太原理工大学学报
Journal of Taiyuan University of Technology
ISSN 1007-9432,CN 14-1220/N

《太原理工大学学报》网络首发论文

题目：跨粒度子图对比学习与注意力融合的药物-基因关系预测

作者：胡冬冬，彭杨，谭暑秋，朱小飞

网络首发日期：2024-03-19

引用格式：胡冬冬，彭杨，谭暑秋，朱小飞．跨粒度子图对比学习与注意力融合的药物-基因关系预测[J/OL]．太原理工大学学报.
https://link.cnki.net/urlid/14.1220.N.20240319.1500.002

网络首发时间：2024-03-19 20:59:19
网络首发地址：https://link.cnki.net/urlid/14.1220.N.20240319.1500.002

太原理工大学学报
Journal of Taiyuan University of Technology

跨粒度子图对比学习与注意力融合的药物-基因关系预测

胡冬冬，彭 杨，谭暑秋，朱小飞
(1. 重庆理工大学 计算机科学与工程学院, 重庆 400054)

摘　要：阐明药物和基因之间的相互联系是药物开发中的一个重要课题。目前，基于随机游走算法的图神经网络方法在解决药物与基因交互关系识别上已经取得了不错的效果，但是当前的方法，单一子图的方法往往容易忽略掉全局图的信息，不能够很好的聚合节点的信息，同时，药物和基因的节点表示采用简单的融合方式，不能够有效的利用节点表示的信息，用于交互关系的分类。针对上述问题，本文提出了跨粒度对比学习与注意力融合的药物-基因交互关系预测方法，一方面采用跨粒度的对比学习方法，得到远距离和近距离的节点信息，同时采用对比学习的结构增加对药物和基因节点的区分。另一方面利用注意力融合机制，充分挖掘节点中隐含的信息，将远近距离信息进行注意力融合。在 2 个真实数据集上的实验结果表明该模型比基线模型具有更好的分类效果。

关键词：对比学习；图表示学习；关系图神经网络；注意力机制；基因-药物关系预测
中图分类号：TP391

Drug-Gene Interaction Prediction Method through Cross Granular Subgraph Contrastive Learning and Attention Mechanism Fusion

HU Dongdong，PENG Yang，TAN Shuqiu，ZHU Xiaofei
(1. College of Computer Science and Engineering, Chongqing University of Technology, Chongqing 400054,China)

Abstract: Clarifying the interconnections between drugs and genes is an important topic in drug development. At present, the graph neural network method based on the random walk algorithm has achieved great results in identifying drug-gene interaction relationships. However, existing methods using single graph neural network modeling, can't aggregate the information of neighbor nodes so well. In addition, most methods use a simple method for the node representation of drugs and genes fusing, which can't effectively use the information represented by nodes for the classification of interaction relationships. To address the above issues, this article proposes a cross granularity contrastive learning and attention fusion method for predicting drug gene interaction relationships. On the one hand, a cross granularity contrastive learning method is adopted to obtain node information from both distant and close distances, on the other hand, by utilizing attention fusion mechanisms, hidden information in nodes can be fully mined, and attention fusion can be performed on distance information. The experimental results on two real datasets show that our model has better classification performance than the baseline.

Keywords: contrastive learning; graph representation learning; relational graph neural network; attention mechanism; Gene-Drug interaction prediction

药物-基因交互关系预测（DGI）[1]的目的在于从现有的交互关系出发，挖掘出潜在的药物和基因之间的交互关系[2]，药物和基因之间的相互作用的研究对药物发现和药物重定位至关重要，虽然实验的方法仍然是最可靠的方法，但是因其高昂的实验费用，不能做到逐一对药物和基因进行实验[3]，因此计算预测的方法逐渐成为大家的共识。

基金项目：国家自然科学基金项目(No.62141201)；重庆市自然科学基金项目(CSTB2022NSCQ-MSX1672)；重庆市教育委员会科学技术研究计划重大项目(No.KJZD-M202201102)。
作者简介：胡冬冬(1998-)，男，硕士，研究方向为图神经网络、自然语言处理;彭阳(1996-)，男，硕士，研究方向为图神经网络、自然语言处理;谭暑秋（1982-)，女，博士，讲师，计算机学会(CCF)会员，研究方向为机器学习和图像处理;朱小飞(1979-)，男(通讯作者)，博士，教授，计算机学会(CCF)高级会员，研究方向为自然语言处理、数据挖掘与信息检索(zxf@cqut.edu.cn).

当前的研究工作，大部分主要集中在预测药物和基因之间是否能够结合的相互作用上[4-5]，用来发现对蛋白质靶点具有高亲和力的药物。但药物和基因之间的作用关系可能存在多种关系[7]，例如部分药物可以刺激基因的表达（激动剂），而部分药物相反，能够抑制基因的表达（拮抗剂），因此仅仅判断药物和基因是否能够结合， Rao 等人[8]将作用关系进行了进一步的划分，为揭示药物和基因更深层次的关系打下了基础。

当前对于基因和药物交互关系的预测主要是基于神经网络的方法，一部分基于网络的方法是高度依赖于人工标注的特征[9-10]，但基于人工标注往往需要投入巨大的成本，大部分情况下高质量的特征是难以得到的，另一部分的不依赖特征的方法[11-12]，往往需要利用全局的信息来进行训练，因此出现了基于子图的的方法，CoSMIG[8]模型提出了交互式子图的学习方式，交互式的从边和节点之间进行卷积。但是该方法有两个主要缺点：（1）基于单一子图的方法往往容易忽略掉全局图的信息，导致预测效果不好。（2）同时现有的方法对于待预测药物/基因节点表示的融合，仅采用简单拼接方式，不能够充分的挖掘节点表示中隐含的信息。

针对以上两个问题，本文提出了一个基于跨粒度子图对比学习和注意力融合方法。我们采用跨粒度子图方并在此基础上将子图对比学习方法引入 DGI 问题中，并对该方法进行适应性改进，用来解决单一子图易丢失信息的问题。同时采用注意力融合的方式，代替已有方法中拼接方式，对子图的远近距离信息进行注意力融合。主要思想是，首先利用随机游走算法获得不同粒度的子图[13]，为了避免采用子图的方法而忽略全局图中的信息，我们采用跨粒度的子图，来对全局图进行信息的补偿，然后，我们对不同粒度的子图分别用关系图卷积网络[14]对不同关系的邻居节点进行卷积，之后对同一 batch 内的中心节点进行跨粒度的对比学习，最后对中心药物节点表示和中心基因节点表示进行交叉注意力的融合，用来预测中心药物节点表示和中心基因节点表示。

本文的主要贡献分为以下三个方面：

a) 针对现有基于单一子图的方法容易忽略掉全局图信息，本文提出跨粒度子图学习方法，利用多粒度学习被忽略的信息，并将节点分类领域的子图对比学习方法引入DGI问题中并对其进行适应性改进。
b) 针对现有的方法采用简单拼接融合特征的方式，不能够充分的挖掘节点表示中隐含的信息，充分挖掘节点中隐含的信息，将远近距离信息进行注意力融合。
c) 本文提出的方法在 2 个基准数据集上进行了大量的实验，实验结果证明该方法的有效性。

1 相关工作

1.1 图卷积神经网络

图卷积神经网络[15](GCN)的核心思想是将图中的节点表示为一个向量，并通过在相邻节点之间进行卷积操作来更新这些节点的表示。现有的方法主要是基于神经网络的方法，一部分方法从手动构造的特征出发，构建深度学习模型框架对药物和基因表示进行学习，比如 DeepDTA[9]模型使用卷积神经网络(CNN)[16]从原始蛋白质序列和药物的化学式学习特征表示。

受推荐系统的启发，部分学者将药物和基因看成用户和商品，采用推荐中的基于矩阵补全方法[17]来进行 DGI 预测。例如，MC[18]模型将药物和基因的交互矩阵采用矩阵分解方法补全矩阵中缺失的交互关系。F-EAE[19]和 GRALS[20]等模型在 MC 方法的基础上，对矩阵分解的方法进行了优化。使交互关系的预测更加的准确。

由于图卷积神经网络(GCN)能够很好的处理结构数据和学习局部特征，一部分学者开始将图卷积神经网络引入到药物和基因关系预测中。例如 GC-MC[21]模型将交互矩阵中的交互关系转化为图卷积神经网络中的链路预测问题。sRGCNN[22]方法在节点信息聚合时进一步的考虑节点的相似性，减少了噪声的引入。但是 GC-MC 模型和 sRGCNN 模型都是从整个交互图中进行学习，计算的过程比较复杂。

1.2 多粒度子图的数据增强

图数据增强[23]是一种长见的用于提高模型性能的技术，它通过对原始图数据进行变换和扩充来生成新的训练样本。常见的图数据增强方法有对节点特征的扰动、对边的扰动以及子图抽取等[24]，由于在 DGI 问题中高质量的节点特征难以获得，同时从二分图采集的子图从中心节点发散，因此对关键边的删除可能会造成信息聚合的中断，而现有的基于子图抽取的方式则可以解决信息聚合中断的问题[8]，但是采用单一的子图抽取方式，在子图稀疏的条件下往往不能很好的得到中心药物/基因节点的表示。

采用多粒度子图的方式已经在链路预测、节点分类等[25]问题上取得了模型上效果的提升，因此本文将在 DGI 问题中引入跨粒度子图的方法。

1.3 图对比学习

对比学习(Contrastive Learning)[26]是一种无监督学习方法，通过比较同一组数据中的样本来学习数据的内在结构和特征表示。对比学习的目的是让相似的数据点靠近，将不相似的数据点分开，从而可以有效地捕捉数据的特征表示。Tao[27]等人采用超图对比学习的方法，构建超边从全局和局部视角进行对比学习。Liu 等人[25]将基于子图的对比学习应用于图结构的表示学习中，取得了良好的效果。但其对比学习方法中采用 readout 函数，获得子图的表示，但这并不适合于 DGI 问题，同时该方法中，并没有考虑同一 batch 中类内负样本对作为损失。本文将对上述方法做出改进，在 DGI 问题中引入基于子图的对比学习方法用以加强节点表示。

2 问题定义

本文的研究内容可以看成是二分图交互关系的分类。现在设 $\mathcal{G}=(V,E)$，$\mathcal{G}$ 表示二分图，其中 V 代表节点集合，由互不相交的两个集合构成，这两个集合分别为含有 m 个药物节点集合 $D=(d_1,d_2,...,d_m)$ 和含有 n 个基因节点的集合 $G=(g_1,g_2,...,g_n)$，E 表示为边的集合，$e_{ij}\in E$，表示边的两端节点分属于两个不同的节点集合，其中 $i\in G, j\in D$。为了对边的类型进一步的区分，我们用 $R=(r_1,r_2...r_k)$ 表示为基因和药物的交互类型，图中的每一条边被分为 R 中的某一种关系，可以用映射函数 $\phi(e_{ij})\in R$ 表示，本文中我们将待预测关系的药物和基因节点定义为中心药物/基因节点，分别用 C_0, C_1 表示。本文的目标是从已知的药物节点 D 和基因节点 G 构成的二分图 G 出发，预测出未知的边 $e_{ij}\notin E$ 所属于的类型 R，其中 $i\in G, j\in D$。

3 方法与模型

3.1 基于重启随机游走的子图抽取

对于药物和和基因的交互图如图 1 所示，药物和基因的交互关系网络复杂，一种药物可以和多种基因进行交互，同时一种基因同时也可以和多种药物产生交互。为了减轻图神经网络计算的时间复杂度，我们采用子图的结构，代替利用整图结构进行卷积操作，同时我们采用多跳子图的方式，尽可能的利用整图的信息。具体而言，对于给定需要预测交互类型的中心药物节点 d_i 和中心基因的节点 g_j，分别从 d_i 和 g_j 出发采用不同跳数的重启随机游走算法采集出不同跳数的子图，重启随机游走算法如公式(1)所示：

$$p=c\mathrm{A}\mathrm{D}^{-1}p+(1-c)e, \tag{1}$$

其中 c 表示的是重启概率，控制着随机游走下一步是随机选取的邻居节点还是到重启的节点，p 为列向量，p_i 代表选取节点 i 的概率，A 代表邻接矩阵，D 为 A 的度矩阵。e 为重启向量，其中 $e_i=1$ 表示为出发节点，$e_i=0$ 表示为非出发节点。

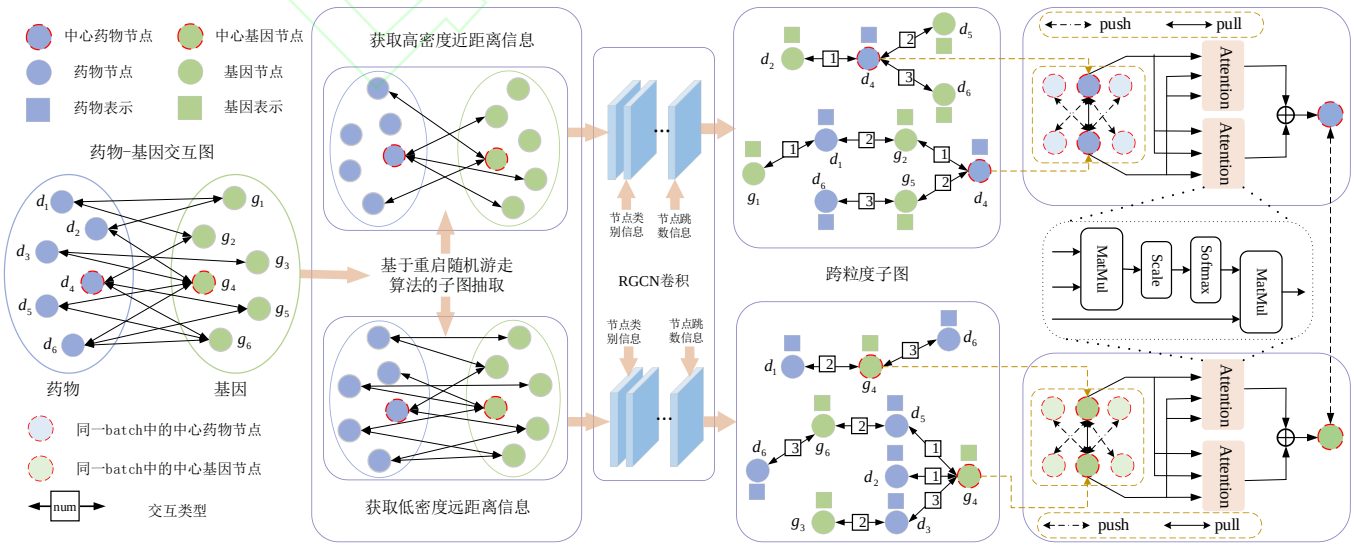

图 1 跨粒度子图对比学习与注意力融合模型结构图

Fig.1 Architecture of cross granularity subgraph contrastive learning and attention fusion model

采用重启随机游走算法，可以得到从药物/基因中心结点出发，到各个节点的概率，在每一跳中选取概率较大的前 p 个节点作为子图的节点，由重启随机游走的线路和抽取到的节点，可以构成子图。近距离的节点往往具有高密度的信息帮助模型预测中心药物和基因节点的信息，同时远距离的节点能加入辅助的节点信息，因此我们采集不同跳数采集得到不同粒度的子图。由于 DGI 问题是一个二部图，从中心节点出发奇数跳得到子图的两端为药物-基因节点。因此本文中采用的跳数为奇数，同时出于时间复杂度的考虑，我们采用的跳数分别为一跳和三跳，得到的子图如图 1 所示。

3.2 节点编码与聚合

该模块由两部分内容组成，节点编码和节点信息聚合。节点编码部分将节点的类别信息和跳数信息建模得到初始节点嵌入表示。节点信息聚合部分利用 RGCN 对节点的信息进行聚合。

3.2.1节点编码

在得到子图的表示后，首先需要得到子图节点的初始嵌入表示，但由于现实中初始节点特征的获取是困难的，而采用随机初始化特征，往往不能够很好的将不同类型的节点和跳数信息进行区分，因此我们采用了 Rao 等人[8]的初始化方法，将药物和节点信息以及跳数信息映射为一个实数，这一个实数中包含了节点信息和跳数信息，再将映射的实数转换为 *One-Hot* 向量作为节点的初始化表示。因此初始化的信息中将药物和和节点信息进行区分同时将节点的跳数信息很好的融合到了初始化节点信息中，节点编码过程可以表示为：

$$f(i,j)=2i+j, \tag{2}$$

$$h^{'}=One\text{-}Hot(f(i,j)), \tag{3}$$

其中 $i\in\{0,1\}$，0 代表药物，1 代表基因。$j=\{0,1,2\cdots n\}$ 代表节点经过的跳数。$h^{'}$ 为图中任意节点的初始表示。

3.2.2节点信息的聚合

由于在图中的边存在不同的关系，为了更好的聚合不同类型的边上的节点，我们采用 RGCN[29]来对不同关系下的节点信息进行聚合：

$$h_u^{(l+1)}=\sigma\left(\sum_{r\in R}\sum_{v\in N_u^r}\frac{1}{c_{ur}}W_r^{(l)}h_v^{(l)}+W_o^{(l)}h_u^{(l)}\right), \tag{4}$$

其中 $\sigma(\cdot)$ 代表激活函数，N_u^r 表示节点 u 在关系 r 下邻居节点的集合，c_{ur} 表示的正则化常数，l 表示卷积的层数，h_u 代表图中的节点 u 的表示，W_r,W_o 分别表示可训练参数，$h_u^{(0)}=h^{'}$。

激活函数中的第一项是对不同关系下的邻居节点进行聚合，第二项是对上一层节点表示的残差连接。因此图中节点的表示将包含不同关系下的信息，同时保留上一层中节点的信息。

3.2.3跨粒度对比学习

对于不同的子图，我们利用 RGCN 得到不同粒度子图节点的表示，传统方法中常采用 readout 函数将子图所有的节点压缩为子图的表示，再对子图的表示进行对比学习，但在 DGI 问题中，我们只关心中心药物/基因节点，因此我们将中心药物/基因节点取出，避免其他节点的干扰。在对比学习中正样本为同一 batch 中不同粒度下对应的中心节点，与现有的子图对比学习中选取的负样本不同，同一 batch 内不同的中心药物节点表示由于同属于药物类，所以中心药物节点表示依然可能相似。因此本文中不仅考虑了类间的负样本对，同时也将类内负样本对作为损失的一部分。单个节点的损失计算如公式(5)所示：

$$\mathcal{L}\left(h_i^{v_0}\right)=-log\frac{e^{\theta\left(h_i^{v_0},h_i^{v_1}\right)/\tau}}{e^{\theta\left(h_i^{v_0},h_i^{v_1}\right)/\tau}+\sum_{j\neq i}\left(e^{\theta\left(h_i^{v_0},h_j^{v_0}\right)/\tau}+e^{\theta\left(h_i^{v_0},h_j^{v_1}\right)/\tau}\right)}, \tag{5}$$

其中 v_0,v_1 表示第 0 个子图和第 1 个子图，$h_i^{v_0}$，$h_i^{v_1}$ 分别表示节点 i 在第 0 个和第 1 个子图中的节点表示，τ 为温度系数，$\theta(\cdot)$ 表示相似性的度量，分母中的第二项为类内负样本对，第三项为类间负样本对，这里我们采用内积的方式作为相似性度量。上面公式给出了第 i 个节点在视图 v_0 中的损失函数，同理我们可以得到节点 i 在视图 v_1 中的损失函数，我们将所有节点的损失的平均值作为两个视图的最终损失，如公式(6)所示：

$$\mathcal{L}_1=\mathcal{L}\left(H^{v_0},H^{v_1}\right)=\frac{1}{2N}\sum_{i=1}^{N}[\mathcal{L}(h_i^{v_0})+\mathcal{L}(h_i^{v_1})] \tag{6}$$

3.3 注意力融合机制

得到不同子图的表示后，因为仅需要对待预测的节点之间的关系进行预测，所以我们仅取出不同子图中的中心节点表示参与后面的计算。不同子图中的中心节点代表了融合了不同跳数子图中聚合的信息，拼接融合的方式往往难以挖掘出含有不同信息中心节点隐含信息。为了更好的获得远近距离信息，本文采用交叉注意力机制，获得不同子图中的中心基因节点信息和中心药物节点表示：

$$Z_0 = softmax(\frac{C_1 C_0^T}{\sqrt{d_T}})C_0,$$
$$Z_1 = softmax(\frac{C_0 C_1^T}{\sqrt{d_T}})C_1, \quad (7)$$

其中 C_0, C_1 为中心药物/基因节点表示，d_T 为注意力缩放系数，Z_0, Z_1 为经过交叉注意力后的中心药物/基因节点表示。

最后将中心药物和基因节点表示拼接，用于对药物和基因交互关系的预测：

$$p = softmax(concat(z_0^d, z_1^g)), \quad (8)$$

上式中 p 为最终在每一类的概率分布。$concat(\cdot)$ 表示向量的拼接操作。z_0^d 表示待预测的药物节点表示，z_1^g 表示待预测的基因节点表示。通过 *softmax* 激活函数得到中心药物节点与中心基因节点表示在各个关系上的概率分布。

3.4 损失函数

根据最终的预测的概率和真实值之间计算交叉熵损失：

$$\mathcal{L}_2 = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K} y_{ij}\log(p_{ij}), \quad (9)$$

其中 N 代表样本的个数，K 代表关系预测中关系的类别数。最终的总损失为对比学习损失 $\mathcal{L}_1$ 和交叉熵损失 $\mathcal{L}_2$ 的加和：

$$\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2. \quad (10)$$

本文模型算法伪代码如下所示：

算法 1：本文模型算法流程图

Input: 药物－基因交互图 G
Output: 药物－基因在各个关系上的概率分布 p
for each epoch **do**
  %子图的抽取
  在 G 中，从中心药物节点出发得到一跳和三跳子图 G_{sub_1}, G_{sub_3};
  %节点信息的聚合;
  $G_{sub\_1}, G_{sub\_3} = RGCN(G_{sub\_1}, G_{sub\_3})$;
  从子图中取出中心药物节点 C_0, C_1;
  计算对比学习损失: $\mathcal{L}_1 = Contrast(C_0, C_1)$;
  %节点信息的融合
  注意力融合节点表示 $Z_0 = Attention(C_0, C_1)$;
  同理可以得到基因节点表示 Z_1;
  %关系预测
  得到概率分布: $p = softmax(concat(z_0^d, z_1^g))$;
  计算交叉熵损失: $\mathcal{L}_2 = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K} y_{ij}\log(p_{ij})$;
  计算最终的损失 $\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2$, 反向传播并更新参数;
end for
return p

4 实验

4.1 数据集

为了评估本文提出模型的有效性，我们使用 Rao[8]等人使用的 DrugBank 数据集和 DGIdb 数据集。DrugBank 数据集：这是从 DrugBank 数据库中获得的药物转录组学数据集。其中药物化合物代谢引起的基因表达上升/下降代表了药物与基因之间的相互作用。因此这个数据集有两种类型的交互，即增加和减少。DGIdb 数据集：该数据从 DGIdb 数据库中获得，包含超过 1664 个基因和 1185 种药物，涉及超过 11366 种药物-基因相互作用和 14 种类型的关系。关于数据集的具体统计，如表 1 所示。

从表 1 中可以看出，DGIdb 数据集中药物和基因节点的数量相差不大，交互的种类比较丰富有 14 种。而 DrugBank 数据集中基因节点的数量远大于药物节点的数量，而且交互关系只有 2 种，但总交互数量却大于 DGIdb 数据集的交互数量。因此从数据集分析中可知 DrugBank 数据集交互关系分类应当比 DGIdb 更难。

表 1 药物-基因交互关系数据集统计

Table 1 Statistics of Drug Gene Interaction Dataset

Dataset	DrugBank	DGIdb
Number of Drug	425	1185
Number of Gene	11284	1664
Interactions	80924	11366
Interaction types	2	14

4.2 基准方法

本文基准的方法主要包括两类的方法：一类是基于矩阵补全类方法（MF-based），另一类是基于图神经网络类方法（GNN-based），本文将与这两类基准方法进行比较，验证模型的有效性。基准方法的结果均来自 CoSMIG[8]。

矩阵补全类方法：

MC[18]：MC 方法将含有缺失值的交互矩阵利用矩阵分解的方法恢复出完整的交互矩阵。

GRALS[19]：GRALS 方法交替进行最小二乘估计将矩阵分解的过程进行优化。

F-EAE[20]：F-EAE 模型提出了利用可交换矩阵层的方法来对补全交互矩阵进行优化。

神经网络类方法：

GC-MC[21]：GC-MC 模型将图神经网络引入到交互图中关系的预测，把交互矩阵的矩阵补全问题转化为图神经网络中节点的链路预测

问题。

sRGCNN[22]：sRGCNN 方法在 GC-MC 方法的基础上，进一步考虑了交互节点的相似性。只在相似的节点之间进行信息的聚合，避免了部分噪声的引入。

PinSage[29]：PinSage 模型采用与 GC-MC 类似的卷积定义，通过分析用户与商品过往的交互信息来预测用户可能感兴趣的商品。

IGMC[30]：IGMC 模型改进了 GC-MC 方法中神经网络的初始化方法，将节点的类别和跳数信息作为初始化信息，缓解了神经网络依赖初始特征的问题。

CoSMIG[8]：CoSMIG 模型采用交互式子图的方式对节点进行学习，将节点和边的信息通过注意力交叉卷积的方式更新节点的表示。

4.3 实验设置

实验中使用 pytorch geometric[31]框架来实现关系图卷积神经网络，根据 DGIdb 数据集上交叉验证结果调整模型的超参数，并在另外的数据集上也使用相同的参数。具体而言，我们设置多粒度随机游走分别为一跳和三跳，关于其他参数如学习率和嵌入维度等对实验的影响，可以参考 6.7 部分。模型在 Nvidia GeForce RTX 2080 GPU 上训练 80 个 epoch。作为一个多关系链路预测问题，我们采用准确率（ACC）作为评价模型预测药物-基因相互作用关系的指标。我们对模型进行了五次训练，并将五次实验结果进行了平均，作为模型最终预测得分。

4.4 总体实验

总体实验在两个数据集上的最终结果如表 2 所示，下面对实验结果进行分析：

其中最优结果加粗标出，feature 列表示的是模型是否使用了人工特征对模型进行初始化，从最后实验结果来看本文的模型即使在没有使用人工初始化特征的情况下也可以达到较好的效果。method 列表示不同方法之间的比较。实验中我们采用分类的准确率（ACC）作为模型最终的评价指标。Test ACC 表示在测试集上的准确率。

从表中数据可以看出本文提出的模型整体都优于基线模型，具体而言，在 DrugBank 数据集的测试集上，本文提出的模型比最好的基线模型提升了 1.4 个百分点，在另外一个 DGIdb 数据集上，本文的模型在测试集的准确率要比最好的基线模型提升 2.4 个百分点。从最终的结果可以看出本文模型的有效性。

表 2 本文模型在准确率（ACC）上与其他基线模型的对比。

Table 2 Comparison of the accuracy (ACC) of this model with other baseline models.

	methonds	DrugBank	DGIdb
		Test ACC /%	Test ACC /%
MF-based	MC	0.518±0.013	0.559±0.009
	GRALS	0.532±0.021	0.578±0.016
	F-EAE	0.566±0.004	0.623±0.003
GNN-based	GC-MC	0.586±0.008	0.601±0.005
	sRGCNN	0.602±0.010	0.689±0.007
	PinSage	0.629±0.004	0.713±0.005
	IGMC	0.634±0.003	0.803±0.006
Subgraph-based	CoSMIG	0.678±0.003	0.852±0.012
	ours	**0.694±0.043**	**0.876±0.032**

本文的模型在 DGIdb 数据集上分类提升的表现要比 DrugBank 数据集上要好 1 个百分点。造成这一现象的原因可能是 DrugBank 和 DGIbd 交互节点的数量和交互关系类别的数量存在较大的差异。在 DGIdb 数据集中药物和基因的交互节点数量相差不大，但 DrugBank 数据中基因节点要比药物节点多 10859 个，同时 DGIdb 平均每个节点连接边的数量约为 4，而 DrugBank 的数量平均每个节点连接边的数量约为 7，DGIdb 平均节点连接边的数量要小于 DGIdb。因此模型从 DGIdb 数据中学习到更加的均衡的节点信息，所以在 DGIdb 数据集上的提升效果会更加的明显。

4.5 消融实验

为了说明本文提出的跨粒度子图对比学习模块和注意力融合模块的有效性，下面就这两个模块分别进行消融实验。如图 2 所示为模型消融实验的结果。

a) w multi-granularity 表示仅使用单个子图进行卷积。
b) w concat 表示将中心节点表示采用拼接的方式作为最终的表示。
c) w gate 表示将中心节点表示采用门控机制进行融合作为最终的表示。

从图 2 中可以看出在仅使用单粒度子图进行表示学习，模型的效果在验证集和测试集上都明显下降，在 DrugBank 测试集上下降了 2.7%，在 DGIdb 测试集上下降了 3.7%，因此说明相比于单粒度子图而言，跨粒度子图的方式能够有效的聚合邻居节点的信息。

对使用拼接融合和门控融合这两种不同的方式进行比较，可以发现这两种融合方式对于模型的最终的效果影响相差不大，拼接融合的方式略好于门控的方式。对注意力融合方式和

拼接融合的方式进行比较，可以看出注意力方式要比拼接融合在两个数据集上好，相对于拼接融合注意力，注意力融合方式在 DrugBank 和 DGIdb 数据集上分别提升了 0.2 个百分点和 1.9 个百分点，相对于拼接融合注意力，注意力融合方式在 DrugBank 和 DGIdb 数据集上分别提升了 0.8 个百分点和 2 个百分点，因此说明注意力融合能够有效的融合节点的信息。

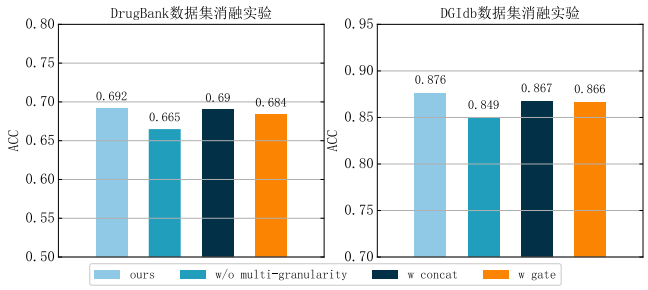

图 2 消融实验
Fig. 2 Ablation Experiment

4.6 参数敏感性分析

为了使模型达到最佳的性能，需要对模型的重要超参数进行调节，本文在 DurgBank、DGIdb 数据集上分别就模型学习率 (learning rate)、RGCN 卷积层数两个参数分别进行实验。

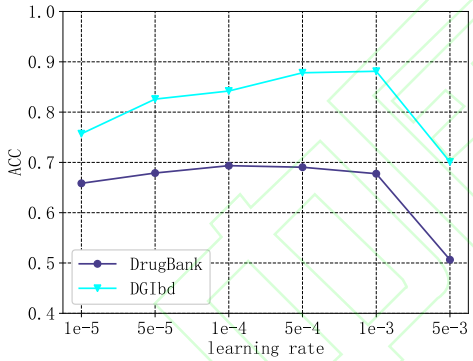

图 3 学习率分析
Fig. 3 Learning rate analysis

学习率敏感性分析：不同的学习率将影响模型收敛的状态，学习率太小，可能导致模型不收敛，而学习率太大可能导致模型损失的震荡。为了挑选出最佳的学习率，实验中我们选取学习率分别为 lr=[1e-5,5e-5,1e-4,5e-4,1e-3,5e-3]，研究学习率对分类准确率的影响。从图 3 中可以看出学习率从右往左逐渐增大，在验证集上准确率总体而言随学习率的增大先增大后减少。造成准确率上升的可能原因是随着学习率的增大模型收敛的速度加快，同时模型不再收敛于局部最优解。造成准确率下降的原因可能是学习率过大造成模型在训练的过程中不能很好的收敛于最优解。因此最终模型选择 1e-3 作为 DGIdb 数据集中学习率参数，选择 1e-4 作为 DrugBank 数据集中学习率参数。

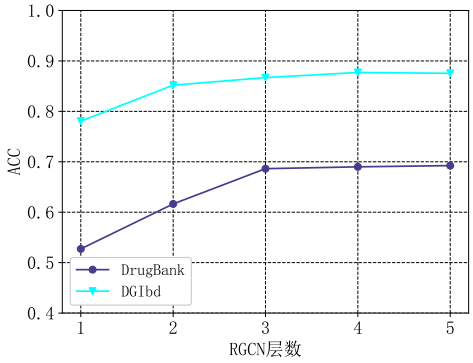

图 4 RGCN 卷积层数分析
Fig.4 RGCN layer analysis

RGCN 卷积层数分析：在 RGCN 中，当图的稀疏度较低时，增加层数可能导致过平滑问题，从而导致模型性能下降。而适当的增加层数有助于模型捕捉更多的结构信息。因此本实验将从层数分别为[1,2,3,4,5]中选择出最佳的 RGCN 层数。最终的试验结果如图 4 所示，可以看到模型层数从 1 到 3 层时，随着模型层数的增加效果逐步提高，这说明在 RGCN 层数增多的过程中，模型可以捕获更多的药物和基因交互的更多信息。3 层以后模型基本不再变化，这说明随着卷积层数的增多，不能一直为模型提供有效的信息，因此我们选择 3 层作为 RGCN 的最终层数。

4.7 每一跳最大节点数量分析

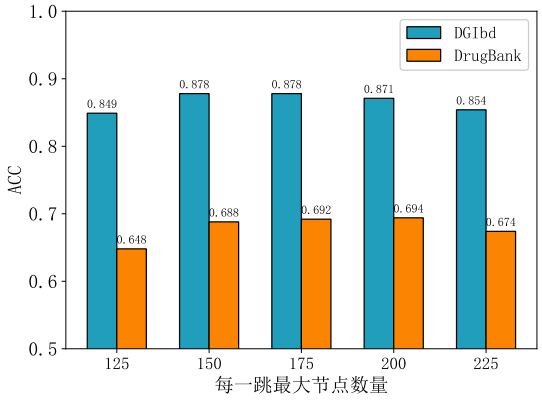

图 5 每一跳最大节点数量分析
Fig. 5 Analysis of the maximum number of nodes per hop

每一跳选取最大的节点数分析：在重启随机游走算法中，每一跳节点需要选取一定数量的节点作为下一跳的出发节点，当每一跳选取节点数量太少，聚合信息将会缺失，而选取节点数量太多，将会对模型的学习带入噪音，产生干扰，因此需要选择合适的节点数量使得模型的效果达到最佳。在每一跳节点最多选取的数量实验中分别选取[125,150,175,200,225]作为节点的数量，实验结果如图 5 所示，在两个数据集中，随着节点数量的增加准确率先上升后

略微下降，模型效果上升的原因是节点数量的增多可以为模型提供更多的有效信息，模型效果降低的原因可能是节点的数量过多导致了获取到了噪音信息。在 DGIbd、DrugBank 两个数据集中每一跳最大节点数量分别达到 175、200 的时候模型达到的效果最佳。

4.8 通用性实验

不同学习率下训练损失随 epoch 变化分析：如图 6 左图所示，随着迭代轮数（epoch）的增加不同的学习率的训练损失（loss）都不断的在下降，在 10 个 epoch 内 loss 都下降较快，当 epoch 在[10-60]之间时 loss 随 epoch 的变化逐渐放缓。在 epoch 为 60 之后，模型已经收敛，不在随 epoch 的变化而变化。

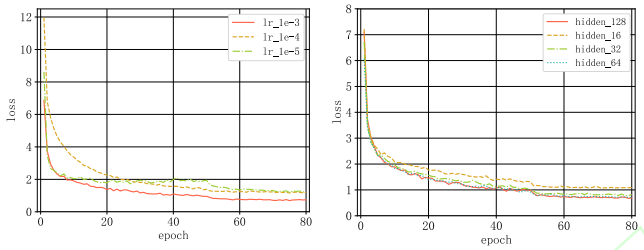

图 6 学习率和隐藏层维度训练损失分析
Fig.6 Analysis of training loss at different learning rates and embedding dimensions

不同隐藏状态维数下训练损失随 epoch 变化分析：如图 6 右图所示，我们可以看到随着 epoch 的增加模型的 loss 逐渐下降，当嵌入的维度为 16 时，模型整体相较于其他的嵌入维度而言效果偏差，导致这一现象的可能原因是嵌入维度太小导致模型学习丢失了较多的信号。虽然不同的嵌入维度，会影响最终的分数，但是总体而言，影响的幅度较小，它说明了本文提出模型的稳定性。

5 结束语

本文提出了一种跨粒度子图对比学习和注意力融合机制的模型。在节点信息聚合方面，相较于原有的基于子图的模型，本文采用了跨粒度子图的方式能够得到比单一子图更多的信息。同时在跨粒度子图的基础上，将基于图对比学习框架进行了适应性改进，应用于本问题中。在节点信息融合方面，相较于已有的拼接融合方式，本文引入了注意力融合的方法，利用子图的远近距离信息，更加充分挖掘节点中隐含的信息。最终的实验结果检验了本文提出的模型在 DGI 问题中关于节点信息聚合和节点信息融合的可行性和有效性。在未来工作中我们将进一步的尝试采用不同的节点融合方式，进一步提高模型的准确率。

参考文献

[1] Jayne-Louise E Pritchard, Tracy A O'Mara, and Dylan M Glubb. Enhancing the promise of drug repositioning through genetics. Frontiers in pharmacology, 8:896, 2017.

[2] Stephen M Strittmatter. Overcoming drug development bottlenecks with repurposing: old drugs learn new tricks. Nature medicine, 20(6):590–591, 2014.

[3] Shawn J Stachel, John M Sanders, Darrell A Henze, Mike T Rudd, Hua-Poo Su, Yiwei Li, Kausik K Nanda, Melissa S Egbertson, Peter J Manley, Kristen LG Jones, et al. Maximizing diversity from a kinase screen: identification of novel and selective pan-trk inhibitors for chronic pain. Journal of Medicinal Chemistry, 57(13):5800–5816, 2014.

[4] Masashi Tsubaki, Kentaro Tomii, and Jun Sese. Compound–protein interaction prediction with end-to-end learning of neural networks for graphs and sequences. Bioinformatics, 35(2):309–318, 2019.

[5] 曹业伟,刘飞.癌症多组学数据深度自编码器整合分型方法[J].计算机工程与应用,2022,58(18):154-161.
Cao Yewei, Liu Fei. A deep autoencoder integration typing method for cancer multi omics data [J]. Computer Engineering and Applications, 2022,58 (18): 154-161.

[6] Maryam Bagherian, Elyas Sabeti, Kai Wang, Maureen A Sartor, Zaneta Nikolovska-Coleska, and Kayvan Najarian. Machine learning approaches and databases for prediction of drug–target interaction: a survey paper. Briefings in bioinformatics, 22(1):247–269, 2021.

[7] Kelsy C Cotto, Alex H Wagner, Yang Yang Feng, Susanna Kiwala, Adam C Coffman, Gregory Spies, Alex Wollam, Nicholas C Spies, Obi L Griffith, and Malachi Griffith. Dgidb 3.0: a redesign and expansion of the drug–gene interaction database. Nucleic acids research, 46(D1):D1068–D1073, 2018.

[8] Rao J, Zheng S, Mai S, et al. Communicative Subgraph Representation Learning for Multi-Relational Inductive Drug-Gene Interaction Prediction[J]. International Joint Conferences on Artificial Intelligence. Vienna, Austria, pp. 3919–25, 2022.

[9] Hakime Oturk, Arzucan Ozgur, and Elif Ozkirimli. Deepdta: deep drug–target binding affinity prediction. Bioinformatics,

34(17):i821–i829, 2018.

[10] Bai P, Miljković F, John B, et al. Interpretable bilinear attention network with domain adaptation improves drug–target prediction[J]. Nature Machine Intelligence, 2023, 5(2): 126-136.

[11] Yunan Luo, Xinbin Zhao, Jingtian Zhou, Jinglin Yang, Yanqing Zhang, Wenhua Kuang, Jian Peng, Ligong Chen, and Jianyang Zeng. A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. Nature communications, 8(1):1–13, 2017.

[12] 张然,王学志,汪嘉葭等.药物-靶点相互作用预测的计算方法综述[J].计算机工程与应用,2023,59(12):1-13.
Zhang Ran, Wang Xuezhi, Wang Jiajia, Meng Zhen. Review of computational methods for predicting drug target interactions [J]. Computer Engineering and Applications, 2023,59 (12): 1-13

[13] Wen Y, Song X, Yan B, et al. Multi-dimensional data integration algorithm based on random walk with restart[J]. BMC bioinformatics, 2021, 22(1): 1-22.

[14] Schlichtkrull M, Kipf T N, Bloem P, et al. Modeling relational data with graph convolutional networks[C]//The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings 15. Springer International Publishing, 2018: 593-607.

[15] Bhatti U A, Tang H, Wu G, et al. Deep learning with graph convolutional networks: An overview and latest applications in computational intelligence[J]. International Journal of Intelligent Systems, 2023, 2023: 1-28.

[16] Cong S, Zhou Y. A review of convolutional neural network architectures and their optimizations[J]. Artificial Intelligence Review, 2023, 56(3): 1905-1969.

[17] Zhang C, Chen H, Zhang S, et al. Geometric inductive matrix completion: A hyperbolic approach with unified message passing[C]//Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining. 2022: 1337-1346.

[18] Li X, Zhang H, Zhang R. Matrix completion via non-convex relaxation and adaptive correlation learning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(2): 1981-1991.

[19] Liu K, Xue F, He X, et al. Joint multi-grained popularity-aware graph convolution collaborative filtering for recommendation[J]. IEEE Transactions on Computational Social Systems, 2022, 10(1): 72-83.

[20] Hartford J, Graham D, Leyton-Brown K, et al. Deep models of interactions across sets[C]//International Conference on Machine Learning. PMLR, 2018: 1909-1918.

[21] Fan Y, Chen M, Pan X. GCRFLDA: scoring lncRNA-disease associations using graph convolution matrix completion with conditional random field[J]. Briefings in Bioinformatics, 2022, 23(1): bbab361.

[22] Defferrard M, Bresson X, Vandergheynst P. Convolutional neural networks on graphs with fast localized spectral filtering[J]. Advances in neural information processing systems, 2016, 29.

[23] Zhu Y, Xu Y, Yu F, et al. Graph contrastive learning with adaptive augmentation[C]//Proceedings of the Web Conference 2021. 2021: 2069-2080.

[24] Zhao T, Liu Y, Neves L, et al. Data augmentation for graph neural networks[C]//Proceedings of the aaai conference on artificial intelligence. 2021, 35(12): 11015-11023.

[25] Liu Y, Zhao Y, Wang X, et al. Multi-scale subgraph contrastive learning[C]//Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence. 2023: 2215-2223.

[26] Chen M, Huang C, Xia L, et al. Heterogeneous graph contrastive learning for recommendation[C]//Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining. 2023: 544-552.

[27] Tao W, Liu Y, Lin X, et al. Prediction of multi-relational drug–gene interaction via Dynamic hyperGraph Contrastive Learning[J]. Briefings in Bioinformatics, 2023, 24(6): bbad371.

[28] Zhang P, Tu S, Zhang W, et al. Predicting cell line-specific synergistic drug combinations through a relational graph convolutional network with attention mechanism[J]. Briefings in Bioinformatics, 2022, 23(6): bbac403.

[29] Ying R, He R, Chen K, et al. Graph convolutional neural networks for web-scale recommender systems[C]//Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining. 2018: 974-983.

[30] Zhang M, Chen Y. Inductive Matrix Completion Based on Graph Neural Networks[C]//International Conference on Learning Representations. 2019.

[31] Fey M ,Lenssen E J .Fast Graph Representation Learning with PyTorch Geometric.[J].CoRR,2019,abs/1903.02428.