因序列、蛋白基因表达相似性等信息的增多，异质信息网络为提高 DTI 的预测能力提供了新的机遇，一些更加高效的算法孕育而出。Liu 等[13]引入药物-药物、靶标-靶标相似度来推断药物靶标关系，把药物和靶标相似矩阵转化为隐层特征向量，在低维空间中找寻节点的关系。其优势在于，即使药物靶标的相互作用有限，也可以预测新药物和新靶标。Hao 等[14]使用经矩阵分解的低维矩阵与相似信息矩阵构建药物和靶标网络图，根据融合两网络图后的信息预测 DTI。Olayan 等[15]在异质图上通过随机游走构造最大与均值两种类型的矩阵，结合多种相似性矩阵预测药物靶标关系。Mohamed 等[16]利用 Trimodel 模型在药物和靶标构建的知识图谱上得到所有药物和靶标信息的低维向量表征，使用向量表征预测 DTI。由于引入了多种药物(靶标)相似性，与以往同质信息网络上的 DTI 预测方法相比，异质信息网络上的 DTI 预测方法在标准数据集上测试得到的准确率与速率都有显著的提升。但是，现有的 DTI 预测方法还存在一些尚待解决的问题，譬如对正负样本高度不平衡的大型数据集进行预测时精确度较低，以及在较大型数据集上的表现力较弱等问题，解决这些问题需要引入新的方法(算法)。

2.2 图卷积神经网络模型

图卷积神经网络模型填补了目前神经网络无法处理非欧氏结构的空白，其利用图结构、节点特征和边特征学习图中节点的低维表征向量 $\boldsymbol{h}_v$ 以及整个图的表征向量 $\boldsymbol{h}_G$[17]。现有的图卷积神经网络模型大多是消息传递网络(Message Passing Neural Networks，MPNNs)架构，这种架构主要包括两阶段，即消息传递阶段与图整体表征向量读出阶段，其中消息传递阶段也常被称为聚合过程[18]。消息传递阶段中较为流行的方法是基于邻居的聚合方法，这种策略通过多次迭代聚合邻居节点与边的表征信息来不断更新节点的表征向量。图卷积网络用层数表示迭代次数，随着迭代次数的增加，聚合的节点邻居信息就越多。k 次迭代能够聚合到距离当前节点 k 个跳步内的邻居信息[19]，此时图卷积网络的层数为 k。由于第 k 层的信息是由第 $k-1$ 层聚合而来，因此图卷积神经网络的第 k 层聚合可表示为：

$$\boldsymbol{h}_v^{(k)}=COMBINE^{(k)}(AGGREGATE(\boldsymbol{h}_u^{(k-1)},\forall \boldsymbol{u}\in N(\boldsymbol{v})),\boldsymbol{h}_v^{(k-1)}) \tag{1}$$

$$\boldsymbol{h}_G=READOUT(\{\boldsymbol{h}_v^{(k)}\mid \boldsymbol{v}\in \boldsymbol{V}\}) \tag{2}$$

其中，$\boldsymbol{h}_v^{(k)}$ 表示第 k 层节点 v 的特征信息，$\boldsymbol{N}(v)$表示节点的邻居节点，$\boldsymbol{h}_G$ 为全图的特征表示方法，$\boldsymbol{V}$ 表示边的集合。

聚合阶段是图卷积神经网络的重要步骤，可根据聚合方法的不同划分形成不同类型的图卷积神经网络。图卷积网络可根据聚合方法分为谱方法与空间方法两类。第一个谱卷积神经网络是 Bruna 等[20]基于图谱理论提出的。随后 Defferrard 等[21]采用切比雪夫方法对卷积核进行参数化并提出切比雪夫网络(ChebNet)。Kipf 等[9]对 ChebNet 进行了简化并对权重矩阵做归一化处理，提出了 GCN。

谱卷积神经网络虽存在聚合的信息来源不一定是邻域节点的问题，但 ChebNet 与 GCN 已经开始从空间域考虑并定

义节点权重矩阵，为研究人员考虑在节点域定义聚合函数奠定了基础。为了解决 GCN 在大型图中计算拉普拉斯矩阵费时的问题，GraphSAGE[17]在节点邻域内随机采样并使用可训练的聚合函数替代拉普拉斯矩阵。FastGCN[22]进一步改善了采样算法，不再局限于节点邻域内的样本，而是直接对每一层的感受野进行随机采样。聚合函数的不同也形成了不同类型的图卷积神经网络。目前图卷积神经网络的聚合函数可自适应于任务和具体的图结构，使得此种网络类型不仅可以应用于交通网络与生物网络，还可以应用于社交网络、推荐系统以及反欺诈等方面。

3　GCNDTI 方法

本文提出的 GCNDTI 方法由编码器、解码器与训练优化 3 部分构成。编码器以节点与邻居节点信息以及局部拓扑结构为输入，输出此节点的低维向量表征。解码器根据药物节点和靶标节点的相关低维信息，由向量空间投影得到节点间的概率评分。训练优化通过预测值与真实值之间的差值优化模型参数。图 1 给出了 GCNDTI 方法的示意图，其中图 1(a)给出了 GCNDTI 所使用的异质信息网络实例，图 1(b)—图 1(d)分别为编码器、解码器与训练优化的示意图。

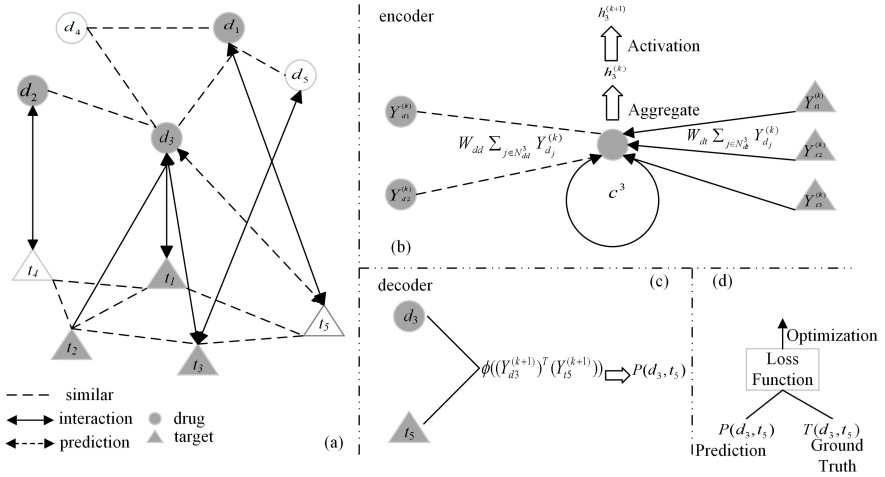


图 1　GCNDTI 流程图

Fig. 1　Flow chart of GCNDTI

3.1　编码嵌入

在大规模的图结构中，节点的低维向量嵌入常用于图分析以及预测问题的特征输入，因为在低维向量上的数据处理比在高维向量上更加简单有效[16]。因此，GCNDTI 方法编码器的主要任务是聚合邻居信息并获得节点低维信息。

在聚合过程中，通过对当前节点的自身特征信息及其邻居节点的特征信息进行卷积操作来获得下一层的节点特征信息。图卷积神经网中第 k 层的节点聚合流程如图 1(b)所示，聚合公式如下：

$$\boldsymbol{h}_v^{(k)}=\sum_r\sum_{j\in N_r^i}c_r^{ij}\boldsymbol{W}_r^{(k)}\boldsymbol{Y}_j^{(k)}+c_r^i\boldsymbol{Y}_i^{(k)} \tag{3}$$

$$\boldsymbol{Y}_i^{(k+1)}=\phi(\boldsymbol{h}_i^{(k)}) \tag{4}$$

其中，$\boldsymbol{W}_i^{(k)}$ 为第 k 层关系 r 类型下的权重参数，$\boldsymbol{h}_i^{(k)}$ 为节点 i 在第 k 层神经网络的隐藏层状态，c_r^{ij} 与 c_r^i 为归一化常数，$c_r^{ij}=1/\sqrt{|N_r^i||N_r^j|}$，$c_r^i=1/|N_r^i|$，$N_r^i$ 为在 r 关系下节点 v_i 的邻居集合，ϕ 为非线性激活函数(选用 Rectified Linear Unit(ReLU)函数)，$\boldsymbol{Y}_i^{(k)}$ 与 $\boldsymbol{Y}_i^{(k+1)}$ 分别为更新前后的节点特征。式(4)中，当 $k=0$ 时，$Y_i^{(0)}=\boldsymbol{X}_i$，$\boldsymbol{X}_i$ 为节点 $i(i\in E)$的初始特征信息，E 为节点的集合。若节点没有特征信息，则可以使用 one-hot 形式表示特征信息[23]。

编码过程可根据边类型的不同赋予其不同的权值，以聚合邻居节点信息。边类型 r 不同，权重参数 $\boldsymbol{W}_r$ 也就不同，这是因为异质信息网络中节点间关系(有向边)表示的实际意义不同。这些关系可利用“起点节点类型-终点节点类型”信息对表示为“药物-药物”“药物-靶标”“靶标-药物”“靶标-靶标”4 种，因此每层都有 4 种权重参数，文中分别用 $\boldsymbol{W}_{dd}$，$\boldsymbol{W}_{dt}$，$\boldsymbol{W}_{td}$，$\boldsymbol{W}_{tt}$ 表示。在聚合时，对于相同类型的关系使用同一权重参数进行聚合，不同类型的关系分别使用各自类型相关的权重参数进行聚合。例如，在聚合过程中，如果当前节点为“药物”类型节点，则该节点与聚合的邻居节点的关系包括“药物-药物”以及“药物-靶标”等。对于图 1(b)中的节点 d_3，其聚合计算包括距离其 k 跳步的邻居节点与当前节点信息(k 表示图卷积神经网络叠加的层数)。

3.2　解码

药物-靶标关系预测任务是一个关系(边)缺失预测问题，而解码步骤就是尝试构造节点间缺失的边，解码过程如图 1(c)所示。本文使用向量空间投影方法作为解码器，采用两个已编码的节点特征信息作为输入，解码后输出两节点间的预测评分 P。若计算节点 v_i 和节点 v_j 的评分 P^{ij}，则解码的计算式如下：

$$P^{ij}=\sigma(\boldsymbol{Y}_i^{\mathrm{T}}\boldsymbol{Y}_j) \tag{5}$$

其中，$\boldsymbol{Y}_i$ 为节点 i 嵌入后的低维信息，$\boldsymbol{Y}_j$ 为节点 j 嵌入后的低维信息，σ 为激活函数。σ 函数一般选用 $Sigmoid$ 函数，可使输出的评分结果值在(0,1)区间。

3.3　训练优化

训练优化过程是为了优化图卷积神经网络中的权值参数，使最终预测值接近于标注数据(GroundTruth)。需要优化的权值参数指图卷积神经网络每层的权值参数，即每层的 $\boldsymbol{W}_{dd}$，$\boldsymbol{W}_{dt}$，$\boldsymbol{W}_{td}$，$\boldsymbol{W}_{tt}$。为了使解码得到的预测值趋近于标注数

据，本文引入负样本并使用交叉熵函数作为损失函数来不断优化每层的4个权值参数。

在GCNDTI中，引入负样本(未知DTI)的目的是对模型参数进行调整，使异质信息网络图中正样本(已知DTI)的概率评分尽可能高，同时降低负样本的概率评分。本文选取了与验证集合中正样本数量相同的负样本。负样本的选取方法是随机选取与节点 v_i 不具有作用关系的节点 v_m 构造负样本边 (v_i, v_m)，其中 v_m 的取样满足分布 p[24]。构造负样本后，分别标注正负样本数据为1和0，然后使用交叉熵函数优化训练的权值参数，使正负样本趋近于标注数据。参数优化过程如图1(d)所示，具体表达式如下：

$$L_r^{(i,j)} = -\log(P_r^{(i,j)}) - \mathbb{E}_{m\sim p_r(j)} \log(1-P_r^{(i,m)}) \tag{6}$$

$$L = \sum_{(i,r,j)\in\mathbb{R}} L_r^{(i,j)} \tag{7}$$

其中，$P_r^{i,j}$ 为GCNDTI方法正样本 (v_i, v_j) 的预测值，$P_r^{i,m}$ 为此方法负样本 (v_i, v_m) 的预测值，$\mathbb{R}$ 表示所有存在的边，L 为所有节点集合的损失值。

4 实验结果与分析

本文在Yamanishi_08和DrugBank_FDA数据集上采用AUPR(Area Under the Precision-Recall Curve)值对比了GCNDTI方法与其他方法在药物靶标任务中的优劣，并验证了GCNDTI方法发现的新的未知药物-靶标关系。

4.1 包含药物与靶标信息的异质信息网络

药物与靶标有关信息的增多为各种机器学习方法在异质信息网络中预测DTI提供了数据基础。目前异质信息网络常用于药物靶标预测任务的输入，DTINet[6]，DDR[15]与NeoDTI[25]等方法都是通过提取异质信息网络中节点或者边的特征信息来提高预测精度。异质信息网络也非常适用于图卷积神经网络模型，由于图卷积神经网络不具有平移不变性的缺陷，因此可以在每个节点上聚合信息，忽略节点输入顺序的特点。

表1 5个数据集的统计信息

Table 1 Summary infomations of the five datasets

Dataset	NR	GPCR	IC	E	DB_FDA
Drug	54	223	210	445	1482
Target	26	96	204	664	1408
DTI	90	635	1476	2926	9881
A_d	1.67	2.85	7.03	6.58	6.67
A_t	3.46	6.68	7.24	4.41	7.02
P2N/%	6.67	3.03	3.57	1.00	0.48

注：A_d 表示每个药物平均作用的靶标数目；A_t 表示每个靶标平均作用的药物数目；P2N表示数据集中正负样本的比率

本文所用异质信息网络由药物相似矩阵、靶标相似矩阵和药物-靶标作用关系矩阵构成，其中药物-靶标作用关系信息来自Yamanishi_08[26]和DrugBank_FDA数据集。这两个数据集都是药物-靶标预测方法进行评估时经常采用的数据集。Yamanishi_08数据集由KEGGBRITE[27]，BREBDA[28]，SuperTarget[29]以及DrugBank[30]数据库的已知DTI构成，根据蛋白靶标类别的不同可将该数据集分为4个类别的子数据集，分别是酶(Enzymes，E)、核受体(Nuclear Receptors，NR)、离子通道(Ion Channels，IC)和G蛋白偶联受体(G Protein-Couple Receptors，GPCR)。DrugBank_FDA数据集是DrugBank数据集(5.0.3版本)中由美国食品药物管理局(Foodand Drug Administration，FDA)批准药物的数据子集。5个数据集的详细统计信息如表1所列。

异质信息网络中的药物与靶标相似信息来自Olayan等[15]构造的信息矩阵。药物与靶标相似信息是由多种相似信息经过相似筛选与融合得到的。Olayan等将药物副作用、药物基因关系、蛋白基因序列、靶标基因关系等信息经过相似筛选与相似网络融合(Similarity Network Fusion，SNF)[31]后得到含较少噪声的药物和靶标相似信息。筛选与融合流程如图2所示。

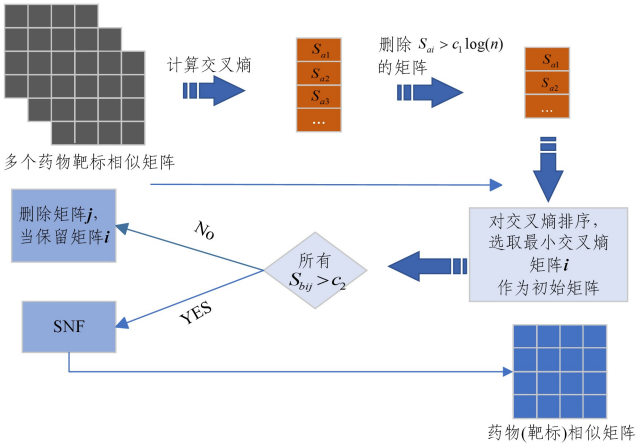


图2 相似筛选与融合过程

Fig. 2 Process of similarity selection and fusion

4.2 评价类型与标准

在目前的研究中，研究人员所关注的信息已经不仅仅是简单地对部分未知作用关系的预测，还包括对新药物以及新靶标作用关系的预测。因此，本文在3种预测类型下通过实验对比GCNDTI与其他方法的预测能力，3类预测类型分别为：1)新药物预测类型(S_d)，即只保留该药物和其他药物之间的相似性信息，预测该药物和所有靶标的作用关系；2)新靶标预测类型(S_t)，即只保留该靶标和其他靶标之间的相似性信息，预测该靶标与所有药物的作用关系；3)未知部分DTI(S_i)，即掩盖(抹去)部分药物靶标作用关系，预测所掩盖的作用关系信息。

自Mei等[11]提出将AUC(Area Under the ROC Curve)和AUPR作为药物靶标模型的评价指标后，其就被多种药物靶标预测模型所采用[32]。这两种评价标准中，AUC适用于各类中正负样本相对平衡的数据，而AUPR能够度量样本结果相关性和获得真正相关(True Positive)结果的数量，具有在不平衡的样本中作为评价指标的能力。文献[12]也证实了在正负样本数据高度不平衡的实验中AUPR比AUC更加敏感。分析本文的实验数据集，本文实验所用的Yamanishi_08数据集正负样本比全部低于1:9，DrugBank_FDA中甚至低于1:100，因此皆属于高度不平衡数据集。两数据集样本的具体信息如表1中的P2N(正负样本比值)所示。基于以上分析，本文以AUPR作为评价指标对GCNDTI方法与其他方法进行比较和分析。

本文在Yamanishi_08与DrugBank_FDA数据集上对GCNDTI方法进行了评估，并与TriModel[16]，DDR[15]，NRLMF[11]和BLM-NII[10]这4种方法进行了对比。表2列出了

所有方法在 AUC 与 AUPR 标准下的结果，分别按预测类型标记为 S_d，S_t，S_i。

4.3　实验设置

实验数据集来源于 Yamanishi_08 与 DrugBank_FDA，一共包括了 2414 种药物、2398 种靶标、15008 条作用关系边(DTI)。同时，利用药物与靶标相似信息矩阵[15]辅助模型训练。

在实验过程中，同样采用 TriModel[16]，DDR[15]等模型交叉验证的方法对模型进行训练和评估。对于不同预测类型，GCNDTI 方法所选用的训练集、验证集和测试集也不相同，具体如下：

(1)S_d 预测类型。训练集包含了 80%药物节点相连的边(DTI)，测试集与验证集分别只包含 10%药物节点相连的边(DTI)。

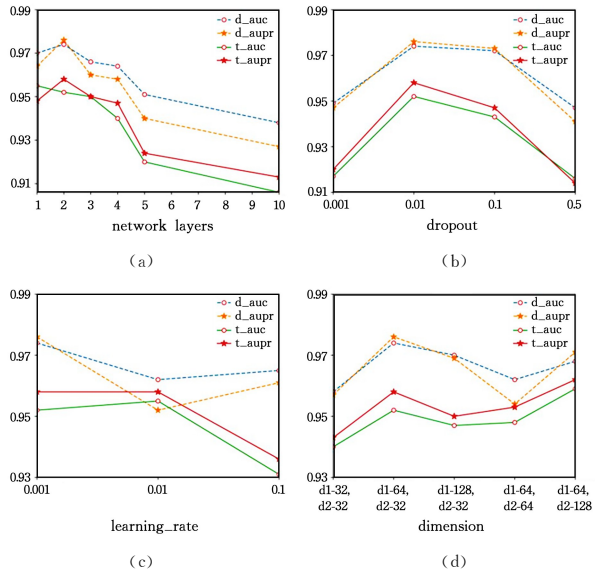
(2)S_t 预测类型。训练集包含了 80%靶标节点的边(DTI)，测试集与验证集分别只包含 10%靶标节点的边(DTI)。

(3)S_i 预测类型。训练集包含 80%的边(DTI)，测试集与验证集分别只包含 10%的边(DTI)。

本次实验重要的超参数包括卷积神经网络的层数 L、学习率 α、丢弃率 β 以及图卷积网络中每层的维度等。为了解不同参数对模型的影响，本文在 IC 数据集上对上述 4 种类型的超参数做消融实验，结果如图 3 所示。图 3(a)－图 3(d)分别为图卷积神经网络层数、丢弃率、学习率以及每层维度数的实验。最后本文选取在 S_d 与 S_t 预测类型下，AUC 和 AUPR 的最大值作为模型的最终超参数，分别为 $L=2$，$\alpha=0.001$，$\beta=0.01$，$d_1=64$ 和 $d_2=32$。

过平滑现象是图卷积神经网络无法像卷积神经网络那样通过叠加网络层数来提高模型性能的主要原因。图 3(a)中，预测结果在第二层达到最优效果后，随着网络层数增多，无论是 AUC 还是 AUPR 值都开始不断下降。Li 等[33]也表示，图卷积神经网络的过平滑现象会使得局域特征的值都收敛到相同的值上，导致图卷积神经网络性能下降。虽然 ResGCN-28[34]使用残差、稠密连接和膨胀卷积解决了图卷积网络在某些网络中的梯度消失以及过平滑问题，但链路预测、节点分类以及聚类等领域还未出现有效的解决方法。

在训练过程中，采用 Adam[35]作为模型的优化算法。为了防止出现过拟合，GCNDTI 方法引入了随机丢弃策略、权重衰减以及 mini-batch 分块方法。在分块时，考虑到每种数据集中样本数量不同，因此在每种数据上训练时实际分块的大小也不相同。



注：虚线为药物，实线为靶标

图 3　网络层数、丢弃率、学习率以及维度对 GCNDTI 模型的影响

Fig. 3　Influence of number of network layers, dropping rate, learning rate and dimension on GCNDTI model

4.4　实验分析

在 Yamanishi_08 与 DrugBank_FDA 数据集上对比实验结果的 AUC 值，结果如表 2 所列。从表 2 可知，在 IC，E 和 DrugBank_FDA 数据集中，GCNDTI 方法的 AUC 值与 TriModel 和 DDR 方法的最优值基本相同，而 GCNDTI 方法的 AUC 值较低的情况集中在 NR 与 GPCR 数据集上，这是因为这两个数据集中已知药物和靶标作用关系数目偏少，即 A_d 与 A_p 数值较少，可能会导致权值参数训练优化不充分，使得模型的预测能力下降。相比之下，由于 IC，E 和 DrugBank_FDA 数据集的数据量较多，GCNDTI 方法的 AUC 值也随之提高。

表 2　药物靶标模型在 5 种数据集中 3 个指标 S_d，S_t，S_i 的结果对比

Table 2　Comparison of three indexes (S_d, S_t, S_i) of state-of-the-art models on five datasets

M	Model	NR			GPCR			IC			E			DB_FDA		
Config.		S_d	S_t	S_i	S_d	S_t	S_i	S_d	S_t	S_i	S_d	S_t	S_i	S_d	S_t	S_i
AUC	BLM-NII	0.88	0.85	0.91	0.85	0.87	0.88	0.83	0.89	0.91	0.73	0.89	0.96	0.71	0.75	0.90
	NRLMF	0.88	0.83	0.93	0.87	0.92	0.95	0.80	0.93	0.98	0.75	0.90	0.95	0.89	0.80	0.93
	DDR	0.90	**0.88**	0.92	0.91	**0.93**	0.96	0.94	0.97	0.98	0.84	0.92	0.97	0.91	0.86	0.96
	TriModel	0.89	0.85	**0.99**	**0.92**	0.86	**0.99**	0.93	**0.98**	**0.99**	0.95	0.96	**0.99**	**0.94**	**0.94**	**0.99**
	GCNDTI	**0.96**	0.85	0.88	0.90	0.80	0.93	**0.97**	0.95	0.97	**0.97**	**0.98**	0.97	0.92	0.92	0.90
AUPR	BLM-NII	0.35	0.41	0.62	0.37	0.37	0.53	0.37	0.61	0.83	0.22	0.73	0.86	0.03	0.05	0.12
	NRLMF	0.49	0.45	0.72	0.35	0.55	0.69	0.30	0.71	0.79	0.28	0.76	0.89	0.28	0.23	0.32
	DDR	0.71	0.64	0.83	0.63	0.61	0.79	0.69	0.80	0.92	0.73	0.82	0.92	0.44	0.39	0.61
	TriModel	0.87	0.77	0.84	0.81	0.73	0.80	0.76	0.87	0.95	0.78	0.83	0.96	0.59	0.62	0.64
	GCNDTI	**0.96**	**0.84**	**0.85**	**0.94**	**0.91**	**0.92**	**0.97**	**0.95**	**0.97**	**0.98**	**0.98**	**0.97**	**0.87**	**0.82**	**0.76**

注：state-of-the-art 模型中数据源自 TriModel；黑体与阴影部分为最佳结果；M 为评价标准

5 个数据集对比实验结果的 AUPR 值如表 2 所列。从表 2 可知，在 15 项预测任务中，GCNDTI 方法的 AUPR 值都高于其他 5 种对比方法。与次优的 TriModel 方法相比，在所有预测任务中，GCNDTI 方法的 AUPR 值平均提高了 5%以上，其中 S_d 预测任务中 GCNDTI 方法比 TriModel 方法的 AUPR 值平均提高了 18%，S_t 预测任务中 AUPR 值平均提

高了 13.6%，S_i 预测任务中 AUPR 值平均提高了 5.6%。另外，GCNDTI 方法中的节点信息只有两跳步内的已知邻居节点信息经聚合过程得到，并没有使用异构网络中的大量未知信息以及相关度较弱的节点信息，减弱了异构网络图中其他节点信息的干扰。因此，该方法在高度不平衡的数据集中具有显著的表现力，特别对新药物 DTI 与新靶标 DTI 的预测表现更佳。

GCNDTI 方法不仅在正样本稀疏的数据集上具有良好的表现力，而且在数据集规模较大时 AUPR 值仍然表现较好。TriModel，DDR，NRLMF 和 BLM-NII 方法在数据集 E 与 IC 上预测任务时，实验结果的 AUPR 值普遍高于 60%，TriModel 方法甚至高于 80%，然而在较大型数据集 DrugBank_FDA 上执行同样的预测任务时，AUPR 值大幅度降低。对比数据集 E 和 DrugBank_FDA 的实验结果，4 种方法的 AUPR 值平均降低了 24.2%，33.3%，34.5%和 54.2%。对相同的实验任务在相同的数据集上进行对比，GCNDTI 方法的 AUPR 值分别仅平均降低 15.2%，降低幅度为其他 4 种方法的 1/3～1/2。由此可知，GCNDTI 方法受数据分布和数据量的影响较小。

GCNDTI 方法优于其他方法的原因是其在训练过程中采用非线性方法融合后的矩阵信息并对不同边类型赋予不同参数权重来进行 DTI 预测。与 GCNDTI 方法相比，TriModel，DDR，NRLMF 和 BLM-NII 方法都存在一定的缺陷。BLM-NII 利用的药物和靶标信息为单个药物和靶标相似矩阵，可能存在信息不足的问题。NRLMF 方法利用了多个药物和靶标相似矩阵的简单线性组合，但简单的线性组合无法处理相似矩阵内的冗余信息。DDR 和 TriModel 方法利用了多种矩阵非线性融合得到药物和靶标信息，因此两种方法的实验结果的 AUC 值和 AUPR 值都优于 DNLMF 和 BLM-NII 方法，但是 DDR 和 TriModel 方法都只利用了部分相连边信息和固定的权重融合邻居特征信息，无法根据边的类型赋予其不同权值。本文提出的 GCNDTI 方法根据边的类型赋予其不同权值，并能通过训练优化增强作用关系边的权值，因此具有更优的表现能力。

4.5 未知药物-靶标关系的发现与验证

良好的药物靶标预测方法不仅对上述 3 类预测任务具有显著的表现力，而且还具有发现数据集中未知(无标注)DTI 的能力[16]。未知药物-靶标关系指原本在 5 个药物-靶标数据集中不存在的作用关系(标记为 0)，但通过预测得到的一类药物-靶标关系。这些作用关系虽然在各个测试数据集中不存在，但实际上可能存在于更全、更新的生物数据库中。

未知药物-靶标关系的发现是通过 GCNDTI 预测的概率矩阵得到的。在解码器得到概率矩阵后，剔除所有数据集中已存在的作用关系，将保留下来的具有前 top-N 概率评分的药物-靶标关系作为本文方法所发现的未知药物-靶标关系。为了降低所确定的未知 DTI 的随机性，本文累计计算出了 10 次预测矩阵的概率评分，利用加权求和方法计算出最终评分。未知药物-靶标关系的验证是在 KEGG[27]，DrugBank[30]，SuperTaget[29]，CTD[36]，T3DB[37] 和 ChEMBL[38] 6 个专业的生物数据库中进行的，这些数据库更新较快，数据更全面，可能包含这些未知药物-靶标关系。在实际的未知 DTI 的发现与验证中，GCNDTI 方法在 5 个数据集上预测(发现)的 top-5 未知 DTI 关系信息如表 3 所列，其中在相关数据库中能够搜寻到的药物-靶标关系用粗体表示(表示这些 DTI 关系得到验证)。表 3 最后一列(Evidence)列出了包含此未知关系的生物数据库名称。由表 3 可知，对 5 个数据库采用 GCNDTI 方法，根据概率评分的高低得到的 top-5 未知药物-靶标关系共有 25 对，其中有 14 对在专业生物数据库上得到验证。

表 3 各个数据集中 top-5 新药物-靶标关系的验证

Table 3 Validation of the top-5 scored combination for each unknown DTIs of the investigated datasets

Dataset	#	Drug Name	Drug Id	Target Name	Target Id	Evidence
NR	1	**Palmitic acid**	**D05341**	**ESR1**	**has:2099**	ChEMBL
	2	Tazarotene	D01132	NR0B1	has:190	None
	3	**Progesterone**	**D00066**	**ESR2**	**has:2100**	DrugBank
	4	Etretinate	D00316	NR0B1	has:190	None
	5	Palmitic acid	D05341	ESR2	has:2100	None
GPCR	1	**Atenolol**	**D00235**	**ADRB3**	**has:155**	SuperTarget
	2	**Amiodarone**	**D02910**	**ADRB3**	**has:155**	DrugBank
	3	**Amiodarone**	**D02910**	**ADRB2**	**has:154**	DrugBank
	4	**Bisoprolol**	**D02345**	**ADRB3**	**has:155**	SuperTarget
	5	Nilutamide	D00965	ADRA1D	has:146	None
IC	1	Carbachol	D00524	CHRNA1	has:1134	None
	2	Caffeine	D00528	CHRNA5	has:1138	None
	3	**Nicotine**	**D03365**	**CHRNA1**	**has:1134**	KEGG
	4	**Carbachol**	**D00524**	**CHRNA5**	**has:1138**	CTD
	5	**Nicotine**	**D03365**	**CHRNA5**	**has:1138**	KEGG
E	1	**Methoxsalen**	**D00139**	**CYP1A1**	**has:1543**	DrugBank
	2	**Metyrapone**	**D00419**	**CYP1A1**	**has:1543**	CTD
	3	**Salicylic acid**	**D00097**	**PTGS2**	**has:5743**	DrugBank
	4	Anastrozole	D00960	CYP46A1	has:10858	None
	5	Caffeine	D00528	CYP2A7	has:1549	None
DB_FDA	1	**Amitriptyline**	**DB00321**	**HTR3A**	**P46098**	ChEMBL
	2	Desipramine	DB01151	HTR3A	P46098	None
	3	**Dopamine**	**DB00988**	**HTR2A**	**P28223**	CTD
	4	Epinastine	DB00751	HTR1A	P08908	None
	5	Loxapine	DB00408	HTR2B	P41595	None

GCNDTI方法预测的未知药物-靶标关系也可在文献中找到证明。例如，在GPCR数据集中，Amiodarone（碘酰酮，一种用于延长各部心肌组织动作以及降低窦房结自律性的药物）与ADRB3（β3肾上腺素能受体，一种具有调节心血管与信号转导功能的靶标蛋白质）并没有作用关系，但GCNDTI预测了Amiodarone与ADRB3具有作用关系。这对作用关系除了可在DrugBank中查询得到，文献[39]也证实了作用关系的存在。

为了比较未知药物-靶标关系发现与验证性能的优劣，在top-N参数相同的情况下，本文在NR，GPCR，IC和E数据集上预测DTI，并对比了GCNDTI，TriModel，NRLMF和BLM-NII获得的验证结果数量的大小（命中数量）。若只考虑top-5的未知药物-靶标关系，由于概率评分值可能在预测中存在微小变化，导致top-5未知药物-靶标关系集不稳定，因此本文以top-10为对比实验的top-N参数。在Si预测类型下，4种方法的top-10未知药物-靶标关系的命中数量结果如表4所列。从表中可以看出，GCNDTI方法在预测未知药物-靶标命中数量上与最优方法TriModel相当。

表4　在S_i预测情况下，4种方法top-10预测中得到验证的数目

Table 4　Verified number of four methods in top-10 predications under S_i

	NR	GPCR	IC	E
BLM-NII	3	7	3	7
NRLMF	5	6	6	**9**
TriModel	**7**	**8**	**7**	7
GCNDTI	6	**8**	**7**	7

结束语　本文提出了一种基于图卷积神经网络模型的方法（GCNDTI）来预测DTI。该方法采用图卷经神经网络在包含多种药物（靶标）相关信息的异质信息网络上学习得到能精确表达每个节点拓扑特征及邻居特征信息的低维向量表征，并利用这些向量信息来预测节点间概率的评分。实验结果证明，在AUPR标准下，相比其他4种方法，GCNDTI方法在DTI预测中对新药物与新靶标预测有显著的效果，其对部分未知作用关系的预测也与目前的最优方法相当，并且随着样本总量的增加，GCNDTI的性能降低幅度较小 。

在本文研究的基础上，未来可从以下3个方面进行进一步的研究。

（1）引入多种经过相似筛选与融合处理的矩阵信息或者直接使用未经过处理的多种相似矩阵信息作为本文方法的输入信息，通过药物和靶标矩阵信息的处理多样性来提高预测能力。

（2）扩展现有异质信息网络的节点类型，引入副作用节点、疾病节点与基因节点等，通过节点种类与边数量的增加来改进预测性能。

（3）将图卷积神经网络模型应用于与药物靶标关系预测相似的其他领域，如基因和疾病预测、药物副作用预测等。

参 考 文 献

[1] NOVAC N. Challenges and opportunities of drug repositioning [J]. Trends in Pharmacological Sciences, 2013, 34(5): 267-272.

[2] EZZAT A, WU M, LI L X, et al. Computational prediction of drug-target interactions using chemogenomic approaches: an empirical survey[J]. Brief Bioinform, 2019, 20(4): 1337-1357.

[3] CHENG A, COLEMAN R, SMYTH K, et al. Structure-based maximal affinity model predicts small-molecule druggability[J]. Nature Biotechnology, 2007, 25(1): 71-75.

[4] MUHAMMAD S, FATIMA N. In silico analysis and molecular docking studies of potential angiotensin-converting enzyme inhibitor using quercetin glycosides [J]. Pharmacognosy Magazine, 2015, 11(Suppl 1): S123.

[5] KEISER M, ROTH B, ARMBRUSTER B, et al. Relating protein pharmacology by ltigand chemistry[J]. Nature Biotechnology, 2007, 25(2): 197.

[6] LUO YN, ZHAO X B, ZHOU J T, et al. A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information[J/OL]. Nature Communications. https://www. biorxiv. org/content/biorxiv/early/2017/01/13/100305. full. pdf.

[7] XU B B, CEN K T, HUANG J J, et al. A Survey on Graph Convolutional Neural NetWork[J]. Chinese Journal of Computers, 2020, 43(5): 755-780.

[8] SCARSELLI F, GORI M, TSOI A, et al. The graph neural network model[J]. IEEE Transactions on Neural Networks, 2008, 20(1): 61-80.

[9] KIPF T N, WELLING M. Semi-supervised classification with graph convolutional networks[J]. arXiv: 1609. 02907, 2016.

[10] BLEAKLEY K, YAMANISHI Y. Supervised prediction of drug-target interactions using bipartite local models[J]. Bioinformatics, 2009, 25(18): 2397-2403.

[11] MEI J P, KWOH C K, YANG P, et al. Drug-target interaction prediction by learning from local information and neighbors[J]. Bioinformatics, 2013, 29(2): 238-245.

[12] XIA Z, WU L Y, ZHOU X B, et al. Semi-supervised drug-protein interaction prediction from heterogeneous biological space [J]. BMC Systems Biology, 2010, 4: S2-S6.

[13] LIU Y, WU M, MIAO C Y, et al. Neighborhood Regularized Logistic Matrix Factorization for Drug-Target Interaction Prediction[J]. PLOS Computational Biology, 2016, 12(2).

[14] HAO M, BRYANT S, WANG Y L. Predicting drug - target interactions by dual-network integrated logistic matrix factorization [J]. Scientific Reports, 2017, 7(1).

[15] OLAYAN R, ASHOOR H, BAJIC V, et al. DDR: efficient computational method to predict drug-target interactions using graph mining and machine learning approaches[J]. Bioinformatics, 2018, 34(7): 1164-1173.

[16] MOHAMED S K, NOVÁEK V, NOUNU A. Discovering protein drug targets using knowledge graph embeddings[J]. Bioinfomatics, 2020, 36(2): 603-610.

[17] HU W H, LIU B W, GOMES J, et al. Pre-training Graph Neural Networks[J]. arXiv: 1905, 12265, 2019.

[18] ZHOU J, CUI G Q, ZHANG Z Y, et al. Graph neural networks: A review of methods and applications[J]. arXiv: 1812. 08434, 2018.

[19] HAMILTON W, YING Z T, LESKOVES J. Inductive representation learning on large graph[C]// Advance in Neural Informa-

tion Processing System. 2017:1024-1034.
[20] BRUNA J, ZARAMBA W, SZLAM A, et al. Spectral networks and locally connected networks on graphs[J]. arXiv:1312.6203, 2013.
[21] DEFFERRAR M, BRESSON X, VANDERGHETNS P. Convolutional neural networks on graphs with fast localized spectral filtering[C]// Advance in Neural Information Processing System. 2016:3844-3852.
[22] CHEN J, MA T F, XIAO C. Fastgcn: fast learning with graph convolutional networks via importance sampling[J]. arXiv: 1801.10247, 2018.
[23] ZITNIK M, AGRAWAL M, LESKOVEC J. Modeling polypharmacy side effects with graph convolutional networks[J]. Bioinformatics, 2018, 34(13): i457-i466.
[24] MIKOLOV T. Distributed Representations of Words and Phrases and their Compositionality[J]. Advances in Neural Information Processing Systems, 2013, 26: 3111-3119.
[25] WAN F P, HONG L X, XIAO A, et al. NeoDTI: neural integration of neighbor information from a heterogeneous network for discovering new drug-target interactions[J]. Bioinformatics, 2018, 35(1): 104-111.
[26] YAMANISH Y, ARAKI M, GUTTERIDG A, et al. Prediction of drug-target interaction networks from the integration of chemical and genomic spaces[J]. Bioinformatics, 2008, 24(13): i232-i240.
[27] KANEHISA M, GOTO S, HATTOR M, et al. From genomics to chemical genomics: new developments in KEGG[J]. Nucleic acids research, 2006, 34(suppl_1): D354-D357.
[28] SCHOMBURG I, CHANG A, EBELING C, et al. BRENDA, the enzyme database: updates and major new developments[J]. Nucleic acids research, 2004, 32(suppl_1): D431-D433.
[29] GUNTHER S, KUHN M, DUNKEL M, et al. SuperTarget and Matador: resources for exploring drug-target relationships[J]. Nucleic Acids Res, 2008, 36(Database issue): D919-D922.
[30] WISHART D, KNOX C, GUO A C, et al. DrugBank: a knowledgebase for drugs, drug actions and drug targets[J]. Nucleic Acids Research, 2008, 36(Database issue): D901-D906.
[31] WANG B, MEZLINI A, DEMIR F, et al. Similarity network fusion for aggregating data types on a genomic scale[J]. Nature Methods, 2014, 11(3): 333-337.
[32] YU D H, GUO M Z, LIU X Y, et al. Predicted Results Evaluation and Query Verification of Drug-Target Interaction[J]. Journal of Computer Research and Development, 2019, 56(9): 1881-1888.
[33] LI Q M, HAN Z C, WU X M. Deeper insights into graph convolutional networks for semi-supervisedlearning[C]// Thirty-Second AAAI Conference on Artificial Intelligence. 2018.
[34] LI G H, MATTHIAS M, ALI T, et al. DeepGCNs: Can GCNs go as deep as CNNs? [C]// Proceedings of the IEEE International Conference on Computer Vision. 2019.
[35] KINGMA D, BA J. Adam: A method for stochastic optimization [J]. arXiv:1412.6980, 2014.
[36] DAVIS A, GRONDIN C, JOHNSON R, et al. The comparative toxicogenomics database: update 2017[J]. Nucleic Acids Research, 2016, 45(D1): D972-D978.
[37] WISHART D, ARNDT D, PON A, et al. T3DB: the toxic exposome database[J]. Nucleic Acids Research, 2015, 43(Database issue): D928-D934.
[38] GAULTON A, BELLIS L, BENTO A, et al. ChEMBL: a large-scale bioactivity database for drug discovery[J]. Nucleic Acids Research, 2012, 40(Database issue): D1100-D1107.
[39] DISATNIK M H, SHAINBER A. Regulation of beta-adrenoceptors by thyroid hormone and amiodarone in rat myocardiac cells in culture[J]. Biochemical Pharmacology, 1991, 41(6/7): 1039-1044.

GAO Chuang, born in 1995, master. His main research interests include graph convolutional neural network and recommendation system.

LI Jian-hua, born in 1977, Ph.D, associate professor, is a member of China Computer Federation. His main research interests include computer drug design and data mining.