文章编号:1671-9352(2023)12-0022-09 DOI:10.6040/j.issn.1671-9352.1.2022.8766

标签指导的多尺度图神经网络蛋白质作用关系预测方法

王新生,朱小飞*,李程鸿
(重庆理工大学计算机科学与工程学院, 重庆 400054)

摘要:提出了一种标签指导的多尺度图神经网络蛋白质作用关系(label guided multi-scale graph neural network protein-protein interactions, LGMG-PPI)预测方法,不仅增强了数据的表征能力,还引入了标签信息指导学习。首先,通过图数据增强得到多尺度图表示,并将多尺度图表示输入图神经网络得到多尺度蛋白质表示,再引入对比学习进一步提高蛋白质表征能力;其次,构造自学习的标签关系图,学习标签之间的关系,得到标签的特征表示;最后,通过标签的特征表示,对蛋白质作用关系的预测进行指导。在 3 个公开的数据集上进行了实验,与最优基准方法相比,LGMG-PPI 方法具有更好的性能,相比最优基准方法,在 SHS27k、SHS148k 和 STRING 这 3 个数据集上的 micro-F_1 分数分别提升了 2.01%、0.94%和 0.93%。

关键词:蛋白质作用关系; 图神经网络; 数据增强;标签关系图
中图分类号:TP391 **文献标志码**:A
引用格式:王新生,朱小飞,李程鸿.标签指导的多尺度图神经网络蛋白质作用关系预测方法[J]. 山东大学学报(理学版),2023,58(12):22-30.

Label guided multi-scale graph neural network for protein-protein interactions prediction

WANG Xinsheng, ZHU Xiaofei*, LI Chenghong
(School of Computer Science and Engineering, Chongqing University of Technology, Chongqing 400054, China)

Abstract: A protein-protein interactions prediction method based on label guided multi-scale graph neural network is proposed, which not only enhances the representation ability of data, but also introduces label information to guide learning. Firstly, the multi-scale graph representation is obtained by graph data augmentation, and the multi-scale graph representation is input into graph neural network to obtain multi-scale protein representation, and comparative learning is introduced to further improve the protein characterization ability. Secondly, the self-learning label relation graph is constructed to learn the relationship between labels and obtain the feature representation of labels. Finally, the prediction of protein-protein interactions is guided by the feature representation of labels. Experiments are carried out on three public datasets. Compared with the optimal benchmark method, the proposed method has better performance. Specifically, compared with the best baseline method, the micro-F_1 scores on the three datasets SHS27k, SHS148k and STRING increase by 2.01%, 0.94% and 0.93% respectively.

Key words: protein-protein interactions; graph neural network; graph data augmentation; graph relation graph

0 引言

蛋白质间的相互作用在生化过程中起着关键的作用[1],如 DNA 复制、转录、翻译和跨膜信号转导等,因此检测蛋白质作用关系(protein-protein interactions, PPIs)及其类型,对了解生物体在正常和疾病状态下的细胞生物学过程至关重要,同时这类研究也有助于治疗靶点的识别[2]和新药物的设计[3]等。在早期的 PPIs 研究工作

收稿日期:2022-09-29; **网络出版时间**:2023-09-12 09:38:30
网络出版地址:https://link.cnki.net/urlid/37.1389.N.20230908.1142.002
基金项目:国家自然科学基金资助项目(62141201); 重庆市技术创新与应用发展专项资助项目(cstc2020jscx-dxwtBX0014); 重庆市教委语言文字科研项目重点项目(yyk20103); 重庆理工大学研究生创新项目(gzlcx20223227)
第一作者简介:王新生(1997—),男,硕士研究生,研究方向为图神经网络和自然语言处理. E-mail:wxscc0610@ 2020.cqut.edu.cn
*通信作者简介**:朱小飞(1979—),男,教授,博士,研究方向为自然语言处理、数据挖掘与信息检索. E-mail:zxf@ cqut.edu.cn

中,使用的是基于实验室的方法,主要包括酵母双杂交筛选[4]、蛋白质芯片[5]和质谱蛋白复合物鉴定[6]等。尽管基于实验的 PPIs 预测方法有广泛的应用,但仍有许多不足之处。首先,实验室实验通常耗时且劳动密集,因此导致 PPIs 的识别效率低下;其次,由于实验室实验条件的限制,因此基于实验室方法生成的 PPIs 数据不完整[7],且已经证实预测结果的错误率较高[8]。为了克服这些缺点,人们提出了各种计算模型,以便更系统地、准确地识别 PPIs,这些计算模型背后的主要思想是利用以往的研究数据,确定先前已知的、具有相互作用的蛋白质对,为设计确认 PPIs 的新实验提供有价值的知识。借助早期实验管理的广泛可用性 PPIs 数据,计算模型的研究得到快速的发展,其中具有代表性的就是以深度学习为基础的计算模型。

早期有关深度学习算法的预测 PPIs 研究主要使用卷积神经网络(convolution neural network, CNN)[9]提取蛋白质的局部特征或使用循环神经网络(recurrent neural network, RNN)[10]保存上下文的长距离依赖信息。这类深度学习算法存在许多问题,如:不能有效地过滤和聚集蛋白质的局部特征,无法保留重要的上下文和序列的氨基酸信息;没有利用蛋白质对之间的相互影响。Chen 等[11]通过构建一个端到端的 PPIs 预测框架,在对 PPIs 建模时考虑了蛋白质序列的上下文和顺序化信息,并且建立的 PIPR 体系结构可以灵活地应用于不同的 PPIs 任务。随着图神经网络(graph neural network, GNN)的发展,Lv 等[12]通过构造蛋白质作用网络图,引入图神经网络来进行预测,这种方法不仅考虑到了蛋白质对之间的影响,还通过蛋白质对之间的关系增强自身的特征表示,进一步提升了 PPIs 预测方法的性能,但该方法有 2 个主要不足之处:(1) 仅基于原始的数据集构造蛋白质作用网络图及蛋白质特征表示,未对原始数据集进行充分的探索,从而导致训练出的模型泛化能力不足;(2) 蛋白质之间往往存在多种作用关系,这些作用关系可能存在相互关联的信息,但该方法未曾考虑这方面的信息。

针对上述的 2 个问题,本文提出一个自学习标签指导的多尺度图神经网络 PPIs(LGMG-PPI)预测方法,不仅提升了模型的泛化能力,还引入标签信息作为指导,进一步提升模型的分类效果。首先,构造蛋白质相互作用网络,通过图数据增强得到多尺度的蛋白质作用网络;其次,引入图神经网络,通过邻居节点加强自身的特征表示,学习到多尺度的蛋白质特征表示;然后,为了消除不同尺度蛋白质特征表示的差异,引入对比学习来进一步提升蛋白质特征表示的泛化能力;进一步,通过构造自学习的标签关系图,学习标签之间的联系,得到标签的特征表示;最后,用学习到的标签特征去指导蛋白质间作用关系表示的学习。

本文的贡献主要包括 3 个方面:(1) 提出了一种 LGMG-PPI 预测方法,通过图数据增强和图神经网络学习到多尺度的蛋白质特征表示,并且引入对比学习,提升了蛋白质特征表示的泛化能力;(2) 引入标签信息,通过构造自学习的标签关系图,学习到标签之间的关系,进而指导蛋白质相互作用关系的学习;(3) 在 3 个公开的数据集上进行实验,以验证本文所提出方法的有效性,并与最优的基准方法进行分类效果对比。

1　相关工作

近几年,GNN 在各种不规则数据任务中的应用取得了巨大成功,如节点分类、蛋白质属性预测等,目前大多数 GNN 一般可分为谱域 GNN 和空域 GNN。谱域 GNN 是基于图谱理论学习节点的表示,根据卷积定理将空域上图的卷积转化为频率上的卷积。例如,Bruna 等[13]使用图拉普拉斯算子设计傅里叶域的图卷积算法,Defferrard 等[14]采用切比雪夫多项式作为卷积滤波器,提高了滤波效率。空域 GNN 是利用空间近邻直接在图上定义卷积操作。Velickovic 等[15]利用注意力机制聚集邻域表征,得到图注意力网络(graph attention networks, GAT);更进一步,Xu 等[16]提出不仅要聚集邻域的表征,还需要保留节点上的隐藏状态信息,得到图同构网络(graph isomorphism network, GIN)。GNN 作为图表示学习的重要工具,尽管经过不断改进,效果也越来越好,但是仍然面临着 GNN 过度平滑的问题。过度平滑是指当 GNN 的层数增加时,GNN 中的节点表示会收敛到一个固定的点,彼此之间变得不可区分,这种现象限制了 GNN 的深度,从而阻碍了 GNN 的表征能力。目前,解决该问题的一个常用方法是将 GNN 的层数作为一个超参数,测试不同层数 GNN 的效果,最终确定 GNN 的层数。

2　方法

首先对 PPIs 进行形式化的描述,然后对本文所提出的 LGMG-PPI 预测方法进行详细描述。图 1 展示了

LGMG-PPI 的整体架构,主要由多尺度图神经网络模块和自学习的标签关系图模块组成。

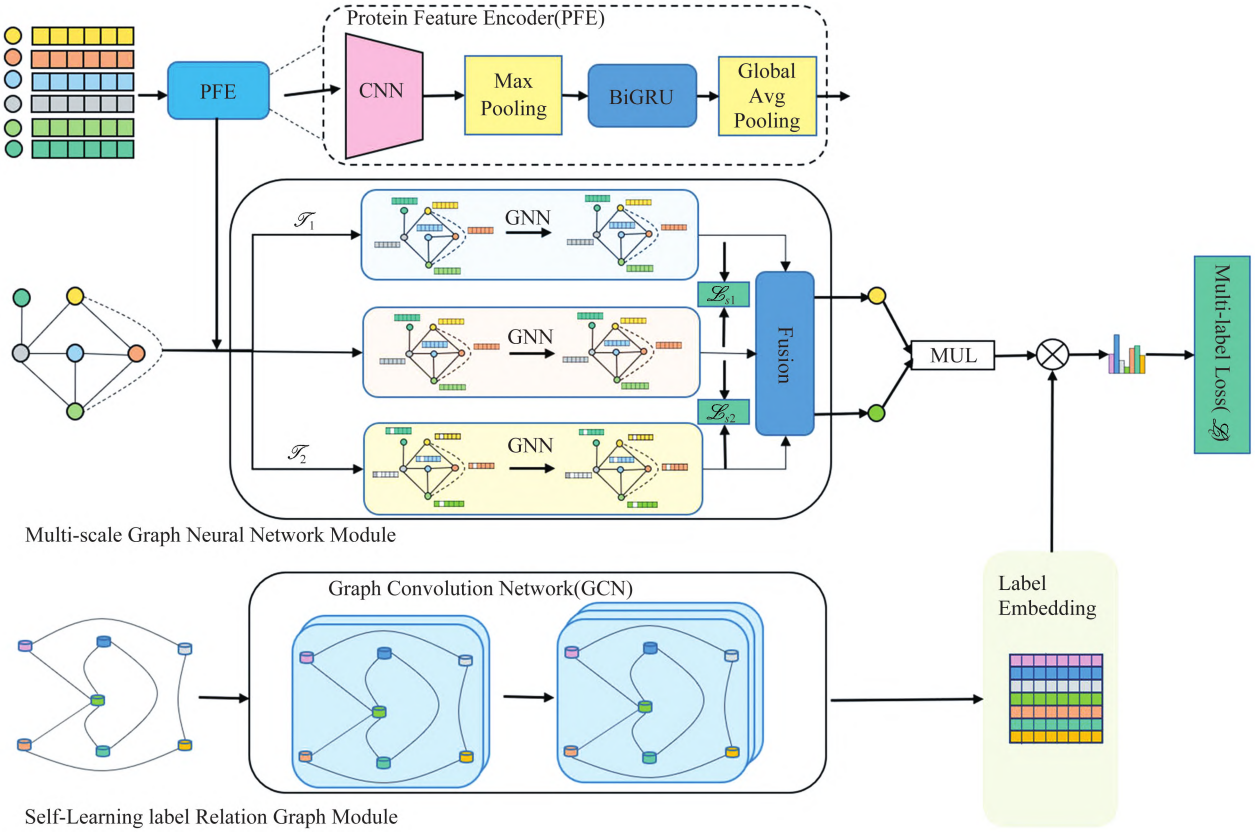

图 1 LGMG-PPI 模型框架图

Fig.1 Framework of model LGMG-PPI

2.1 蛋白质相互作用关系

蛋白质是由氨基酸构成序列,常见氨基酸有 20 种。定义蛋白质集合 $\mathscr{P}=\{p_1,p_2,\cdots,p_n\}$,其中 $p_i=\{a_1,a_2,\cdots,a_m \mid a_j\in\mathscr{A}\}$,$\mathscr{A}$ 是氨基酸集合。定义 $\mathscr{X}=\{x_{ij}=\{p_i,p_j\} \mid i\neq j,\ p_i,p_j\in\mathscr{P},\ I(x_{ij})\in\{0,1\}\}$ 为 PPIs 集合,其中 I 表示 2 个蛋白质之间是否存在关系,若 $I(x_{ij})=1$,表示蛋白质 p_i 和蛋白质 p_j 之间存在作用关系;若 $I(x_{ij})=0$,表示蛋白质 p_i 和蛋白质 p_j 之间不存在作用关系,或者表示在目前的研究工作中未发现二者之间存在作用关系。通过上述的定义,本文将蛋白质作为节点,PPIs 作为连边,构造 PPIs 图 $\mathscr{G}=(\mathscr{P},\mathscr{X})$。

PPIs 仅仅表示蛋白质之间是否存在相互作用关系,但蛋白质之间可能存在多种作用关系,本文的任务就是预测蛋白质之间存在的这些多种作用关系,是一个多标签分类任务。本文定义 PPIs 的标签集合为 $\mathscr{L}=\{l_1,l_2,\cdots,l_t\}$,其中 t 表示有 t 种作用关系。

2.2 蛋白质特征编码器

为了充分提取蛋白质的局部特征和全局特征,本文设计一种蛋白质特征编码器(protein feature encoder, PFE),它主要包含局部特征编码器和全局特征编码器这 2 个模块。

2.2.1 局部特征编码器

局部特征编码器包括 CNN[9]、最大池化层(global max pooling, GMP)。输入蛋白质序列 $p_i\in\mathscr{P}$,通过局部特征编码器,得到蛋白质的局部特征表示 h_i,

$$h_i=f_{\mathrm{GMP}}(f_{\mathrm{CNN}}(p_i;\theta_{\mathrm{CNN}}))。\tag{1}$$

2.2.2 全局特征编码器

全局特征编码器包括双向门控循环单元(bidirectional gate recurrent unit, BiGRU)和全局平均池化层(global avg pooling, GAP)。

将基于局部特征编码器的特征表示 h_i 输入全局编码器,进一步提取蛋白质特征的全局特征表示,最终得到

具有局部信息和全局信息的蛋白质特征表示 $x_i \in X$,

$$x_i = f_{\mathrm{GAP}}(f_{\mathrm{BiGRU}}(h_i;\theta_{\mathrm{BiGRU}}))。\tag{2}$$

2.3　多尺度图数据增强模块

之前的研究工作[17]表明,适当地扰动图数据能够有效地增强图数据的泛化能力。多尺度图数据增强(multi-scale graph data augmentation, MS-GDA)模块主要包含 2 种图数据增强函数。定义原始图 $G=(\boldsymbol{X},\boldsymbol{A})$,其中节点特征 $\boldsymbol{X}\in\mathbf{R}^{N\times F}$ 和邻接矩阵 $\boldsymbol{A}\in\mathbf{R}^{N\times N}$。

模块基于原始图 $G=(\boldsymbol{X},\boldsymbol{A})$ 从 2 个不同的视角应用随机图数据增强函数 $\mathcal{T}_1$ 和 $\mathcal{T}_2$,分别得到 $G_1=(\boldsymbol{X},\boldsymbol{A}_1)$ 和 $G_2=(\boldsymbol{X}_2,\boldsymbol{A})$。

$$G_1=(\boldsymbol{X},\boldsymbol{A}_1)=\mathcal{T}_1(G=(\boldsymbol{X},\boldsymbol{A})),\tag{3}$$

$$G_2=(\boldsymbol{X}_2,\boldsymbol{A})=\mathcal{T}_2(G=(\boldsymbol{X},\boldsymbol{A}))。\tag{4}$$

这里的图数据增强函数 $\mathcal{T}_1$ 和 $\mathcal{T}_2$ 是标准的图扰动策略[17],下面介绍 $\mathcal{T}_1$ 和 $\mathcal{T}_2$ 如何进行图扰动。

针对原始图 $G=(\boldsymbol{X},\boldsymbol{A})$,对其连边进行扰动,随机地删除原始图拓扑结构的连边,最终得到 $G_1=(\boldsymbol{X},\boldsymbol{A}_1)$。具体表示为:

$$\mathcal{T}_1:\mathscr{X}_1=\varepsilon_1\cdot\mathscr{X},\tag{5}$$

$$\varepsilon_1\sim\mathrm{Bernoulli}(N,1-\delta_1),\tag{6}$$

其中:$\mathscr{X}$ 表示原始图的连边集合;Bernoulli 表示伯努利分布;$\delta_1\in(0,1)$ 是一个超参数,表示删除连边的比率。

对原始图 $G=(\boldsymbol{X},\boldsymbol{A})$ 的节点特征进行扰动,随机地将原始图节点特征的某些列置为 0,最终得到 $G_2=(\boldsymbol{X}_2,\boldsymbol{A})$。具体表示为:

$$\mathcal{T}_2:\boldsymbol{X}_2=\varepsilon_2\cdot\boldsymbol{X},\tag{7}$$

$$\varepsilon_2=\begin{cases}0, & \mathrm{Unifrom}(0,1)<\delta_2\\ 1, & \mathrm{Unifrom}(0,1)>=\delta_2,\end{cases}\tag{8}$$

其中:Unifrom 表示均匀分布;$\varepsilon_2\in\mathbf{R}^F$, $\delta_2\in(0,1)$ 是一个超参数,表示节点特征置为 0 的比率。

2.4　多尺度图神经网络表示学习

2.4.1　图编码器

GNN 是当前解决图结构数据的主流深度学习模型,主要通过中心节点和邻居节点的关系聚合邻居节点的特征,进而增强自身节点的特征表示。具体来讲,通过 k 次聚合、更新的迭代,节点表示聚合其 k 跳邻居节点的表示。具有 k 次迭代的 GNN 表示如下:

$$a_v^k=\mathrm{AGG}^k(\{z_u^{k-1}:u\in\mathcal{N}(v)\}),\quad z_v^k=\mathrm{UPDATE}^k(z_v^{k-1},a_v^k),\tag{9}$$

其中:$\mathcal{N}(v)$ 表示节点 v 的邻居集合;$z_v^{(k)}$ 表示节点 v 第 k 次迭代的特征表示。

本文采用 GIN 进行图表示的学习,输入数据是 $G=(\boldsymbol{X},\boldsymbol{A})$。具体表示为

$$z_v^{(k)}=\mathrm{MLP}^{(k)}\left((1+\omega^{(k)})\cdot z_v^{(k-1)}+\sum_{u\in\mathcal{N}(v)}z_u^{(k-1)}\right),\tag{10}$$

其中:ω 是一个可学习参数或者常数;初始化 $z_v^0=x_v$, $x_v\in\boldsymbol{X}$,最终学习到节点 v 的特征表示 $z_v\in\boldsymbol{Z}_0$。

同理,分别将经过数据增强后的图 $G_1=(\boldsymbol{X},\boldsymbol{A}_1)$ 和 $G_2=(\boldsymbol{X}_2,\boldsymbol{A})$ 输入到 GIN 中,学习到 $\boldsymbol{Z}_1\in\mathbf{R}^{N\times F}$ 和 $\boldsymbol{Z}_2\in\mathbf{R}^{N\times F}$ 这 2 种不同的节点特征表示。

2.4.2　多尺度图融合

基于上述的学习,得到了不同尺度的节点特征表示。为了得到表征能力更好的节点表示,下面将融合上述的 3 种节点特征,具体表示如下:

$$\boldsymbol{Z}'=f_{\mathrm{Fusion}}([\boldsymbol{Z}_0,\boldsymbol{Z}_1,\boldsymbol{Z}_2]),\tag{11}$$

其中 f_{Fusion} 表示融合函数,可以为求和、取平均等,$\boldsymbol{Z}'\in\mathbf{R}^{N\times F}$。

通过最终得到蛋白质关系图的节点特征表示 $\boldsymbol{Z}'$,得到边的特征表示 $\boldsymbol{E}\in\mathbf{R}^{R\times F}$,具体表示如下:

$$e_{ij}=z_i'\odot z_j',\tag{12}$$

其中:$e_{ij}\in\boldsymbol{E}$;$\odot$ 表示哈达玛积;$z_i'\in\mathbf{R}^F$ 和 $z_j'\in\mathbf{R}^F$ 分别表示节点 i 和节点 j 的特征表示。

2.5 自学习的标签关系图

2.5.1 标签关系图

以往的工作往往通过挖掘数据中的各类标签的信息，人为地构造标签关系。Chen 等[18]通过分析不同类别标签之间出现的频率构造标签共现图，但这种方式需要大量的数据才能保证构造的标签关系正确性，并且当新的实验数据与原本的数据的标签分布不一致时，会导致分类效果的骤降。

本文采用一种自学习的方式得到标签之间的关系表示，构造自学习的标签关系图（self-learning label relation graph，SL-LRG）。首先，设置一个可学习参数 $\boldsymbol{A}_L \in \mathbf{R}^{T\times T}$，$T$ 表示标签的类别个数，初始化 $\boldsymbol{A}_L$ 为单位矩阵，将 $\boldsymbol{A}_L$ 作为标签关系图的初始拓扑结构；其次，通过预训练模型 BERT[19]获取标签名称的嵌入表示，作为节点的特征表示，具体表示如下：

$$\boldsymbol{X}_L = \mathrm{BERT}(L_{\mathrm{NAME}}), \tag{13}$$

其中：$L_{\mathrm{NAME}} \in \mathbf{R}^{T}$表示标签名称，$\boldsymbol{X}_L \in \mathbf{R}^{T\times D}$表示标签名称的词向量。最终得到了标签关系图 $G_L=(\boldsymbol{A}_L,\boldsymbol{X}_L)$。

为了进一步利用好标签之间的关系，加强标签节点的表示，引入 GNN 进行学习。采用图卷积神经网络（graph convolution network，GCN），输入为 $G_L=(\boldsymbol{A}_L,\boldsymbol{X}_L)$，输出为标签节点的特征表示 $\boldsymbol{Z}_L \in \mathbf{R}^{T\times F}$，具体表示如下：

$$\boldsymbol{Z}_L^{(l)} = \sigma(\widetilde{\boldsymbol{D}}^{-\frac{1}{2}}\boldsymbol{A}_L\widetilde{\boldsymbol{D}}^{-\frac{1}{2}}\boldsymbol{Z}_L^{(l-1)}\boldsymbol{W}^{(l-1)}), \tag{14}$$

其中：初始化 $\boldsymbol{Z}_L^{(0)}=\boldsymbol{X}_L$；$\widetilde{\boldsymbol{D}}$ 为度矩阵；$\boldsymbol{W}^{(l-1)}$ 是一个可学习的参数矩阵；σ 表示 sigmoid 激活函数。值得注意的是，$\boldsymbol{A}_L$ 是一个初始化为单位矩阵的可学习参数矩阵，在模型训练的过程中通过梯度回传更新参数 $\boldsymbol{A}_L$，进而学习到数据中隐含的标签关系，达到自学习标签关系图的目的。

2.5.2 标签指导学习

基于标签关系图的节点特征表示 $\boldsymbol{Z}_L \in \mathbf{R}^{T\times F}$，进一步修正蛋白质关系图的连边特征表示 $\boldsymbol{E} \in \mathbf{R}^{R\times F}$，具体表示如下：

$$\boldsymbol{Z}_E = \boldsymbol{E}\times\boldsymbol{Z}_L^{\mathrm{T}}, \tag{15}$$

其中 $\boldsymbol{Z}_E \in \mathbf{R}^{R\times T}$为蛋白质关系图连边特征的最终表示。

2.6 损失函数

2.6.1 自监督学习任务

在现实生活中，数据中往往包含噪音，而这些噪音会使模型无法准确表示原始数据分布，严重影响模型的学习效果。为了解决这个问题，本文在模型中引入了一个自监督学习任务，目的是增加辅助任务来提高主要学习任务的准确性，提高模型的性能。

在模型中，通过多尺度图表示学习模块得到了 $\boldsymbol{Z}_0 \in \mathbf{R}^{N\times F}$、$\boldsymbol{Z}_1 \in \mathbf{R}^{N\times F}$、$\boldsymbol{Z}_2 \in \mathbf{R}^{N\times F}$这 3 种视图的节点特征表示。定义积极样本对$(z_{0,i},z_{1,i})$，建立如下损失函数：

$$\ell(z_{0,i},z_{1,i}) = \log\frac{e^{\theta(z_{0,i},z_{1,i})/\tau}}{e^{\theta(z_{0,i},z_{1,i})/\tau}+\sum\limits_{k=1,k\neq i}^{N} e^{\theta(z_{0,i},z_{1,k})/\tau}}, \tag{16}$$

其中：$z_{0,i}\in\boldsymbol{Z}_0$，$z_{1,i}\in\boldsymbol{Z}_1$，$\theta(z_{0,i},z_{1,i})$，分别表示计算 $z_{0,i}$和 $z_{1,i}$的余弦相似度；τ 表示温度参数。

同理，积极样本对$(z_{1,i},z_{0,i})$的损失函数如下：

$$\ell(z_{1,i},z_{0,i}) = \log\frac{e^{\theta(z_{1,i},z_{0,i})/\tau}}{e^{\theta(z_{1,i},z_{0,i})/\tau}+\sum\limits_{k=1,k\neq i}^{N} e^{\theta(z_{1,i},z_{0,k})/\tau}}。 \tag{17}$$

$\boldsymbol{Z}_0 \in \mathbf{R}^{N\times F}$和$\boldsymbol{Z}_1 \in \mathbf{R}^{N\times F}$这 2 种视图的所有积极样本对损失函数如下：

$$\mathscr{L}_{s1} = \frac{1}{2N}\sum_{i=1}^{N}[\ell(z_{0,i},z_{1,i})+\ell(z_{1,i},z_{0,i})]。 \tag{18}$$

同理，$\boldsymbol{Z}_0 \in \mathbf{R}^{N\times F}$和$\boldsymbol{Z}_2 \in \mathbf{R}^{N\times F}$这 2 种视图的所有积极样本对的损失函数如下：

$$\mathscr{L}_{s2} = \frac{1}{2N}\sum_{i=1}^{N}[\ell(z_{0,i},z_{2,i})+\ell(z_{2,i},z_{0,i})]。 \tag{19}$$

2.6.2　监督学习任务

基于学习到的蛋白质关系图的连边特征表示 $\boldsymbol{E}\in\mathbf{R}^{R\times F}$,进行监督学习任务的模型训练,具体表示如下:

$$p_{ij}=\mathrm{Softmax}(e_{ij}),\tag{20}$$

$$\hat{y}_{ij}=\mathrm{argmax}(p_{ij}),\tag{21}$$

$$\mathscr{L}_c=\sum_{c=1}^{t}\left(\sum_{(i,j)\in\mathscr{X}_{\text{train}}}-y_{ij}^{c}\log\hat{y}_{ij}^{c}-(1-y_{ij}^{c})\log(1-\hat{y}_{ij}^{c})\right),\tag{22}$$

其中:$e_{ij}\in\boldsymbol{E}$,t 表示标签类别个数,$\mathscr{X}_{\text{train}}$表示连边集合的训练集。

结合自监督学习任务和监督学习任务,最终得到的损失函数表示为:

$$\mathscr{L}_{\text{total}}=\mathscr{L}_c+\lambda_1\mathscr{L}_{s1}+\lambda_2\mathscr{L}_{s2},\tag{23}$$

其中 λ_1 和 λ_2 是超参数。

3　实验

3.1　数据集

本文沿用以往工作[12]的数据集设置,使用 STRING[20]数据库中的 PPIs 数据来评估模型。此外,Chen 等[11]从 STRING 中抽取了 2 个子数据集,分别为 SHS27k 和 SHS148k。本文将这 2 个数据也加入到实验中,用来评估模型效果。3 种数据集的具体信息如表 1 所示,其中原始数据集是蛋白质网络关系图,节点代表蛋白质,连边代表蛋白质之间存在作用关系;其次,由于蛋白质由氨基酸序列组成,因此本文统计了每个数据集中组成蛋白质的氨基酸序列平均长度。

表 1　数据集统计

Table 1　The statistics of datasets

数据集	节点数	连边数	氨基酸数(Avg)	标签数
SHS27k	1 690	7 624	571	7
SHS148k	5 189	44 488	597	7
STRING	15 335	593 397	604	7

3.2　实验设置和评价指标

从数据集中随机挑选 20%的数据作为测试集,为了消除数据划分的随机性对 PPI 方法性能的影响,在 3 种不同的随机种子下重复实验结果。本文使用基于氨基酸序列的蛋白质特征,参考 Chen 等[11]使用的氨基酸嵌入方法来表示每个氨基酸。模型采用 Adam 算法更新所有的可训练参数的评价标准沿用文献[12]。

3.3　基准方法

3.3.1　机器学习基准方法

本文选择 3 种具有代表性的机器学习(machine learning, ML)算法作为基准方法,分别是支持向量机(support vector machine, SVM)[21]、逻辑回归(logistic regression, LR)[22]和随机森林(random forest, RF)[23]。

3.3.2　深度学习基准方法

本文选择 4 种 PPIs 预测任务的深度学习(deep learning, DL)算法,分别是 DPPI[24]、DNN-PPI[25]、PIPR[11]和 GNN-PPI[12]。DPPI 是一种新的学习框架,仅利用蛋白质的序列信息对 PPIs 进行建模和预测;DNN-PPI 采用 CNN 和 LSTM 进行编码,并且分别对蛋白质特征进行建模,最终得到蛋白质对之间的关系表示;PIPR 是一个端到端框架,其在 Siamese 结构中加入了深度残差卷积神经网络,并利用局部特征和上下文信息对 PPIs 进行预测;GNN-PPI 首次提出利用 GNN 学习蛋白质特征,其通过构建蛋白质作用关系网络,利用相邻的蛋白质节点的表示增强自身的特征表示,进而预测 PPIs。

3.4　对比实验

表 2 展示了不同计算方法在不同数据集上的性能,为 3 次不同随机种子下的 micro-F_1 均值±标准差。通过观察分析可得出以下结果:

(1) 深度学习算法的性能总体上优于机器学习算法的,表明基于深度学习的技术在封装蛋白质对各种类型的信息(如氨基酸组成及其共现情况)并自动提取适合学习目标的鲁棒信息方面具有优越性;其次,随

着数据集的增大,各类方法的性能也随之增加,原因是数据量的增加使得模型学习更充分,模型的泛化能力更强。

(2) 与最优的基准方法 GNN-PPI 相比,LGMG-PPI 在所有类型的数据上具有更好的预测效果,且效果更加稳定。其中 micro-F_1 分数在 SHS27k 数据集上提升了 2.01%,在 SHS148k 数据集上提升了 0.94%,在 STRING 数据集上提升了 0.93%。

表 2 不同模型在不同数据集上的 micro-F_1

Table 2 The micro-F_1 of different models on different datasets 单位:%

方法	SHS27k	SHS148k	STRING
SVM	75.35±1.05	80.55±0.23	—
RF	78.45±0.88	82.10±0.20	88.91±0.08
LR	71.55±0.93	67.00±0.07	67.74±0.16
DPPI	73.99±5.04	77.48±1.39	94.85±0.13
DNN-PPI	77.89±4.97	88.49±0.48	83.08±0.11
PIPR	83.31±0.75	90.05±2.59	94.43±0.10
GNN-PPI	87.91±0.39	92.26±0.10	95.43±0.10
LGMG-PPI	**89.68±0.10**	**93.13±0.03**	**96.32±0.04**

3.5 消融实验

为了进一步分析模型中各个模块的作用,通过删减不同的模块进行实验,进而验证各个模块的有效性,本文设置了以下消融实验。

(1) w/o MS-GDA($\mathscr{T}_1$):表示去除多尺度图数据增强模块中 $\mathscr{T}_1$ 类型的数据增强,即不使用扰动图连边的数据增强方法。

(2) w/o MS-GDA($\mathscr{T}_2$):表示去除多尺度图数据增强模块中 $\mathscr{T}_2$ 类型的数据增强,即不使用扰动图节点特征的数据增强方法。

(3) w/o MS-GDA:表示完全去除多尺度图数据增强模块,即不使用图数据增强策略。

(4) w/o SL-LRG:表示去除标签关系图模块,即不使用标签信息来进行指导学习。

实验结果如表 3 所示。从实验结果来看,扰动图节点特征的数据增强方法略优于扰动图连边的数据增强方法,且 2 种图数据增强方法都是有益于模型的,说明图数据增强方法通过扰动原始图数据可增强模型的泛化能力。此外,当去除标签关系图模块后,模型在所有数据集上的效果均有降低,说明引入标签关系图模块能够学习到标签之间的隐含关系,进而得到标签的隐藏状态,对最终的预测结果进行指导。总体来讲,LGMG-PPI 各个子模块都是有益于整个模型的。

表 3 消融实验

Table 3 The ablation study 单位:%

方法	SHS27k	SHS148k	STRING
LGMG-PPI	**89.68±0.10**	**93.13±0.03**	**96.32±0.04**
w/o MS-GDA($\mathscr{T}_1$)	89.35±0.05	93.05±0.11	96.14±0.17
w/o MS-GDA($\mathscr{T}_2$)	89.23±0.10	92.95±0.04	96.28±0.17
w/o MS-GDA	88.97±0.09	92.80±0.14	95.92±0.03
w/o SL-LRG	89.39±0.12	92.87±0.04	96.04±0.02

3.6 自学习标签关系图有效性实验

3.6.1 拓扑结构有效性实验

自适应标签图通过引入自学习的拓扑结构,进而学习标签特征。为了验证拓扑结构的有效性,不引入自学习的拓扑结构,用多层感知机(multi-layer perceptron, MLP)替代 GCN,即公式(14)替换为 $\boldsymbol{Z}_L=f_{\mathrm{MLP}}(\boldsymbol{X}_L)$。

实验结果如图 2 所示。从实验结果来看,引入标签的拓扑结构的效果明显更好,说明 PPIs 预测任务的标签间存在某些联系,而通过自学习标签关系图能够很好地学习到标签间的隐含关系,进一步证明了 LGMG-PPI的有效性。

3.6.2　节点特征有效性实验

自学习标签关系图节点特征的初始表示是词的嵌入表示,本文使用预训练模型 BERT[19] 得到词的嵌入表示。通过实验比较 BERT 和 One-Hot 嵌入表示下的模型效果,评估模型在不同词嵌入表示下的性能。

实验结果如图 3 所示。从图中可以看出,当使用不同的词嵌入作为 GCN 的输入时,多标签识别精度不会受到显著影响,说明模型所实现的效果提高并不完全来自于词嵌入所衍生的语意信息。此外,使用强大的词嵌入表示可以带来轻微的性能提升,可能的原因是从大型文本语料库中学习的词嵌入保留了一定的语意信息,而这些词嵌入在嵌入空间中存在一定的联系,模型可以利用这些隐式联系进一步提升模型的预测能力。

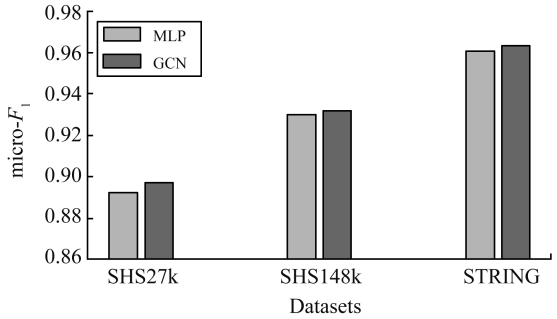

图 2　SL-LRG 拓扑结构有效性验证

Fig.2　Verify the validity of topology structure of SL-LRG

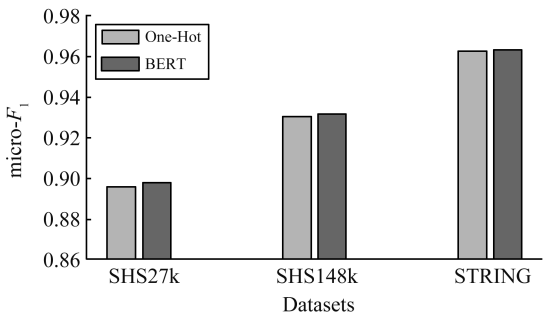

图 3　SL-LRG 节点特征有效性验证

Fig.3　Verify the validity of node feature of SL-LRG

4　总结与展望

本文提出一种 LGMG-PPI 预测方法,通过图数据增强得到多种尺度的图表示,并将这些多种尺度的图输入图神经网络,得到多种尺度的蛋白质特征表示,并且引入对比学习,进一步提高蛋白质的表征能力。此外,构造自学习的标签关系图,学习标签之间的关系,进而得到标签的信息表示,对最终的蛋白质关系的预测进行指导学习。在 3 个公开数据集上的实验结果表明 LGMG-PPI 方法在预测蛋白质作用关系任务上的有效性,且预测效果优于最优的基准方法。

参考文献:

[1] STELZL U, WORM U, LALOWSKI M, et al. A human protein-protein interaction network: a resource for annotating the proteome[J]. Cell, 2005, 122(6):957-968.

[2] PETTA I, LIEVENS S, LIBERT C, et al. Modulation of protein-protein interactions for the development of novel therapeutics [J]. Molecular Therapy, 2016, 24(4):707-718.

[3] SKRABANEK L, SAINI H K, BADER G D, et al. Computational prediction of protein-protein interactions[J]. Molecular Biotechnology, 2008, 38(1):1-17.

[4] FIELDS S, STERNGLANZ R. The two-hybrid system: an assay for protein-protein interactions[J]. Trends in Genetics, 1994, 10(8):286-292.

[5] TONG A H Y, DREES B, NARDELLI G, et al. A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules[J]. Science, 2002, 295(5553):321-324.

[6] HO Y, GRUHLER A, HEILBUT A, et al. Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry[J]. Nature, 2002, 415(6868):180-183.

[7] RAO V S, SRINIVAS K, SUJINI G N, et al. Protein-protein interaction detection: methods and analysis[J]. International Journal of Proteomics, 2014, 2014(1):147648-147659.

[8] HUANG H, BADER J S. Precision and recall estimates for two-hybrid screens[J]. Bioinformatics, 2009, 25(3):372-378.

[9] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6):84-90.

[10] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8):1735-1780.

[11] CHEN M, JU C J T, ZHOU G, et al. Multifaceted protein-protein interaction prediction based on *Siamese* residual RCNN [J]. Bioinformatics, 2019, 35(14):i305-i314.

[12] LV G F, HU Z Q, BI Y G et al. Learning unknown from correlations: graph neural network for inter-novel-protein interaction prediction[C]//Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence. San Francisco: Margan Kaufmann, 2021: 3677-3683.

[13] BRUNA J, ZAREMBA W, SZLAM A, et al. Spectral networks and locally connected networks on graphs[C]//International Conference on Learning Representations. New Orleans: OpenReview.net, 2014.

[14] 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(12):3844-3852.

[15] VELICKOVIC P, CUCURULL G, CASANOVA A, et al. Graph attention networks[J]. Stat, 2017, 1050(20):10.48550.

[16] XU K, HU W H, LESKOVEC J, et al. How powerful are graph neural networks? [C]//International Conference on Learning Representations. New Orleans: OpenReview.net, 2019.

[17] YOU Y, CHEN T, SUI Y, et al. Graph contrastive learning with augmentations[J]. Advances in Neural Information Processing Systems, 2020, 33:5812-5823.

[18] CHEN Z M, WEI X S, WANG P, et al. Multi-label image recognition with graph convolutional networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 5177-5186.

[19] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding [C]//NAACL.Minneapolis:Association for Computational Linguistics, 2019: 4171-4186.

[20] SZKLARCZYK D, GABLE A L, LYON D, et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets[J]. Nucleic Acids Research, 2019, 47(D1): D607-D613.

[21] GUO Y, YU L, WEN Z, et al. Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences[J]. Nucleic Acids Research, 2008, 36(9):3025-3030.

[22] SILBERBERG Y, KUPIEC M, SHARAN R. A method for predicting protein-protein interaction types[J]. PLoS One, 2014, 9(3):e90904.

[23] WONG L, YOU Z H, LI S, et al. Detection of protein-protein interactions from amino acid sequences using a rotation forest model with a novel PR-LPQ descriptor[C]//International Conference on Intelligent Computing. Fuzhou: Springer, 2015: 713-720.

[24] HASHEMIFAR S, NEYSHABUR B, KHAN A A, et al. Predicting protein-protein interactions through sequence-based deep learning[J]. Bioinformatics, 2018, 34(17):i802-i810.

[25] LI H, GONG X J, YU H, et al. Deep neural network based predictions of protein interactions using primary sequences[J]. Molecules, 2018, 23(8):1923.

(编辑:于善清)