基于多重相似性和增强注意力预测药物-靶标相互作用

王伟[1a,b],余梦雪[1a],孙斌[1a],万仕彤[1a],刘栋[1a,b],周运[1a,b],张红军[2],王鲜芳[3]

(1.a.河南师范大学 计算机与信息工程学院;b.河南省教育人工智能与个性化学习重点实验室,河南 新乡 453007;
2.河南理工大学 鹤壁工程技术学院,河南 鹤壁 458030;3.河南工学院 计算机科学与技术学院,河南 新乡 453000)

摘 要:在新药发现和药物重定位研究中,发现药物与靶标之间的相互作用是重要的研究内容.针对药物与靶标相互作用网络,提出一种基于多重相似性和增强注意力机制的图卷积神经网络模型(RSGCN)预测药物-靶标相互作用.首先,提出了多重相似性来捕捉网络结构特征,以充分利用节点间的直接或间接关系.然后,通过 PCA 降维减少相似性噪声对实验结果的影响.最后,采用图卷积神经网络(graph convolution neural network,GCN)获得节点嵌入表示,并融入基于注意力的增强层,通过增强注意力机制获得节点间的注意力权重,能够高效地预测药物与靶标之间的相互作用.在黄金标准数据集上的实验结果表明 RSGCN 模型具有较好的性能.

关键词:图卷积神经网络(GCN);多重相似性;PCA;增强注意力机制;药物-靶标相互作用

中图分类号:TP181 **文献标志码**:A **文章编号**:1000-2367(2025)02-0099-09

药物重定位(或称为药物再利用)是一种对已批准药物发现其新用途的重要策略,为药物研发提供了一种快速高效的新途径.药物研发是一个系统工程,其中药物-靶标的发现和验证是重要的研发过程之一,确定药物-靶标相互作用(DTI)是药物开发和药物重新定位早期阶段的关键步骤[1-2].然而,由于采用湿实验方法鉴定药物-靶标相互作用(DTI)存在研发费用高且研发周期长的问题,因此,采用计算方法从大量候选药物中筛选潜在的 DTI,能够减轻费用昂贵和耗时的湿实验研究工作,提高药物发现的效率.

近年来,图神经网络的快速发展将深度学习应用扩展到了图领域,并已应用于基于网络的药物发现[3].基于网络的方法通常包括两个步骤:网络构建和 DTI 预测.这些方法不仅考虑药物之间的关系,而且考虑靶点之间的关系,因此基于网络的方法受到越来越多的关注[4],通过网络嵌入整合药物和靶点的各种信息,可以进一步提高 DTI 预测的准确性[5-7].一些研究工作在此领域取得了重要进展.例如,SHANG 等[8]提出了一种基于多视图网络嵌入方法来预测潜在的 DTI,整合了药物和靶标的异质信息.MHADTI[9]是一种基于多视角异构信息网络的药物-靶点相互作用预测,利用药物和靶点的多源信息构建不同的相似性网络.DTI-HETA[10]利用图卷积神经网络获得药物和靶标的嵌入表示.为了突出不同邻域节点对中心节点聚集图卷积信息的贡献,在节点嵌入过程中引入了图注意机制.KronRLS-MKL[11]能够集成多个异构信息源,适用于任意不同的网络规模,此外,它通过返回权重来自动选择更相关的内核.RTHNEDT[12]是一种基于关系拓扑的网络嵌入方法来预测药物与靶标的相互作用,该模型在带有标签的网络和未带标签的网络都能获得较好的预测性能.HNEDTI[13]模型对药物相似度矩阵和靶标相似度矩阵分别设置两个相似度阈值参数,过滤相似度较低的边,然后用已知的药物相关网络和靶标相关网络构建药物与靶标异质网络.GMDTI[14]是一种基于异

收稿日期:2023-06-28;**修回日期**:2023-08-03.

基金项目:国家自然科学基金(62072160;62072157);河南省科技攻关项目(242102211045;242102210001).

作者简介(通信作者):王伟(1975—),男,河南新乡人,河南师范大学副教授,博士,研究方向为机器学习、生物信息学和数据挖掘,E-mail:weiwang@htu.edu.cn.

引用本文:王伟,余梦雪,孙斌,等.基于多重相似性和增强注意力预测药物-靶标相互作用[J].河南师范大学学报(自然科学版),2025,53(2):99-107.(Wang Wei,Yu Mengxue,Sun Bin,et al.Prediction of drug-target interaction based on multiple similarity and enhanced attention mechanisms on graph convolution neural network[J].Journal of Henan Normal University(Natural Science Edition),2025,53(2):99-107.DOI:10.16366/j.cnki.1000-2367.2023.06.28.0003.)

质信息并融合元路径信息的图神经网络模型,用于预测药物-靶标相互作用.PDML[15]是一种基于弱标记和多信息融合的药物-靶标相互作用方法.

在本研究中,提出了一种基于多重相似性和增强注意力机制的图卷积神经网络模型预测药物与靶标相互作用(RSGCN).与现有的药物-靶标相互作用预测方法不同,模型使用多重相似性模块优化药物和靶标节点的原始特征向量,捕获网络结构特征,以充分获取节点之间的直接或间接关系.具体而言,模型采用相似性融合的方法,获得药物和靶标的相似性网络.通过重启随机游走和余弦相似性方法获得药物和靶标的多重相似性网络.然后,模型使用 PCA 来降低维度,以减少相似性噪声对计算结果的影响.接下来,通过图卷积神经网络(GCN)提取网络节点特征,以获得节点的低维表示.为了更有效地发现相邻节点之间的关系,模型引入增强注意力机制来获取节点之间的注意力,最终预测药物与靶标之间的相互作用,实验结果与现有的药物-靶标相互作用预测方法相比,RSGCN 方法具有良好的性能.模型如图 1 所示.

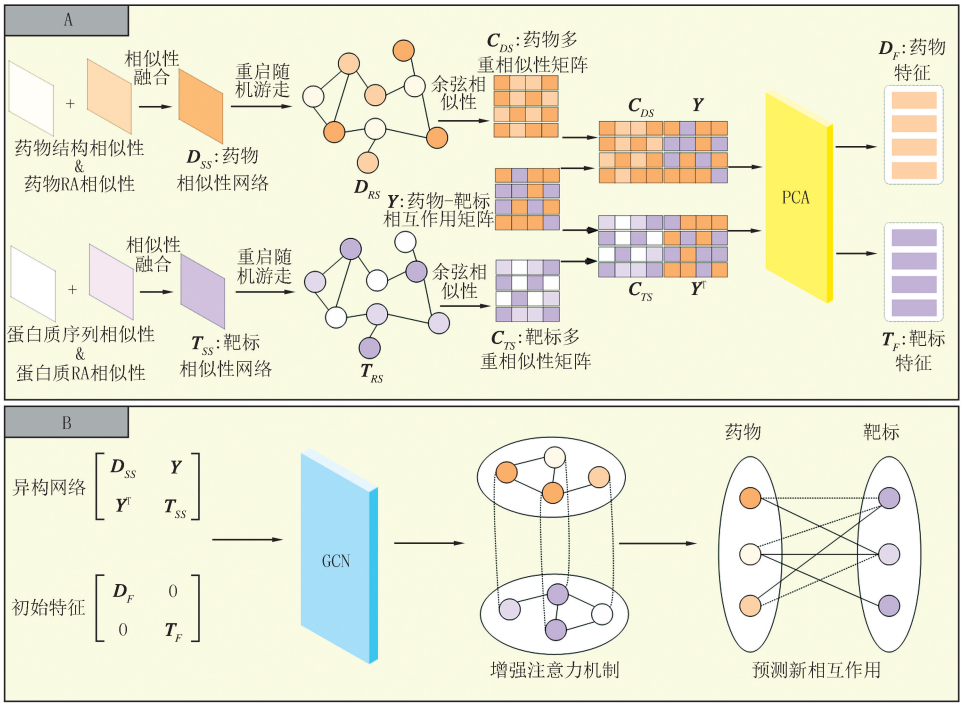

图1 药物-靶标相互作用预测模型框架结构图

Fig.1 Framework structure diagram of drug-target interaction prediction model

1 方 法

1.1 问题定义

将药物和靶标表示为网络中两种不同类型的节点.药物的节点集定义为 $D=\{d_1,d_2,\cdots,d_{N_d}\}$.类似地,靶标节点集定义为 $T=\{t_1,t_2,\cdots,t_{N_t}\}$.其中,$N_d$ 表示药物数量,N_t 表示靶标数量.网络中的边是药物和靶标之间的关联,可以表示为邻接矩阵 $\boldsymbol{Y}\in\mathbf{R}^{N_d\times N_t}$.矩阵 $\boldsymbol{Y}$ 是一个只有 0 和 1 的矩阵,这里,$\boldsymbol{Y}_{i,j}=1$ 指一种药物与一种靶标相互作用.相反,$\boldsymbol{Y}_{i,j}=0$ 表示潜在关联或者无关联.本文的任务是获得一个药物 - 靶标得分矩阵 $\boldsymbol{F}^*$.得分越高相互作用的可能性越大,根据得分预测出潜在的药物-靶向相互作用可为药物发现实验提供实验指导,以提高药物研发效率.

1.2 相似性融合

相似性融合可以获得足够的节点信息,有助于提高模型的预测能力.药物相似性网络通过融合药物结构相似性矩阵和药物 $\boldsymbol{R}$ 相似性(resource allocation similarity)矩阵构建.

$$\boldsymbol{D}_{SS}=d\times\boldsymbol{D}_S+(1-d)\times\mathbf{R}_d, \tag{1}$$

$\boldsymbol{D}_{SS}\in\mathbf{R}^{N_d\times N_d}$,$N_d$ 表示药物数量.d 是药物结构相似性矩阵权重系数,$\boldsymbol{D}_S$ 是药物结构相似性矩阵,$\boldsymbol{D}_S\in$

$\mathbf{R}^{N_d\times N_d}$，实验采用 SIMCOMP[16] 计算化合物 d_i 和 d_j 之间的化学结构相似性，公式如下所示：

$$\boldsymbol{D}_S(d_i,d_j)=|\ d_i\cap d_j\ |\ /\ |\ d_i\cup d_j\ |,\tag{2}$$

$|\ d_i\cap d_j\ |$ 表示两种化合物 d_i 和 d_j 之间的共同子结构的大小，$|\ d_i\cup d_j\ |$ 表示两种化合物 d_i 和 d_j 具有子结构的大小.

$\boldsymbol{R}_d$ 是 $\boldsymbol{R}$ 相似性矩阵，$\boldsymbol{R}_d\in\mathbf{R}^{N_d\times N_d}$. $\boldsymbol{R}$ 相似性是对已知的相互作用关系中的邻居节点信息来计算相似性得分.首先为网络中的每个节点分配一个初始单元，然后依据网络中节点间的拓扑关系，把初始单元均匀分发给相邻节点.初始节点与目标节点之间路径上分配的资源总和就是两节点之间的 $\boldsymbol{R}$ 相似性.例如，节点 d_i 和 d_j 节点之间 $\boldsymbol{R}$ 相似性公式如下：

$$\boldsymbol{R}(d_i,d_j)=\sum\nolimits_{r=N(d_i)\cap N(d_j)}\frac{1}{|\ N(r)\ |},\tag{3}$$

$\boldsymbol{R}(d_i,d_j)$ 的值越大，表示起始节点 d_i 与目标节点 d_j 之间的相似性越高.其中 $N(d_i)$ 和 $N(d_j)$ 分别表示节点 d_i 和节点 d_j 的邻居节点集，r 和 $N(d_i)\cap N(d_j)$ 表示节点 d_i 和节点 d_j 共有的节点集，$|\ N(r)\ |$ 表示 r 的邻居节点个数.

靶标相似性网络 $\boldsymbol{T}_{SS}$ 是通过融合蛋白质序列相似性 $\boldsymbol{P}_S$ 和蛋白质 $\boldsymbol{R}$ 相似性计算得到，其计算公式如下：

$$\boldsymbol{T}_{SS}=t\times\boldsymbol{P}_S+(1-t)\times\boldsymbol{R}_t,\tag{4}$$

$\boldsymbol{T}_{SS}\in\mathbf{R}^{N_t\times N_t}$，$N_t$ 表示靶标数量.t 是蛋白质序列相似性矩阵权重系数，$\boldsymbol{P}_S$ 是蛋白质序列相似性矩阵，实验使用标准化版本的 Smith-Waterman 得分[17] 计算蛋白质之间的序列相似性.两种蛋白质 t_i 和 t_j 之间的 Smith-Waterman 得分如下：

$$\boldsymbol{P}_S(t_i,t_j)=S_W(t_i,t_j)/\sqrt{S_W(t_i,t_i)}\ \sqrt{S_W(t_j,t_j)}.\tag{5}$$

类似药物相似性融合，这里 $\boldsymbol{R}_t$ 指靶标 $\boldsymbol{R}$ 相似性矩阵，节点 t_i 和节点 t_j 之间 $\boldsymbol{R}$ 相似性如公式所示：

$$\boldsymbol{R}(t_i,t_j)=\sum\nolimits_{r=N(t_i)\cap N(t_j)}\frac{1}{|\ N(r)\ |}.\tag{6}$$

1.3　重启随机游走(RWR)

本模型通过 RWR 来获取节点间的相似性.靶标相似性网络 $\boldsymbol{T}_{SS}$ 和药物相似性网络 $\boldsymbol{D}_{SS}$ 通过 RWR 算法得到靶标特征矩阵 $\boldsymbol{T}_{RS}$ 和药物特征矩阵 $\boldsymbol{D}_{RS}$，$\boldsymbol{D}_{RS}\in\mathbf{R}^{N_d\times N_d}$，$\boldsymbol{T}_{RS}\in\mathbf{R}^{N_t\times N_t}$. RWR[18] 的方法定义为：

$$r_i^k=c\boldsymbol{W}r_i^{k-1}+(1-c)\boldsymbol{e}_i,\tag{7}$$

其中 $\boldsymbol{W}=[i,j]$ 是归一化后转移概率矩阵，$\boldsymbol{e}_i\in\mathbf{R}^{n+1}$ 是第 i 个节点初始向量($\boldsymbol{T}_{SS}$ 或 $\boldsymbol{D}_{SS}$ 的行向量)，c 是重启概率，r_i^k 是 RWR 计算后得到的相似度向量.

1.4　余弦相似性

为充分利用节点之间的关系，模型采用余弦相似性计算方法，公式如下：

$$\mathrm{sim}(n,m)=n\cdot m/\|\ n\ \|\times\|\ m\ \|=\sum\nolimits_{i=1}^{N}(n_i\times m_i)/\sqrt{\sum\nolimits_{i=1}^{N}n_i^2}\times\sqrt{\sum\nolimits_{i=1}^{N}m_i^2},\tag{8}$$

这里 n_i 和 m_i 分别表示节点 n 和节点 m 的分量.经过 RWR 得到的靶标矩阵 $\boldsymbol{T}_{RS}$ 和药物矩阵 $\boldsymbol{D}_{RS}$，利用余弦相似性得出靶标多重相似性矩阵 $\boldsymbol{C}_{TS}$ 和药物多重相似性矩阵 $\boldsymbol{C}_{DS}$，其中 $\boldsymbol{C}_{DS}\in\mathbf{R}^{N_d\times N_d}$，$\boldsymbol{C}_{TS}\in\mathbf{R}^{N_t\times N_t}$.

1.5　特征矩阵融合

为了更好地聚集网络节点信息，模型将靶标多重相似性矩阵 $\boldsymbol{C}_{TS}$ 和药物多重相似性矩阵 $\boldsymbol{C}_{DS}$ 分别与药物-靶标相互作用矩阵 $\boldsymbol{Y}$ 相融合，获得靶标特征矩阵 $\boldsymbol{S}_T$ 和药物特征矩阵 $\boldsymbol{S}_D$，公式如下：

$$\boldsymbol{S}_D=[\boldsymbol{C}_{DS}\quad\boldsymbol{Y}],\tag{9}$$

$$\boldsymbol{S}_T=[\boldsymbol{C}_{TS}\quad\boldsymbol{Y}^{\mathrm{T}}],\tag{10}$$

其中，$\boldsymbol{S}_D\in\mathbf{R}^{N_d\times(N_d+N_t)}$，$\boldsymbol{S}_T\in\mathbf{R}^{N_t\times(N_d+N_t)}$.$\boldsymbol{Y}\in\mathbf{R}^{N_d\times N_t}$，$\boldsymbol{Y}^{\mathrm{T}}\in\mathbf{R}^{N_t\times N_d}$. N_d 表示药物数量，N_t 表示靶标数量.

1.6　主成分分析(PCA)

模型采用 PCA 降噪和特征提取，经过多重相似性和邻接矩阵 $\boldsymbol{Y}$ 相结合，噪声往往是不可避免的.为了减少相似性噪声对计算结果的影响，引入了 PCA 算法.PCA 算法能够消除冗余特征，有助于提高处理数据的速度，提升算法的效率.如下所示.

对药物 $\boldsymbol{S}_D$ 矩阵去平均值(即去中心化),即每一位特征减去各自的平均值.计算去中心化矩阵的协方差矩 $\boldsymbol{S}_{DD}$:

$$\boldsymbol{S}_{DD}=\frac{1}{N_d+N_t}\boldsymbol{S}_D{}^{\mathrm{T}}\boldsymbol{S}_D, \tag{11}$$

其中 $\boldsymbol{S}_{DD}\in\mathbf{R}^{((N_d+N_t)\times(N_d+N_t))}$.

模型用特征值分解方法求协方差矩阵 $\boldsymbol{S}_{DD}$ 的特征值与特征向量:

$$\boldsymbol{S}_{DD}=\boldsymbol{Q}\sum\boldsymbol{Q}^{-1}, \tag{12}$$

其中,$\boldsymbol{Q}$ 是矩阵 $\boldsymbol{S}_{DD}$ 的特征向量矩阵,$\sum$ 是一个对角矩阵,其对角线上的元素即特征值.

对特征值从大到小排序,选择其中最大的 l 个,$l=(N_d+N_t)\times k$.然后将其对应的 l 个特征向量分别作为列向量组成特征向量矩阵 $\boldsymbol{P}_d$.获得去噪后的特征矩阵 $\boldsymbol{D}_F$:

$$\boldsymbol{D}_F=\boldsymbol{S}_D\times\boldsymbol{P}_d. \tag{13}$$

同理,靶标得到特征矩阵 $\boldsymbol{T}_F$,$\boldsymbol{T}_F\in\mathbf{R}^{(N_t\times l)}$,$\boldsymbol{D}_F\in\mathbf{R}^{(N_d\times l)}$.

1.7 图卷积神经网络(GCN)

模型通过 GCN 来获得节点的嵌入表示.为了在降噪后的矩阵中使 GCN 更容易找到原始节点之间关联,获得更多节点信息,模型中引入了异构网络.在模型中将药物相似性矩阵 $\boldsymbol{D}_{SS}$、靶标相似性矩阵 $\boldsymbol{T}_{SS}$ 和药物-靶标相互作用矩阵 $\boldsymbol{Y}$ 构建成异构网络,由邻接矩阵 $\boldsymbol{A}_{dt}$ 表示,其构建方法如下:

$$\boldsymbol{A}_{dt}=\begin{bmatrix}\boldsymbol{D}_{SS} & \boldsymbol{Y}\\ \boldsymbol{Y}^{\mathrm{T}} & \boldsymbol{T}_{SS}\end{bmatrix}, \tag{14}$$

这里 $\boldsymbol{A}_{dt}\in\mathbf{R}^{((N_d+N_t)\times(N_d+N_t))}$.

对于初始特征,构建方法如下:

$$\boldsymbol{D}_T=\begin{bmatrix}\boldsymbol{D}_F & 0\\ 0 & \boldsymbol{T}_F\end{bmatrix}, \tag{15}$$

其中 $\boldsymbol{D}_T\in\mathbf{R}^{((N_t+N_d)\times 2l)}$.

邻接矩阵 $\boldsymbol{A}_{dt}$ 与初始特征 $\boldsymbol{D}_T$ 的图卷积计算公式如下:

$$\boldsymbol{F}=(\boldsymbol{I}+\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{A}_{dt}\boldsymbol{D}^{-\frac{1}{2}})\boldsymbol{D}_T\boldsymbol{W}_e, \tag{16}$$

其中,$\boldsymbol{F}\in\mathbf{R}^{((N_d+N_t)\times LFN)}$ 矩阵 $\boldsymbol{W}_e$ 表示训练权重矩阵,$\boldsymbol{W}_e\in\mathbf{R}^{(2l\times LFN)}$,$\boldsymbol{I}$ 表示单位矩阵,$\boldsymbol{D}$ 表示 $\boldsymbol{A}_{dt}$ 的对角矩阵;通过激活函数将偏差矩阵 $\boldsymbol{B}$ 引入到矩阵 $\boldsymbol{F}$ 中,得到特征矩阵 $\boldsymbol{H}$:

$$\boldsymbol{H}=\mathrm{ReLU}(\boldsymbol{F}+\boldsymbol{B}), \tag{17}$$

其中 $\boldsymbol{H}\in\mathbf{R}^{((N_d+N_t)\times LFN)}$,$\boldsymbol{B}\in\mathbf{R}^{((N_d+N_t)\times LFN)}$.

1.8 增强注意力机制

为了使相似的药物(或靶标)节点在特征空间中更好地邻居聚合.本模型引入了增强注意力机制[19],通过学习邻居的权重实现其加权聚合,用于测量邻居节点对节点的影响.a_{ij} 表示节点之间的注意力系数,如下所示:

$$e_{ij}=\mathrm{ReLU}(\boldsymbol{W}_{h_i}+\boldsymbol{W}_{h_j}), \tag{18}$$

$$a_{ij}=\exp(e_{ij})/\sum_{j\in\mathbf{N}_i}\exp(e_{ij}), \tag{19}$$

$$\boldsymbol{H}_i=\sum_{j\in\mathbf{N}_i}a_{ij}h_i, \tag{20}$$

其中,N_i 表示节点 i 的邻居节点集,e_{ij} 表示节点 j 的特征对节点 i 的重要性,$\boldsymbol{W}$ 表示权重矩阵,ReLU 表示激活函数.

1.9 损失函数

模型根据药物和靶标的数量,将特征矩阵 $\boldsymbol{H}$ 分割成表示药物的特征矩阵 $\boldsymbol{H}_d$ 和靶标矩阵 $\boldsymbol{H}_t$,计算出药物-靶标得分矩阵 $\boldsymbol{F}^*$:

$$\boldsymbol{F}^{*}=\boldsymbol{H}_{d}\boldsymbol{W}_{d}\boldsymbol{H}_{t}^{\mathrm{T}}, \tag{21}$$

其中 $\boldsymbol{F}^{*}\in\mathbf{R}^{(N_d\times N_t)}$，$\boldsymbol{W}_d$ 是可训练矩阵且 $\boldsymbol{W}_d\in\mathbf{R}^{(LFN\times LFN)}$.

$\boldsymbol{W}_{np}$ 用于使迭代过程中的预测误差最小化，L 为损失函数[18]：

$$L=\boldsymbol{W}_{np}+\frac{1}{2}\|\boldsymbol{W}_e\|^2+\frac{1}{2}\|\boldsymbol{W}_d\|^2+\frac{1}{2}\|\boldsymbol{B}\|^2, \tag{22}$$

$$\boldsymbol{W}_{np}=\sqrt{\sum\nolimits_{ij;\Phi_{p,ij}=1\text{ or }\Phi_{n,ij}=1}(M'_{ij}-M_{ij})/\sum\nolimits_{ij}(\boldsymbol{\Phi}_{p,ij}+\boldsymbol{\Phi}_{n,ij})}, \tag{23}$$

其中，$\boldsymbol{\Phi}_p$ 和 $\boldsymbol{\Phi}_n$ 分别代表随机挑选的正负样本矩阵.在 $\boldsymbol{\Phi}_p$ 中若某个位置的值为1，就代表此元素为正样本.而在 $\boldsymbol{\Phi}_n$ 中某个位置的值为 1，就代表此元素为负样本.M'_{ij} 代表模型预测标签，M_{ij} 为真实标签.

1.10　RSGCN 算法

RSGCN 是一种基于多重相似性和增强注意力机制的图卷积神经网络模型来预测药物与靶标相互作用，其伪代码如表 1 所示.

表 1　基于多重相似性和增强注意力机制模型算法

Tab. 1　Algorithm based on multiple similarity and enhanced attention model

RSGCN 算法

输入　药物-靶标相互矩阵 $\boldsymbol{Y}$；药物结构相似性矩阵 $\boldsymbol{D}_S$；蛋白质序列相似性矩阵 $\boldsymbol{P}_S$；

输出　药物-靶标得分矩阵 $\boldsymbol{F}^{*}$；

1：计算药物和靶标相似性融合网络 $\boldsymbol{D}_{SS}$ 和 $\boldsymbol{T}_{SS}$；

$\boldsymbol{D}_{SS}=d\times\boldsymbol{D}_S+(1-d)\times\boldsymbol{R}_d$

$\boldsymbol{T}_{SS}=t\times\boldsymbol{P}_S+(1-t)\times\boldsymbol{R}_t$

2：RWR；重启概率 $c=0.8$；

3：$r_i^k=c\boldsymbol{W}r_i^{k-1}+(1-c)\boldsymbol{e}_i$；

4：return 药物特征矩阵 $\boldsymbol{D}_{RS}$ 和靶标特征矩阵 $\boldsymbol{T}_{RS}$；

5：多重相似性：

$$\mathrm{sim}(n,m)=n\cdot m/\|n\|\times\|m\|=\sum\nolimits_{i=1}^{N}(n_i\times m_i)/\sqrt{\sum\nolimits_{i=1}^{N}n_i^2}\times\sqrt{\sum\nolimits_{i=1}^{N}m_i^2};$$

6：return 药物和靶标多重相似性矩阵 $\boldsymbol{C}_{DS}$ 和 $\boldsymbol{C}_{TS}$；

7：特征矩阵融合：

$$\boldsymbol{S}_D=[\boldsymbol{C}_{DS}\quad\boldsymbol{Y}],\boldsymbol{S}_T=[\boldsymbol{C}_{TS}\quad\boldsymbol{Y}^{\mathrm{T}}]$$

return 药物特征矩阵 $\boldsymbol{S}_D$ 和靶标特征矩阵 $\boldsymbol{S}_T$；

8：PCA；$\boldsymbol{S}_D$；$\boldsymbol{S}_T$；概率 $k=0.9$；

9：$\boldsymbol{S}_D$ 和 $\boldsymbol{S}_T$ 去中心化；

10：计算协方差：

$$\boldsymbol{S}_{DD}=\frac{1}{N_d+N_t},\boldsymbol{S}_{TT}=\frac{1}{N_d+N_t}\boldsymbol{S}_T^{\mathrm{T}}\boldsymbol{S}_T,$$

11：特征值与特征向量：

$$\boldsymbol{S}_{DD}=\boldsymbol{Q}\sum\boldsymbol{Q}^{-1},\boldsymbol{S}_{TT}=\boldsymbol{Q}\sum\boldsymbol{Q}^{-1}$$

12：对特征值从大到小排序，选择其中最大的 l 个，$l=(N_d+N_t)\times k$.将其对应的 l 个特征向量分别作为列向量组成药物和靶标特征向量矩阵 $\boldsymbol{P}_d$ 和 $\boldsymbol{P}_t$，

$$\boldsymbol{D}_F=\boldsymbol{S}_D\times\boldsymbol{P}_d,\boldsymbol{T}_F=\boldsymbol{S}_T\times\boldsymbol{P}_t;$$

13：return 去噪后的药物特征矩阵 $\boldsymbol{D}_F$ 和去噪后的靶标特征矩阵 $\boldsymbol{T}_F$；

14：构建矩阵 $\boldsymbol{A}_{dt}$ 和初始特征 $\boldsymbol{D}_T$，

$$\boldsymbol{A}_{dt}=\begin{bmatrix}\boldsymbol{D}_{SS}&\boldsymbol{Y}\\\boldsymbol{Y}^{\mathrm{T}}&\boldsymbol{T}_{SS}\end{bmatrix},\boldsymbol{D}_T=\begin{bmatrix}\boldsymbol{D}_F&0\\0&\boldsymbol{T}_F\end{bmatrix};$$

15：GCN；epoch=1 000；LFN=70；$l_r=0.01$

16：while epoch<1 000 do

17：GCN：$\boldsymbol{F}=(\boldsymbol{I}+\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{A}_{dt}\boldsymbol{D}^{-\frac{1}{2}})\boldsymbol{D}_T\boldsymbol{W}$；

18：$\boldsymbol{H}=\mathrm{ReLU}(\boldsymbol{F}+\boldsymbol{B})$

偏差矩阵 $\boldsymbol{B}$ 引入到矩阵 $\boldsymbol{F}$ 中的特征矩阵 $\boldsymbol{H}$；

19：增强注意力机制：

$$e_{ij}=\mathrm{ReLU}(\boldsymbol{W}h_i+\boldsymbol{W}h_j),$$

$$a_{ij}=\exp(e_{ij})/\sum\nolimits_{j\in\mathbf{N}_i}\exp(e_{ij})$$

$$\boldsymbol{H}_i=\sum\nolimits_{j\in\mathbf{N}_i}a_{ij}h_i;$$

20：计算药物-靶标得分矩阵：$\boldsymbol{F}^{*}=\boldsymbol{H}_d\boldsymbol{W}_d\boldsymbol{H}_t^{\mathrm{T}}$

21：end while

22：输出得分矩阵 $\boldsymbol{F}^{*}$.

1.11　实验数据

本研究采用 Yamanishi 构建的黄金标准数据集，该数据集药物与靶标主要来自 DrugBank[20]、KEGG BRITE[21]、BRENDA[22]和 SuperTarget[23]数据库.数据集包括靶向酶(Es)、离子通道(IC)、G 蛋白偶联受体(GPCR)和核受体(NRs)，共 4 类数据，其中酶(Es)数据集中药物 445 个，靶标 664 个，相互作用对2 926 个；离子通道(IC)数据集中药物 210 个，靶标 204 个，相互作用对 1 476 个；G 蛋白偶联受体(GPCR)数据集中药物 223 个，靶标 95 个，相互作用对 635 个；核受体(NRs)数据集中药物 54 个，靶标 26 个，相互作用对 90 个.

2 结果与讨论

2.1 参数设置

模型中的学习率设置为 0.01.模型采用了学习率自适应的优化算法 Adam.潜在因字数 LFN(latent_factor_num)表示 GCN 嵌入维度,在实验时设置为 5～100,步长为 5,其中 LFN 为 70 时模型性能最高;epoch 是学习算法在整个训练数据集中的迭代次数,在实验时分别设置为 500～2 000,步长为 100,在 epoch 设置为 1 000 时模型性能最高;dropout 分别设置为 0.1～0.9,步长为 0.1,模型在 dropout 为 0.9 时最优;GCN 设置 1 层;在 Es 数据集上进行了以下消融及对照实验.模型采用 10 折交叉验证来评估预测方法的性能.

2.2 相似性融合

为了验证相似性融合对模型的影响,设置了一组消融实验,如表 2 所示,RSGCN(NO-RA)表示采用了药物结构相似性和蛋白质序列相似性;RSGCN(R)表示采用了药物 R 相似性和蛋白质 R 相似性;RSGCN 表示本模型使用了相似性融合.表 2 明显看出在各个评价指标均高于未使用相似性融合,从而说明相似性融合有助于提高模型的预测能力.

表 2 相似性融合实验对比结果

Tab. 2 Comparison results of similarity fusion experiments

Model	AUC	AUPR	F1	MCC	ACC	Recall
RSGCN(NO-RA)	0.980	0.945	0.925	0.924	0.986	0.896
RSGCN(RA)	0.967	0.904	0.883	0.881	0.983	0.843
RSGCN	**0.985**	**0.952**	**0.927**	**0.926**	**0.987**	**0.906**

相似性融合是将药物和靶标的两种相似性矩阵融合成一个网络,对药物和靶标融合时有两个重要的权重系数 t 和 d,其中 t 表示蛋白质序列相似性网络的权重,d 表示药物结构相似性网络的权重.为了得到最优权重系数,将 t 和 d 分别设置为 0.1～0.9,在 IC 数据集上得到实验结果的 AUC 如图 2 所示,t 取 0.2,d 取0.7 时 AUC 达到最高,因此实验中 t 取 0.2,d 取 0.7.并在其他数据集中也筛选出最优的 t 和 d 值,在数据集 NRs 最优 $t=0.1$,$d=0.2$;在数据集 GPCR 最优 $t=0.6$,$d=0.4$,在数据集 Es 最优 $t=0.9$,$d=0.3$.

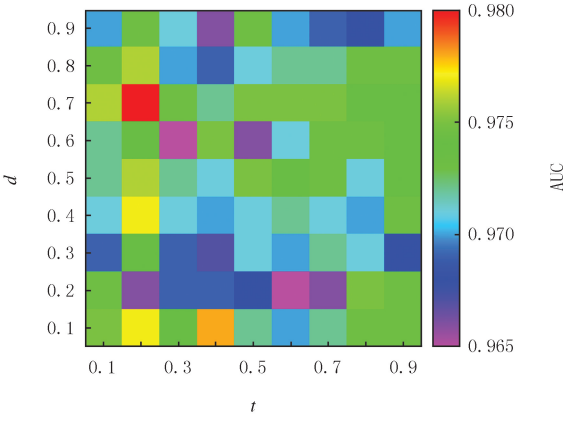

图2 81种组合的AUC分布热图

Fig. 2 81 combinations of AUC distribution heat maps

2.3 多重相似性

为了验证多重相似性的作用,本文进行了消融实验,如表 3 所示,SMGCN(sf+cs)表示使用了相似性融合和余弦相似性;SMGCN(sf+rwr)表示使用了相似性融合和 RWR;SMGCN(sf)表示只使用了相似性融合;RSGCN 是最终的模型,其各项评价指标均达到最高,这是因为多重相似性可以获得更多的节点信息.

表 3 多重相似性实验对比结果

Tab. 3 Comparison results of multiple similarity experiments

Model	AUC	AUPR	F1	MCC	ACC	Recall
SMGCN(sf+cs)	0.962	0.844	0.840	0.838	0.979	0.792
SMGCN(sf+rwr)	0.979	0.947	0.922	0.921	0.986	0.894
SMGCN(sf)	0.968	0.875	0.853	0.852	0.981	0.798
RSGCN	**0.985**	**0.952**	**0.927**	**0.926**	**0.987**	**0.906**

2.4　主成分分析(PCA)

实验进行了使用 PCA 前后的消融实验.如表 4 所示，模型在所有评价指标上均优于没有使用 PCA，这是因为 PCA 有助于消除噪声，获取更加显著的特征.

表 4　主成分分析消融实验结果

Tab. 4　Principal component analysis Scientific control results

Model	AUC	AUPR	F1	MCC	ACC	Recall
No PCA	0.978	0.938	0.922	0.921	0.986	0.897
RSGCN	**0.985**	**0.952**	**0.927**	**0.926**	**0.987**	**0.906**

PCA 中的参数 k，表示维度降为原来的矩阵列数目的 k 倍.将 k 取值范围设置为 0.1～0.9.如图 3 所示，当 k 取 0.9 时，达到了最优实验性能.

2.5　增强注意力机制

为了使相似的药物(或靶标)节点在特征空间中更好地与邻居聚合.通过增强注意力机制获得节点间的注意力权重.

为了验证增强注意力机制的作用，同样设置了一组对照实验，如图 4 所示，模型在各种评价指标上均优于未使用增强注意力机制的结果，这是因为增强注意力机制能够获得更多邻居节点与节点间的影响.

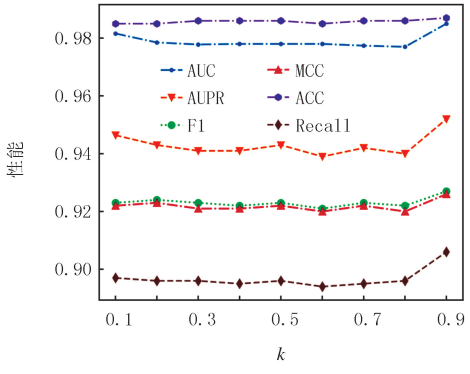

图3　参数k对模型的影响

Fig. 3　The influences of parameter k on the model

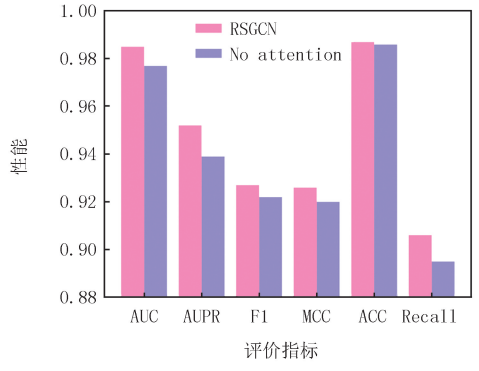

图4　采用增强注意力机制前后的评价指标对比图

Fig. 4　Comparison of evaluation indicators before and after using the attention enhancement mechanism

2.6　与其他方法的比较

本实验与 DTIMGNN[24]、DTICNN[25]、NRLMF[26]、MHADTI[9] 和 BICTR[27] 模型进行了比较.对比结果的 AUC 和 AUPR 如表 5 和表 6 所示，在数据集 Es、IC 和 NRs 上，本模型的性能均优于其他方法.在数据集 GPCR，本模型的性能优于绝大多数其他方法.在数据集 Es 中的 AUC 分别比 DTIMGNN 高 7.2%、DTICNN 高 5.2%、NRLMF 高 4.5%和 MHADTI 高 4.1%.通过上述分析表明，模型在使用了多重相似性和增强注意力机制后，能够更好地捕捉网络节点信息，充分利用节点间直接或间接的特征，实现了更好的预测性能.

表 5　与其他方法的 AUC 结果对比

Tab. 5　Comparison of AUC results with other methods

数据集	DTIMGNN[24]	DTICNN[25]	NRLMF[26]	MHADTI[9]	BICTR[27]	RSGCN
Es	0.913	0.934	0.940	0.944	0.973	**0.985**
IC	0.888	0.892	0.949	0.917	0.968	**0.980**
GPCR	0.863	0.854	0.861	0.881	0.951	**0.940**
NRs	0.860	0.733	0.747	0.910	0.905	**0.950**

3　总　结

本文提出了一种新颖的基于多重相似性和增强注意力机制的图卷积神经网络模型预测药物与靶标相互

作用(RSGCN).首先构建多重相似性矩阵,将药物相似性矩阵和靶标相似性矩阵采用重启随机游走和余弦相似获得药物和靶标的多重相似性特征矩阵,并用 PCA 实现数据降噪声;然后在异构网络中采用 GCN 获取药物和靶标的低维表示,运用增强注意力机制获得节点间的注意力,并通过消融实验和对照实验,验证了多重相似性和增强注意力模块的作用.在多个数据集上实验结果表明,本模型达到了较好的性能,提升了 DTI 预测性能和模型的泛化能力.

表 6 与其他方法的 AUPR 结果对比

Tab. 6 Comparison of AUPR results with other methods

Datasets	DTIMGNN[24]	DTICNN[25]	NRLMF[26]	MHADTI[9]	BICTR[27]	RSGCN
Es	0.849	0.934	0.795	0.937	0.785	**0.952**
IC	0.832	0.889	0.798	0.895	0.808	**0.953**
GPCR	0.855	0.851	0.406	0.860	0.523	**0.887**
NRs	0.843	0.745	0.485	0.915	0.555	**0.844**

参 考 文 献

[1] 彭利红,田雄飞,周立前.基于一致性学习预测药物-靶标相互作用[J].湖南工业大学学报,2020,34(6):27-33.
PENG L H,TIAN X F,ZHOU L Q.Prediction of drug-target interactions based on consistency learning[J].Journal of Hunan University of Technology,2020,34(6):27-33.

[2] 任浩然,邓博韬,李建华,等.药物-靶标相互作用预测平台设计与实现[J].现代计算机,2023,29(5):104-108.
REN H R,DENG B T,LI J H,et al.Design and implementation of drug-target interaction prediction platform[J].Modern Computer,2023,29(5):104-108.

[3] 王红梅,郭真俊,张丽杰.基于图神经网络的药物-靶标相互作用预测研究[J].长春工业大学学报,2021,42(4):318-325.
WANG H M,GUO Z J,ZHANG L J.Drug-target interaction prediction based on graph neural network[J].Journal of Changchun University of Technology,2021,42(4):318-325.

[4] YU G X,WANG Y H,WANG J,et al.Attributed heterogeneous network fusion via collaborative matrix tri-factorization[J].Information Fusion,2020,63:153-165.

[5] LUO Y N,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].Nature Communications,2017,8:573.

[6] ZHENG X D,DING H,MAMITSUKA H,et al.Collaborative matrix factorization with multiple similarities for predicting drug-target interactions[C]//Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining.Chicago:ACM,2013.

[7] 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,2019,35(1):104-111.

[8] SHANG Y F,YE X C,FUTAMURA Y,et al.Multiview network embedding for drug-target Interactions prediction by consistent and complementary information preserving[J].Briefings in Bioinformatics,2022,23(3):bbac059.

[9] TIAN Z,PENG X Y,FANG H C,et al.MHADTI:predicting drug - target interactions via multiview heterogeneous information network embedding with hierarchical attention mechanisms[J].Briefings in Bioinformatics,2022,23(6):bbac434.

[10] SHAO K H,ZHANG Y H,WEN Y Q,et al.DTI-HETA:prediction of drug-target interactions based on GCN and GAT on heterogeneous graph[J].Briefings in Bioinformatics,2022,23(3):bbac109.

[11] NASCIMENTO A C A,PRUDÊNCIO R B C,COSTA I G.A multiple kernel learning algorithm for drug-target interaction prediction[J].BMC Bioinformatics,2016,17:46.

[12] ZHANG L L,OUYANG C P,HU F Y,et al.Relational topology-based heterogeneous network embedding for predicting drug-target interactions[J].Data Intelligence,2023,5(2):475-493.

[13] LU Z L,WANG Y K,ZENG M,et al.HNEDTI:Prediction of drug-target interaction based on heterogeneous network embedding[C]//2019 IEEE International Conference on Bioinformatics and Biomedicine(BIBM).[S.l.]:IEEE,2019:211-214.

[14] 廖懿鸣,欧阳纯萍,刘永彬,等.基于异质信息网络元路径的药物-靶标相互作用预测模型[J].北京大学学报(自然科学版),2022,58(1):37-44.
LIAO Y M,OUYANG C P,LIU Y B,et al.Drug-target interactions prediction based on meta-path of heterogeneous information network[J].Acta Scientiarum Naturalium Universitatis Pekinensis,2022,58(1):37-44.

[15] 彭利红,刘海燕,任日丽,等.基于多标记学习预测药物-靶标相互作用[J].计算机工程与应用,2017,53(15):260-265.
PENG L H,LIU H Y,REN R L,et al.Predicting drug-target interactions with multi-label learning[J].Computer Engineering and Applications,2017,53(15):260-265.

[16] HATTORI M，OKUNO Y，GOTO S，et al.Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways[J].Journal of the American Chemical Society，2003，125(39)：11853-11865.
[17] SMITH T F，WATERMAN M S.Identification of common molecular subsequences[J].Journal of Molecular Biology，1981，147(1)：195-197.
[18] MA Z H，KUANG Z F，DENG L.CRPGCN：predicting circRNA-disease associations using graph convolutional network based on heterogeneous network[J].BMC Bioinformatics，2021，22(1)：551.
[19] HUANG D，AN J Y，ZHANG L，et al.Computational method using heterogeneous graph convolutional network model combined with reinforcement layer for MiRNA-disease association prediction[J].BMC Bioinformatics，2022，23(1)：299.
[20] LAW V，KNOX C，DJOUMBOU Y，et al.DrugBank 4.0：shedding new light on drug metabolism[J].Nucleic Acids Research，2014，42(Database issue)：D1091-D1097.
[21] SCHOMBURG I，CHANG A，PLACZEK S，et al.BRENDA in 2013：integrated reactions，kinetic data，enzyme function data，improved disease classification：new options and contents in BRENDA[J].Nucleic Acids Research，2013，41(Database issue)：D764-D772.
[22] KANEHISA M，GOTO S，HATTORI M，et al.From genomics to chemical genomics：new developments in KEGG[J].Nucleic Acids Research，2006，34(Database issue)：D354-D357.
[23] HECKER N，AHMED J，VON EICHBORN J，et al.SuperTarget goes quantitative：update on drug-target interactions[J].Nucleic Acids Research，2012，40(Database issue)：D1113-D1117.
[24] LI Y，QIAO G Y，WANG K Q，et al.Drug-target interaction predication via multi-channel graph neural networks[J].Briefings in Bioinformatics，2022，23(1)：bbab346.
[25] PENG J J，LI J Y，SHANG X Q.A learning-based method for drug-target interaction prediction based on feature representation learning and deep neural network[J].BMC Bioinformatics，2020，21(Suppl 13)：394.
[26] 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)：e1004760.
[27] PLIAKOS K，VENS C.Drug-target interaction prediction with tree-ensemble learning and output space reconstruction[J].BMC bioinformatics，2020，21(2)：1-11.

Prediction of drug-target interaction based on multiple similarity and enhanced attention mechanisms on graph convolution neural network

Wang Wei[1a,b]，Yu Mengxue[1a]，Sun Bin[1a]，Wan Shitong[1a]，Liu Dong[1a,b]，
Zhou Yun[1a,b]，Zhang Hongjung[2]，Wang Xianfang[3]

(1. a. College of Computer and Information Engineering；b. Key Laboratory of Artificial Intelligence and Personalized Learning in Education of Henan Province，Henan Normal University，Xinxiang 453007，China；2. Hebi Instiute of Engineering and Technology，Henan Polytechnic University，Hebi 458030，China；3. College of Computer Science and Technology Engineering，Henan Institute of Technology，Xinxiang 453000，China)

Abstract：In the research of new drug discovery and drug repositioning，it is important to search for the interactions between drugs and targets. In this study，we propose a graph convolutional network model based on multiple similarities and enhanced attention mechanism to predict drug-target interactions(RSGCN) for the drug-target interaction network. Firstly，we propose to use multiple similarities to optimize the original feature vectors of drugs and targets，capture network structure features，and fully utilize direct or indirect relationships between nodes. Then，we reduce the impact of similarity noise on experimental results through PCA dimensionality reduction. Finally，we use GCN to obtain node embedding representations and incorporate an attention-based enhanced layer to obtain attention weights between nodes，which efficiently predicts interactions between drugs and targets. The experiments use a public gold standard dataset，and the experimental results indicate that the RSGCN model has good performance.

Keywords：graph convolution neural network；multiple similarity；PCA；enhanced attention mechanisms；drug-target interaction

[责任编校　陈留院　杨浦]