中国科学: 信息科学 2024 年 第 54 卷 第 10 期: 2409–2427
SCIENTIA SINICA Informationis
论文

非线性信息瓶颈指导的层次图结构学习方法

孙庆赟[1,2], 罗家逸[1,2], 杨贝宁[1,2], 李建欣[1,2*]

1. 北京航空航天大学计算机学院, 北京 100191
2. 北京市大数据与脑机智能高精尖创新中心, 北京 100191
* 通信作者. E-mail: lijx@buaa.edu.cn

收稿日期: 2024–02–06; 修回日期: 2024–04–30; 接受日期: 2024–06–18; 网络出版日期: 2024–10–09

国家自然科学基金 (批准号: 62225202, 62302023) 资助项目

摘要 近年来, 图神经网络在各个领域的图数据挖掘任务上取得了显著的成功, 已成为领域的研究热点. 图神经网络通过结构传播节点信息, 并以此计算节点的表征, 在大量应用场景上取得了显著的效果提升. 大多数图神经网络模型遵循消息传递机制, 直接将原始图数据作为输入, 假设观测到的图结构准确地描述了节点之间完整的关系. 然而, 真实场景中图数据的产生往往受多种因素影响, 包含大量随机噪声和人为扰动. 这些噪声信息和干扰信息在图神经网络信息聚合的过程中随着图结构传播, 对图表征质量产生严重的影响. 如何度量、识别图数据中的噪声信息, 是领域关注的热点问题之一. 本文从信息论的角度出发, 提出了一种非线性信息瓶颈指导的层次图结构学习方法 NIB-HGSL, 针对图层级分类任务, 为去除结构噪声、学习鲁棒的图表征提供了一个统一通用的框架. NIB-HGSL通过有效信息保留与噪声信息压缩的均衡优化, 可以获得对下游任务来说最关键的层次化最小充分图. 实验结果表明, 本文所提出的 NIB-HGSL方法与其他基线方法相比, 可提高图分类和图回归任务的准确性和鲁棒性.

关键词 图表示学习, 信息瓶颈, 图结构学习, 图神经网络, 图分类

1 概述

近年来, 随着互联网、物联网等信息技术的飞速发展, 图数据规模呈现爆炸式增长, 如微信微博等社交网络图数据, 蛋白质分子等生物图数据, 以及交通通信等基础设施图数据. 由于图数据的强大表达能力和普遍存在性, 对图数据的处理分析已成为数据挖掘领域的研究热点. 图表示学习 (graph representation learning) 通过将图数据映射到低维、稠密的向量空间, 有效地提升了图挖掘的性能与效率[1]. 其中, 图神经网络 (graph neural network, GNN) 是图表示学习的主流方法之一, 在社会计算、商品推荐、交通流量预测等大量应用场景上取得了显著的效果提升[2~4]. 尽管图神经网络展现出了强大

引用格式: 孙庆赟, 罗家逸, 杨贝宁, 等. 非线性信息瓶颈指导的层次图结构学习方法. 中国科学: 信息科学, 2024, 54: 2409–2427, doi: 10.1360/SSI-2024-0049
Sun Q Y, Luo J Y, Yang B N, et al. Hierarchical graph structure learning with nonlinear information bottleneck (in Chinese). Sci Sin Inform, 2024, 54: 2409–2427, doi: 10.1360/SSI-2024-0049

© 2024《中国科学》杂志社 www.scichina.com infocn.scichina.com

https://www.sciengine.com/doi/10.1360/SSI-2024-0049

的能力, 但它们大多假设观察到的结构完美地描述了节点之间的关联关系, 并将图数据的原始结构作为信息传递的途径, 其计算结构与图结构耦合, 对数据的质量极为敏感. 然而, 现实生活中的图数据大多是直接采集的, 这些原始图或是由结构化数据自然形成的 (如社交网络), 或者由一些预定义的规则从原始特征空间中构建的, 而这些规则通常与下游任务无关, 这导致了原始图结构与适用于特定任务的最佳图结构之间存在偏差; 此外, 开放环境下的图数据存在大量随机噪声或人为扰动, 例如欺诈者为了逃避检测会伪装与正常用户的关联. 上述结构噪声信息会导致图神经网络学习到的表征质量严重下降, 鲁棒性差 [5,6]. 随着在工业、金融、风控等关键安全领域的深入应用, 图神经网络模型面临开放的数据环境以及复杂的决策任务, 对其准确性和安全性提出了更高要求. 因此, 如何找到与下游任务密切相关的结构, 避免噪声数据干扰, 提高图表征的表达能力和鲁棒性, 是图神经网络进一步发展和应用亟待解决的问题.

为了优化图结构的质量、提高图表征学习的鲁棒性, 相关研究提出了一系列图结构学习 (graph structure learning) 方法 [7,8], 其中大多数方法是基于对结构的先验假设 (如假设图中存在社区结构) 或显式的结构约束 (如稀疏性 [9]、低秩性 [10] 和特征平滑性 [10]), 并针对下游任务优化邻接矩阵和模型参数. 然而, 这些结构性质假设或显式约束并不一定适用于所有的数据集和任务. 此外, 大部分方法都是针对节点级分类任务设计 [11], 难以适用于图层级的分类或回归任务. 图层级任务主要关注对整个图属性的预测, 包括图分类 (graph classification)、图回归 (graph regression) 等, 在实际应用中均有丰富的场景. 例如: 在社交领域中可通过社团网络进行分类以发现目标群体 [12], 对微博转发网络进行分析以发现谣言信息 [13]; 在生化领域可以通过对分子图进行分类实现对某个分子的性质进行预判 [14]; 在交通领域可以通过对交通路网进行图回归分析以预测其平均到达时间 [15]. 少量图层级的结构学习方法也仅关注于节点间的连接优化, 忽略了对于图层级任务十分重要的层次结构信息 [16]. 而层次化特性是对图层级任务非常重要的特性之一. 例如: 社交网络中存在用户之间的直接连接、共同兴趣的群体、社区结构等不同层次的关系; 分子图数据中存在原子、功能基团、分子等不同层次的结构特性; 因特网、物流网络、金融网络中的节点也通常具有类似于组织结构的层级隶属关系. 因此, 目前仍然缺乏能够统一度量图结构的质量、去除图结构噪声信息, 并保留图结构层次特性的图层级结构学习框架.

回顾上述存在的问题, 结构学习的关键是学习与任务密切相关的底层连接结构. 在信息论中, 信息瓶颈 (information bottleneck, IB) 原则 [17] 为度量数据中与任务有关的信息提供了一个有效的框架. 信息瓶颈原则是对数据压缩率失真理论 (rate distortion theory) 的扩展, 通过在有效信息保留 (relevant information preservation) 和噪声信息压缩 (noisy information compression) 之间进行权衡, 约束输出中与任务无关的信息. 具体来说, 信息瓶颈原则旨在寻求一种对目标信息量最大 (即互信息最大), 而对输入数据信息量最小 (即互信息最小) 的表征. 基于信息瓶颈原则, 学习到的表征对数据噪声更加鲁棒 [18,19]. 本文将信息瓶颈原则扩展到图数据上, 并针对图结构的层次特性, 提出了一个非线性信息瓶颈 [20] 指导的层次图结构学习方法 NIB-HGSL (hierarchical graph structure learning with nonlinear information bottleneck). NIB-HGSL 针对图结构中的不同层级, 在信息瓶颈目标函数的指导下, 通过采用无关特征掩码和结构学习方法, 生成一个新的最小充分图, 为下游任务提炼出更加密切相关的图特征与结构信息, 进而获得更鲁棒的图表征. NIB-HGSL 方法主要包括以下三个步骤: (1) 对于每个结构层次, 最小充分图生成器模块通过对不相关的节点特征进行掩码, 并基于掩码后的特征来学习最小充分图. (2) 图神经网络模块将每个层次的最小充分图作为输入, 学习图表征和其分布. (3) 对各个层次的图表征进行解码并用于分类, 学习其与图标签的联合分布. 整个框架可以用图分类交叉熵损失函数和图表征向量分布的 KL 散度来有效地优化信息瓶颈目标.

本文的主要贡献包括 3 个方面: (1) NIB-HGSL 将信息瓶颈原理引入到层次化的图结构学习中,

https://www.sciengine.com/doi/10.1360/SSI-2024-0049

并结合变分近似获得易于优化的目标函数, 从信息论的角度为图结构质量的信息度量提供了一个更加普适和通用的框架. (2) NIB-HGSL 通过层次图结构学习, 可以保留图数据中不同层级的结构特性并针对性地保留关键信息、去除无关信息. NIB-HGSL 是模型无关的即插即用式框架, 可以与现有图神经网络结合以提高其性能. (3) 在多个图分类和图回归数据集上以及在噪声图以及对抗攻击图上的实验结果表明, 与其他的基线方法相比, NIB-HGSL具备更佳的准确性和鲁棒性.

2 相关工作

本节首先介绍图结构学习方法, 然后简要介绍信息论视角下的图结构度量方法.

2.1 图结构学习方法

图结构学习 [8] 的目标是为输入的图数据学习一个更好的图结构, 并联合优化后的图结构学习表征, 以提高图神经网络模型的表征质量和鲁棒性. 部分工作 (例如 NeuralSparse [11] 和 PTDNet [21]) 将图结构学习建模为一个图稀疏化问题, 并通过删除与任务无关的边来生成新的结构. 这些方法只能生成稀疏图, 不考虑恢复原始图数据中缺失的边. 其他工作通过学习一个新的图结构作为图神经网络的输入, 根据对结构进行建模不同可分为三类方法: 度量学习类、概率建模类、直接优化类. 其中, 度量学习类方法通过对每个节点对特征进行学习获得其边的权重 [11,22], 例如 IDGL [9] 将图结构学习转化为相似度度量学习问题, 迭代学习图结构和节点表征; 概率建模类方法假设图是由一定分布的采样过程生成的, 用可学习的参数对采样边的概率进行建模 [23], 例如 LDS [22] 通过双层优化技术联合学习边的离散概率分布和图神经网络参数; 直接优化类方法将图数据的邻接矩阵作为一个参数, 和图神经网络的参数一起进行优化, 例如 GCN-GT [24] 通过类别标签和平滑性正则同时优化邻接矩阵参数和图神经网络参数.

大多数方法对学习到的图的结构进行一定的约束, 来控制图结构质量, 例如连通性 [25]、低秩 [10]、稀疏性 [9]、特征平滑度 [10] 和同质性 [26]. 此外, 大部分方法针对节点层级的任务设计, 仅有两项工作 HGP-SL [27] 和 VIB-GSL [23] 专门面向图级别的任务进行结构学习. HGP-SL [27] 首先通过节点选择进行池化, 再通过结构学习捕捉节点间关联. VIB-GSL [23] 和本文工作最为接近, 通过变分信息瓶颈来指导图结构的学习, 然而, 其忽略了对于图层级任务极其重要的子结构, 导致大量结构信息丢失. 本文所提方法 NIB-GSL 使用非线性信息瓶颈原则作为指导, 避免了先验性结构假设, 并通过层次性的结构学习保留关键信息, 对不同的图结构和任务具有更强的适应性.

2.2 信息论视角的图结构度量

信息论将信息传递作为一种统计现象来考虑, 为复杂系统的描述提供了一套统一的方式, 越来越多的研究工作将其引入到表征学习中. 信息的度量是信息论研究的基本问题之一, 其中, C. E. Shannon 于 1948 年提出信息熵 (information entropy) [28], 用于度量信息的冗余度 (即复杂性). 为衡量图结构的复杂性, 研究者将熵这一度量信息复杂性的方法扩展到图数据上, 其代表性工作包括冯诺依曼图熵 [29] 和结构熵 [30]. 冯诺依曼图熵 [29] 将图看作一个量子系统, 通过离散的拉普拉斯态和量子态之间的映射, 量化图结构的谱复杂性. 结构熵 [30] 将 K 维结构复杂性定义为确定图结构中可从随机游走中访问的节点的 K 维编码所需的最小总位数. 目前的图熵方法只针对图结构本身的复杂性定义, 难以同时考虑图数据中的属性等特征信息以及下游任务.

互信息 (mutual information, MI) 是在深度学习领域应用较为广泛的一个信息论工具, 它衡量了

https://www.sciengine.com/doi/10.1360/SSI-2024-0049

两个变量之间的相关性. 若两个变量相关性越大, 则其互信息越大, 反之如果两个变量相关独立, 则其互信息为 0. 常用互信息估计方法包括 MINE (mutual information neural estimation) [31]、变分推断 [32]. 其中, MINE 方法通过 DV (Donsker-Varadhan) 表示来逼近 KL 散度, 并通过神经网络作为统计网络来找到一个互信息的较为紧致的下界. 变分推断方法通过变分方法求似然函数的近似分布, 从而获得对互信息的近似估计. 对于图数据的互信息估计来说, 除了上述不同估计方式之外, 还可以分为两类: 直接估计和分解估计. 其中, 大部分现有图上互信息估计方法属于直接估计, 即将图数据作为整体, 将其转化为向量进行互信息行计算. 分解估计在领域内研究较少, 代表性方法为 GIB [33], 这类方法将图数据分解为节点特征矩阵和邻接矩阵, 并对两部分的互信息分别计算并加和作为整体互信息.

近年来, 信息瓶颈原则在深度学习中受到了极大的关注 [17, 34], 部分研究者将其用于解释深度神经网络的训练过程以及指导神经网络学习中. 最近一些工作 [33, 35] 将信息瓶颈原则引入到图学习中. GIB [33] 将一般的信息瓶颈扩展到具有不规则结构的图数据中, 对结构和特征信息进行正则化处理, 以获得鲁棒的节点表征. SIB [35] 通过 “子图 - 图” 与 “子图 - 标签” 间互信息的均衡发现对下游任务最具影响力的关键子图. HGIB [36] 以无监督的方式实现异构图的共识假设. 当前信息论指导的图结构度量方法只针对图数据中观测的结构信息, 极少关注结构中缺失的边. VIB-GSL [23] 首次将信息瓶颈理论引入图结构学习中, 并采取变分近似的方式获取在图数据上易于优化的目标函数. 本文提出的层次图结构学习方法 NIB-HGSL同样从信息论的视角出发, 度量特定下游任务下的图结构质量, 旨在去除结构中的噪声连接、恢复观测结构中缺失的隐藏连接, 提高图表征的鲁棒性.

3 问题背景与定义

本节首先简要概述图结构学习的相关定义, 然后介绍信息瓶颈原则及其在图结构学习中的应用.

图结构学习. 图结构学习 [8] 的目标是同时学习一个优化的图结构和相应的节点/图表征, 以提高图神经网络模型的表征质量和鲁棒性. 本文工作主要专注于图层级分类和回归任务的图结构学习, 可以将其定义为: 设 $G \in \mathbb{G}$ 是一个带有标签 $Y \in \mathbb{Y}$ 的图, 其节点集为 V, 节点特征矩阵为 $\mathbf{X} \in \mathbb{R}^{|V| \times d}$, 邻接矩阵为 $\mathbf{A} \in \mathbb{R}^{|V| \times |V|}$. 本文关注的图结构学习问题可以被表述为: 针对下游的图分类或图回归任务, 输出一个优化的图 $G^* = (\mathbf{X}^*, \mathbf{A}^*)$ 和其相应的图表征 Z^*, 并学习一个分类/预测器 $f : Z^* \to Y$ 获得图标签, 其中 Y 对于图分类任务来说为分类类别, 对于图回归任务来说为回归值.

常见的信息论定义与计算方法. 根据 [37], 对于变量 $X \sim p(X)$, 其香农熵 $H(X) = \mathbb{E}_{X \sim p(X)}[-\log p(X)]$, 交叉熵 $H(p(X), q(X)) = \mathbb{E}_{X \sim p(X)}[-\log q(X)]$, 香农互信息 $\mathcal{I}(X;Y) = H(X) - H(X|Y)$, Kullback-Leibler divergence(KL) 散度 $\mathrm{KL}(p(X)||q(X)) = \mathbb{E}_{X \sim p(X)}[-\log \frac{p(X)}{q(X)}]$.

信息瓶颈原则. 信息瓶颈原则 [17] 是对数据压缩的率失真问题的拓展, 旨在寻求数据拟合和泛化之间的平衡, 使用互信息作为损失函数和实现正则化. 按照信息瓶颈的标准定义, 对于给定的数据 X 的表征 Z 和目标标签 Y, 三元组 (X, Y, Z) 服从马尔科夫链 $\langle Y \to X \to Z \rangle$.

定义1 (信息瓶颈原则) 对于输入数据 X 和它的标签 Y, 信息瓶颈原则的目的是学习最小充分表征 Z, 使得

$$Z = \arg\min_{Z} -\mathcal{I}(Z;Y) + \beta \mathcal{I}(Z;X), \tag{1}$$

其中, β 是拉格朗日乘子 (Lagrangian multiplier), 用于权衡表征的充分性和最小信息量.

层次图结构学习的信息瓶颈原则. 本文提出的层次图结构学习方法方法旨在学习一个层次最小充分图 $G_{\mathrm{IB}} : \{G_{\mathrm{IB}}^{(1)}, G_{\mathrm{IB}}^{(2)}, \cdots, G_{\mathrm{IB}}^{(L)}\}$, 其中 $G_{\mathrm{IB}}^{(l)} = (\mathbf{X}_{\mathrm{IB}}, \mathbf{A}_{\mathrm{IB}})$, 从而以最小的有效信息损失保留了图 G

https://www.sciengine.com/doi/10.1360/SSI-2024-0049

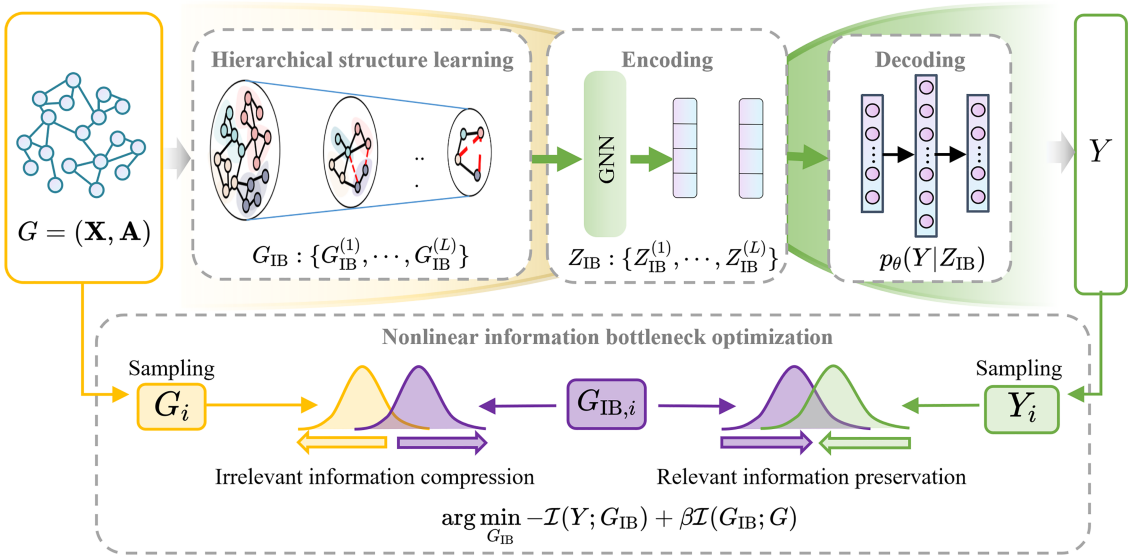


图 1 (网络版彩图) NIB-HGSL 整体框架

Figure 1 (Color online) Framework of NIB-HGSL

的性质, 其定义如下:

定义2 (IB-Graph) 对于一个图 $G = (\mathbf{X}, \mathbf{A})$ 和它的标签 Y, 通过信息瓶颈找到的最优图 G_{IB}: $\{G_{\mathrm{IB}}^{(1)}, G_{\mathrm{IB}}^{(2)}, \cdots, G_{\mathrm{IB}}^{(L)}\}$ 作为最小充分图 IB-Graph:

$$G_{\mathrm{IB}} = \arg\min_{G_{\mathrm{IB}}} -\mathcal{I}(G_{\mathrm{IB}}; Y) + \beta\mathcal{I}(G_{\mathrm{IB}}; G), \tag{2}$$

其中, $G_{\mathrm{IB}}^{(l)} = (\mathbf{X}_{\mathrm{IB}}^{(l)}, \mathbf{A}_{\mathrm{IB}}^{(l)})$ 是优化后的第 l 层图, $\mathbf{X}_{\mathrm{IB}}^{(l)}$ 是其特征集合, $\mathbf{A}_{\mathrm{IB}}^{(l)}$ 是其邻接矩阵.

直观地说, 第一项 $-\mathcal{I}(G_{\mathrm{IB}}; Y)$ 是有效信息保留项, 该项鼓励 G_{IB} 中保留与标签 Y 相关的信息. 第二项 $\mathcal{I}(G_{\mathrm{IB}}; G)$ 是噪声信息压缩项, 该项鼓励学习到的最小充分图中不包含中 G 与标签无关的信息. 而拉格朗日乘子 β 表示信息压缩的程度, 越大的 β 表示 G_{IB} 中保留 G 的信息越少. 假设 $G_n \in \mathbb{G}$ 是 G 中一个任务无关的干扰因素, 即影响观测到的图数据 G, 但与标签 Y 无关的随机变量, 则 G_{IB} 的学习过程遵循马尔科夫链 [18] $\langle (Y, G_n) \to G \to G_{\mathrm{IB}} \rangle$, 也就是 G_{IB} 与 Y 和 G_n 之间只通过 G 进行关联, 这使其只保留了原始图中与任务相关的信息, 对数据中的干扰具有扰动不变性.

4 非线性信息瓶颈指导的层次图结构学习 NIB-HGSL

NIB-HGSL 在信息瓶颈原则的指导下, 通过学习具有多层次的图结构最小充分图 IB-Graph 来保留图数据每一层与任务密切相关的关键信息, 并将其作为图神经网络消息传递的途径, 以此来深入挖掘和发现图结构不同尺度的结构关联, 更好地去除图中各层的噪声信息、保留层次化关键信息. 其整体结构如图 1 所示.

本节首先介绍层次化的图结构学习方法, 然后介绍信息瓶颈原则在不规则层次图结构上的近似上界推导, 最后介绍框架整体优化方法.

https://www.sciengine.com/doi/10.1360/SSI-2024-0049

4.1 层次图结构学习方法

本节介绍层次图结构学习方法, 通过图池化技术递归地聚合节点集以形成超节点, 生成不同结构粒度的层次图, 再通过结构学习对多层次的图结构进行优化, 从而捕捉图中不同粒度的显著信息.

定义 $\mathbf{S}^{(l)} \in \mathbb{R}^{N_l \times N_{l+1}}$ 表示在第 l 层的图池化算子, N_l 和 N_{l+1} 分别代表图在第 l 和第 $l+1$ 层时的节点数量. 聚类时采用软聚类方式, 即节点 v_i 可以同时属于多个类别, $\mathbf{S_{i,j}}$ 表示节点 v_i 隶属节点簇 j 的程度, 并且 $\sum_j \mathbf{S}_{i,j} = 1$. 假设 $\mathbf{A}^{(l)}$ 和 $\mathbf{X}^{(l)}$ 分别代表经过 $l-1$ 层池化后的邻接矩阵和节点特征矩阵, 而 $\mathbf{A}_{\mathrm{IB}}^{(l)}$ 和 $\mathbf{X}_{\mathrm{IB}}^{(l)}$ 分别代表经过结构优化后第 l 层的邻接矩阵和节点特征矩阵, 可以通过以下两个公式生成池化后第 $l+1$ 层的邻接矩阵 $\mathbf{A}^{(l+1)}$ 和节点特征矩阵 $\mathbf{X}^{(l+1)}$:

$$\mathbf{X}^{(l+1)} = (\mathbf{S}^{(l)})^{\mathrm{T}} \mathbf{X}_{\mathrm{IB}}^{(l)}, \ \mathbf{A}^{(l+1)} = (\mathbf{S}^{(l)})^{\mathrm{T}} \mathbf{A}_{\mathrm{IB}}^{(l)} \mathbf{S}^{(l)}, \tag{3}$$

其中 $\mathbf{X}_{\mathrm{IB}}^{(0)}, \mathbf{A}_{\mathrm{IB}}^{(0)}$ 为初始化图 G 的原始节点特征矩阵 $\mathbf{X}$ 和邻接矩阵 $\mathbf{A}$.

直观上, 具有相似特征的节点被分配到相同聚类的概率较高, 因此我们基于注意力机制来获得图池化算子 $\mathbf{S}^{(l)}$, 注意力分数由节点对间的特征相似程度和连接性决定:

$$\phi_{i,j}^{(l)} = \frac{(x_i^{(l)})^{\mathrm{T}} x_j^{(l)}}{\sum_{k \neq i} (x_i^{(l)})^{\mathrm{T}} x_k^{(l)}} (\mathbf{A}_{\mathrm{IB}}^{(l)})_{i,j}^{c}. \tag{4}$$

其中, $x_i^{(l)}, x_j^{(l)}$ 为节点 v_i, v_j 的特征向量, 其中 $(\mathbf{A}_{\mathrm{IB}}^{(l)})^c$ 表示节点在 l 层的 c-hop 连接性, c 为一个超参数, $\phi_{i,j}$ 代表节点对 (v_i, v_j) 间的注意力分数, 衡量了两节点间关联的重要程度. 有了注意力分数矩阵 ϕ, 我们可以得到最终的图池化算子:

$$\mathbf{S}^{(l)} = \mathrm{softmax}(\phi^{(l)} \mathbf{U}^{(l)}), \ \mathbf{U}^{(l)} = \mathrm{GNN}(X_{\mathrm{IB}}^{(l)}, A_{\mathrm{IB}}^{(l)}) \in \mathbb{R}^{N_l \times N_{l+1}}, \tag{5}$$

其中, $\mathrm{GNN}(\cdot)$ 为基础图神经网络编码器. 此处对于带有边/关系特征的图数据, 可以选择考虑边信息或属性的图神经网络作为基础编码器, 如 RGCN [38]、RGAT [39] 等, 以保留边上的特征信息, 进而将这类特征嵌入到学习的图结构中.

通过图池化技术纯化信息的一个关键问题是, 无法正确学习到高层次节点间的结构信息 [40]. 在这项工作中, 我们保留通过聚类所得到的原始结构信息 $\mathbf{A}^{(l+1)}$, 同时引入结构学习来捕获高层次中超节点之间所隐含的结构关系. 首先使用特征掩码策略, 以离散的方式丢弃 $\mathbf{X}^{(l)}$ 中与任务无关特征:

$$\mathbf{X}_{\mathrm{SL}}^{(l+1)} = \{\mathbf{X}_i^{(l+1)} \odot \mathbf{M}, i = 1, 2, .., |V|\}, \tag{6}$$

其中, $\mathbf{M}$ 是一个可学习的二元特征掩码, $\odot$ 是元素间的乘积. 直观地说, 如果一个特定的特征维度与任务不相关, 那么 $\mathbf{M}$ 中相应的权重就会接近于零. 这里使用重参数化技巧 (reparameterization trick) [41] 通过 d 维随机变量进行反向传播, 可以将重参数化为:

$$\mathbf{X}_{\mathrm{SL}}^{(l+1)} = \mathbf{X}_r^{(l+1)} + (\mathbf{X}^{(l+1)} - \mathbf{X}_r^{(l+1)}) \odot \mathbf{M}, \tag{7}$$

其中, $\mathbf{X}_r^{(l+1)}$ 是一个从 $\mathbf{X}^{(l+1)}$ 的经验分布中采样的随机变量.

其次, 对最小充分图的结构进行学习. 如第 2.1 节中所介绍, 通过概率建模是图结构学习的典型建模方式, 具体而言可以选择伯努利分布 (Bernoulli distribution) [22, 42]、类别分布 (categorical distribution) [11, 33]、高斯核函数 (Gaussian kernel) 定义分布 [43] 等. 其中, 类别分布和高斯核函数分布更适合于对于图结构分布有先验知识的情况, 因此本文选择了更具有普适性的伯努利分布进行结

https://www.sciengine.com/doi/10.1360/SSI-2024-0049

构建模. 具体而言, 本文将池化子图间所有可能的边建模为一组相互独立的伯努利随机变量, 其参数是一个可学习的注意力权重:

$$\mathbf{A}_{\mathrm{SL}}^{(l+1)}=\bigcup_{i,j\in V^{l+1}}\left\{a_{i,j}^{(l+1)}\sim\mathrm{Ber}\left(\pi_{i,j}\right)\right\}. \tag{8}$$

对于每一对节点, 结合其表征来优化边的采样概率. $\pi_{i,j}$ 描述了边的对于特定任务的质量, 数值小的表示边更可能是噪声信息, 应该分配更小的权重甚至删除. 对于一对节点 (v_i,v_j), 结合其节点表征计算其边的采样概率. 具体而言, 将节点的特征转化为一个特征隐向量:

$$\mathbf{h}_i^{(l+1)}=\mathrm{NN}\left(\mathbf{x}_{\mathrm{SL}}^{(l+1)}\left(v_i\right)\right), \tag{9}$$

其中 $\mathbf{x}_{\mathrm{SL}}^{(l+1)}\left(v_i\right)$ 为节点 v_i 的特征, $\mathrm{NN}(\cdot)$ 表示一个神经网络, 本工作中使用两层感知机实现. 然后, 使用多头余弦相似度机制, 结合节点特征隐向量和节点表征向量学习连边采样概率:

$$\pi_{i,j}^{(l+1)}=\frac{1}{m}\cos\left(\mathbf{W}_h\cdot\left(\mathbf{x}_i^{(l+1)}|\mathbf{h}_i^{(l+1)}\right),\mathbf{W}_h\cdot\left(\mathbf{x}_j^{(l+1)}|\mathbf{h}_j^{(l+1)}\right)\right), \tag{10}$$

其中, m 为多头的个数, $\mathbf{W}_h$ 为第 h 个头的权重参数. 由于 $\mathbf{A}^l$ 相对于 π 的伯努利分布是不可微的, 因此此处使用伯努利分布的具体松弛 (concrete relaxation) [44] 来更新:

$$\mathrm{Ber}(\pi_{i,j}^{(l+1)})\approx\mathrm{sigmoid}\left(\frac{1}{t}\left(\log\frac{\pi_{i,j}^{(l+1)}}{1-\pi_{i,j}^{(l+1)}}+\log\frac{\epsilon}{1-\epsilon}\right)\right), \tag{11}$$

其中, $\epsilon\sim\mathrm{Uniform}(0,1)$ 和 $t\in\mathbb{R}^+$ 是松弛分布的温度系数. 在具体松弛之后, $a_{i,j}^{l+1}$ 被转化为 $\pi_{i,j}^{l+1}$ 和 ϵ 的一个确定性函数. 具体松弛后的图结构是一个加权的全连接图, 在此之上计算的代价很高. 本文通过屏蔽小于非负阈值的矩阵元素来提取一个对称稀疏的邻接矩阵. 最后得到高一层次的图:

$$\mathbf{A}_{\mathrm{IB}}^{(l+1)}=\lambda\mathbf{A}_{\mathrm{SL}}^{(l+1)}+(1-\lambda)\mathbf{A}^{(l+1)},\ \mathbf{X}_{\mathrm{IB}}^{(l+1)}=\mathbf{X}_{\mathrm{SL}}^{(l+1)}, \tag{12}$$

其中 λ 是一个预定义的超参数, 用于控制优化后的图结构与原始结构的融合程度.

4.2 信息瓶颈目标函数的非线性近似

经过 4.1 中的步骤, 可以得到更新后的层次化图 $G_{\mathrm{IB}}:\{G_{\mathrm{IB}}^{(1)},G_{\mathrm{IB}}^{(2)},\cdots,G_{\mathrm{IB}}^{(L)}\}$. 由于图数据的非欧几里得性质, 其互信息难以计算, 因此, 将信息论应用于图数据的主要挑战之一是对高维联合分布的合理估计 [45]. 公式 (2) 中的信息瓶颈目标难以直接优化, 因此无法直接用于指导层次图结构学习. 接下来, 本文针对图数据的特性, 分别提出了对有效信息保留项和噪声信息压缩项的变分近似上界.

首先, 对于信息瓶颈目标函数中的压缩项 $\mathcal{I}(G_{\mathrm{IB}};G)$, 此项旨在使得 G_{IB} 包含 G 中尽量少的与任务无关的信息, 其上界为 [23,35]:

命题1 ($\mathcal{I}(G_{\mathrm{IB}};G)$ 的上界) 对于图 $G\in\mathbb{G}$、从中学到的最小充分层次图 G_{IB}, 有

$$\mathcal{I}(G_{\mathrm{IB}};G)\leqslant\hat{\mathcal{I}}(G_{\mathrm{IB}};G):=\sum_{l=1}^{L}\mathbb{E}_{p(G_{\mathrm{IB}}^{(l)},G)}\log\frac{p(G_{\mathrm{IB}}^{(l)}|G)}{r(G_{\mathrm{IB}}^{(l)})}, \tag{13}$$

其中, $r(G_{\mathrm{IB}}^{(l)})$ 为 $p(G_{\mathrm{IB}}^{(l)})$ 的变分近似.

https://www.sciengine.com/doi/10.1360/SSI-2024-0049

证明 由于层次化图 G_{IB} 的学习过程中, 每个层级的图 $G_{\mathrm{IB}}^{(l)}$ 结构学习都依赖于上一层的图 $G_{\mathrm{IB}}^{(l-1)}$, 因此 G_{IB} 的学习过程遵循马尔科夫 (Markov) 链 [18]$\langle (Y, G_n) \to G \to G_{\mathrm{IB}} \rangle$, 即 $\langle (Y, G_n) \to G \to G_{\mathrm{IB}}^{(1)} \to G_{\mathrm{IB}}^{(2)} \to \cdots \to G_{\mathrm{IB}}^{(L)} \rangle$. 根据互信息的链式法则 (chain rule for mutual information) [37] 和数据处理不等式 (data processing inequality), 有

$$\mathcal{I}(G; G_{\mathrm{IB}}) = \mathcal{I}(G; G_{\mathrm{IB}}^{(1)}, \cdots, G_{\mathrm{IB}}^{(L)}) = \sum_{l=1}^{L} \mathcal{I}(G_{\mathrm{IB}}^{(l)}; G | G_{\mathrm{IB}}^{(l-1)}) = \sum_{l=1}^{L} \mathcal{I}(G; G_{\mathrm{IB}}^{(l)} | G_{\mathrm{IB}}^{(l-1)}) \leqslant \sum_{l=1}^{L} \mathcal{I}(G; G_{\mathrm{IB}}^{(l)}). \tag{14}$$

对于 $\mathcal{I}(G; G_{\mathrm{IB}}^{(l)})$, 根据互信息的定义, 有

$$I(G_{\mathrm{IB}}^{(l)}; G) = \iint p(G_{\mathrm{IB}}^{(l)}; G) \log \frac{p(G_{\mathrm{IB}}^{(l)} | G)}{p(G_{\mathrm{IB}}^{(l)})} \mathrm{d}G_{\mathrm{IB}}^{(l)} \mathrm{d}G. \tag{15}$$

由于 $p(G_{\mathrm{IB}}^{(l)}) = \int p(G_{\mathrm{IB}}^{(l)} | G) p(G) \mathrm{d}G$ 难以计算, 因此我们使用 $r(G_{\mathrm{IB}}^{(l)})$ 作为 $p(G_{\mathrm{IB}}^{(l)})$ 的变分近似. 根据 KL 散度的非负性, 有 $\mathcal{D}_{\mathrm{KL}}(p(Z) || r(Z)) \geqslant 0 \Longrightarrow \int p(z) \log p(z) \mathrm{d}z \geqslant \int p(z) \log r(z) \mathrm{d}z$.

因此, 有

$$I(G_{\mathrm{IB}}^{(l)}; G) \leqslant \iint p(G_{\mathrm{IB}}^{(l)}, G) \log \frac{p(G_{\mathrm{IB}}^{(l)} | G)}{r(G_{\mathrm{IB}}^{(l)})} \mathrm{d}G_{\mathrm{IB}}^{(l)} \mathrm{d}G = \mathbb{E}_{p(G_{\mathrm{IB}}^{(l)}, G)} \log \frac{p(G_{\mathrm{IB}}^{(l)} | G)}{r(G_{\mathrm{IB}}^{(l)})}. \tag{16}$$

将公式 (16) 代入公式 (14), 有

$$\mathcal{I}(G_{\mathrm{IB}}; G) \leqslant \hat{\mathcal{I}}(G_{\mathrm{IB}}; G) := \sum_{l=1}^{L} \mathbb{E}_{p(G_{\mathrm{IB}}^{(l)}, G)} \log \frac{p(G_{\mathrm{IB}}^{(l)} | G)}{r(G_{\mathrm{IB}}^{(l)})}. \tag{17}$$

其次, 对于信息瓶颈目标函数中的有效信息保留项 $\mathcal{I}(G_{\mathrm{IB}}; Y)$, 此项旨在鼓励保留对图标签的有效信息. 通过对其进行变分近似, 获得 $-\mathcal{I}(G_{\mathrm{IB}}; Y)$ 的上界 [23, 35].

命题2 ($-\mathcal{I}(G_{\mathrm{IB}}; Y)$ 的上界) 对于带有标签 $Y \in \mathbb{Y}$ 的图 $G \in \mathbb{G}$、从中学到的最小充分层次图 G_{IB}, 有

$$-\mathcal{I}(G_{\mathrm{IB}}; Y) \leqslant -\mathbb{E}_{p(Y, G_{\mathrm{IB}})} \left[\log q_\theta \left(Y | G_{\mathrm{IB}} \right) \right] - H(Y), \tag{18}$$

其中, $q_\theta(Y | G_{\mathrm{IB}})$ 是真实后验分布 $p(Y | G_{\mathrm{IB}})$ 的变分近似.

证明

$$-\mathcal{I}(G_{\mathrm{IB}}; Y) = -\iint p(Y, G_{\mathrm{IB}}) \log \frac{p(Y, G_{\mathrm{IB}})}{p(Y) p(G_{\mathrm{IB}})} \mathrm{d}Y \mathrm{d}G_{\mathrm{IB}} = -\iint p(Y, G_{\mathrm{IB}}) \log \frac{p(Y | G_{\mathrm{IB}})}{p(Y)} \mathrm{d}Y \mathrm{d}G_{\mathrm{IB}}, \tag{19}$$

其中, 由马尔科夫链 $\langle (Y, G_n) \to G \to G_{\mathrm{IB}} \rangle$, $p(Y | G_{\mathrm{IB}})$ 可定义为: $p(Y | G_{\mathrm{IB}}) = \int p(Y | G_{\mathrm{IB}}) p(G_{\mathrm{IB}} | G) \mathrm{d}G$. 因为 $p(Y | G_{\mathrm{IB}})$ 难以直接计算, 令 $q_\theta(Y | G_{\mathrm{IB}})$ 作为真实先验分布 $p(Y | G_{\mathrm{IB}})$ 的变分近似, 其中 θ 是变分解码映射的参数. 根据 KL 散度的非负性 $\mathcal{D}_{\mathrm{KL}}(p(Y | G_{\mathrm{IB}}) || q_\theta(Y | G_{\mathrm{IB}})) \geqslant 0$, 即

$$\int p(Y | G_{\mathrm{IB}}) \log p(Y | G_{\mathrm{IB}}) \mathrm{d}Y \mathrm{d}G_{\mathrm{IB}} \geqslant \int p(Y | G_{\mathrm{IB}}) \log q_\theta(Y | G_{\mathrm{IB}}) \mathrm{d}Y \mathrm{d}G_{\mathrm{IB}} \tag{20}$$

将公式 (20) 代入公式 (19), 有

$$\begin{aligned} -\mathcal{I}(G_{\mathrm{IB}}; Y) &\leqslant -\iint p(Y, G_{\mathrm{IB}}) \log \frac{q_\theta(Y | G_{\mathrm{IB}})}{p(Y)} \mathrm{d}Y \mathrm{d}G_{\mathrm{IB}} \\ &= -\iint p(Y, G_{\mathrm{IB}}) \log q_\theta(Y | G_{\mathrm{IB}}) \mathrm{d}Y \mathrm{d}G_{\mathrm{IB}} - H(Y) = -\mathbb{E}_{p(Y, G_{\mathrm{IB}})} \left[\log q_\theta \left(Y | G_{\mathrm{IB}} \right) \right] - H(Y). \end{aligned} \tag{21}$$

https://www.sciengine.com/doi/10.1360/SSI-2024-0049

本文采用充分编码器在表征过程中不存在信息损失的假设 [46], 这也是信息瓶颈已有相关工作 [47] 的常用假设. 对于每一个层级的图 $G_{\mathrm{IB}}^{(l)}$, 将其输入到图神经网络中进行学习其节点表征, 并通过池化层将节点表征转化为图表征 $Z_{\mathrm{IB}}^{(l)}$. 通过这种方法, 可以获得图的层次化表征 $\{Z_{\mathrm{IB}}^{(1)}, Z_{\mathrm{IB}}^{(2)}, \cdots, Z_{\mathrm{IB}}^{(L)}\}$. 进而, 可以通过注意力加权的方式将层次化表征转化为一个统一表征向量 $Z_{\mathrm{IB}} = w_1 Z_{\mathrm{IB}}^{(1)} + w_1 Z_{\mathrm{IB}}^{(2)} + \cdots + w_L Z_{\mathrm{IB}}^{(L)}$. 因此, 对于学习到的层次图结构 G_{IB} 来说, 有

$$\mathcal{I}(G_{\mathrm{IB}}^{(l)}; Y) \approx \mathcal{I}(Z_{\mathrm{IB}}^{(l)}; Y),\ \mathcal{I}(G_{\mathrm{IB}}; G) \approx \mathcal{I}(Z_{\mathrm{IB}}; G). \tag{22}$$

结合式 (18), (13) 和 (22), 可以为层次图结构学习的信息瓶颈目标函数提供可处理的上界:

$$-\mathcal{I}(G_{\mathrm{IB}}; Y) + \beta\mathcal{I}(G_{\mathrm{IB}}; G) \leqslant -\mathbb{E}_{p(Y, Z_{\mathrm{IB}})}\left[\log q_\theta\left(Y|Z_{\mathrm{IB}}\right)\right] - H(Y) + \beta\hat{\mathcal{I}}(Z_{\mathrm{IB}}; G). \tag{23}$$

4.3 非线性信息瓶颈的目标函数优化

基于上节中对信息瓶颈原则中有效信息保留项与噪声信息压缩项的互信息上界近似, NIB-HGSL可以使用基于梯度下降的反向传播技术在训练数据样本上有效地计算信息瓶颈目标函数的上界. 具体来说, 通过对 $G_{\mathrm{IB}} : \{G_{\mathrm{IB}}^{(1)}, G_{\mathrm{IB}}^{(2)}, \cdots, G_{\mathrm{IB}}^{(L)}\}$ 在所有结构层次上优化非线性信息瓶颈目标函数, 获得保留最小充分信息的图结构与对应的层次化图表征:

$$\mathcal{L}_{\mathrm{NIB}} = \sum_{l=1}^{L}\left\{\mathbb{E}_{p(Y, Z_{\mathrm{IB}}^{(l)})}\left[\log q_\theta\left(Y|G_{\mathrm{IB}}^{(l)}\right)\right] - \beta\hat{\mathcal{I}}(Z_{\mathrm{IB}}^{(l)}; G)\right\}. \tag{24}$$

由于有关研究表明, 信息瓶颈曲线 (IB curve) 并不能通过最大化不同 β 下的拉格朗日项而恢复 [48], 因此, 在实际优化时, 本文采用平方拉格朗日项 (squared-IB Lagrangian)[48] 来进一步探索在不同值下的信息瓶颈曲线, NIB-HGSL优化损失函数为

$$\mathcal{L}_{\mathrm{SNIB}} = \sum_{l=1}^{L}\left\{\mathbb{E}_{p(Y, Z_{\mathrm{IB}}^{(l)})}\left[\log q_\theta\left(Y|G_{\mathrm{IB}}^{(l)}\right)\right] - \beta\hat{\mathcal{I}}(Z_{\mathrm{IB}}^{(l)}; G)^2\right\}. \tag{25}$$

由于公式 (23) 中的 $H(Y)$ 是一个与优化无关的常数, 因此在目标函数中舍弃掉了. 在实际训练中, 公式 (25) 中的第一项对于图分类任务来说等价于常用的交叉熵 (cross entropy, CE) 损失函数, 对于图回归任务来说等价于常用的均方误差 (mean square error, MSE) 损失函数. 公式 (25) 中的第二项相当于在优化过程中减少 G_{IB} 对 G 的依赖、增大信息压缩度的正则项. 拉格朗日乘子 β 作为压缩系数调节信息压缩程度, 在 5.5 节分析了其参数敏感度. NIB-HGSL方法的算法伪代码如算法 1 所示.

5 性能评测

本节对 NIB-HGSL方法进行性能评测. 首先介绍实验数据集、基线方法及实验设置, 进一步在图分类任务、图回归任务和图去噪任务上进行验证和分析.

5.1 实验设置

数据集. 对于图分类任务来说, 本文参考图结构学习基准 GSLB [49] 的实验设置, 在四个广泛使用的社交网络数据集 (IMDB-B, IMDB-M, REDDIT-B 和 COLLAB [50]) 和两个生物数据集 (MUTAG [51] 和 PROTEINS [52]) 上进行了验证实验. IMDB-B 和 IMDB-M 数据集中每个图样本代表演员及其在电影中的共现关系关系, 图标签代表电影类型. REDDIT-B 数据集中每个图样本中节点

https://www.sciengine.com/doi/10.1360/SSI-2024-0049

算法 1 NIB-HGSL的训练过程

输入: 图 $G=(\mathbf{X},\mathbf{A})$, 其节点集为 V, 图标签 Y, 超参数 c, L, λ, β, 池化率 α, 训练轮次 Epochs

初始化: $\mathbf{A}_{\mathrm{IB}}^{0}=\mathbf{A},\ \mathbf{X}_{\mathrm{IB}}^{0}=\mathbf{X},\ \mathbf{M}=\mathbf{0},\ n_{(0)}=|V|,\ \mathcal{L}_{\mathrm{SNIB}}=0.$

for $e=1,2,\ldots,$ Epochs **do**

  for $l=1,\ldots,L$ **do**

    $\phi_{i,j}^{(l-1)}=\frac{(x_i^{(l-1)})^T x_j^{(l-1)}}{\sum_{k\neq i}(x_i^{(l-1)})^T x_k^{(l-1)}}(\mathbf{A}_{\mathrm{IB}}^{(l-1)})_{i,j}^{c},\ i\neq j$ 且 $0\leqslant i,j<n_{(l)}$. ▷ 计算注意力分数矩阵

    $\mathbf{U}^{(l-1)}=\mathrm{GNN}(X_{\mathrm{IB}}^{(l-1)},A_{\mathrm{IB}}^{(l-1)})\in\mathbb{R}^{N_{l-1}\times N_l},\ N_l=\alpha N_{l-1}.$

    $\mathbf{S}^{(l-1)}=\mathrm{softmax}(\phi^{(l-1)}\mathbf{U}^{(l-1)})$. ▷ 得到图池化算子

    $\mathbf{X}^{(l)}=(\mathbf{S}^{(l-1)})^{\mathrm{T}}\mathbf{X}_{\mathrm{IB}}^{(l-1)},\ \mathbf{A}^{(l+1)}=(\mathbf{S}^{(l-1)})^{\mathrm{T}}\mathbf{A}_{\mathrm{IB}}^{(l-1)}\mathbf{S}^{(l-1)}.$

    $\mathbf{M}=\mathrm{sigmoid}(\mathbf{M})$.

    $\mathbf{X}_{\mathrm{SL}}^{(l)}=\mathbf{X}_r^{(l)}+(\mathbf{X}^{(l)}-\mathbf{X}_r^{(l)})\odot\mathbf{M}$. ▷ 优化第 l 层图特征

    $\mathbf{h}_i^{(l)}=\mathrm{NN}\left(\mathbf{X}_{\mathrm{SL}}^{(l)}(u)\right)$.

    $\pi_{i,j}^{(l)}=\frac{1}{m}\cos\left(\mathbf{W}_h\cdot\left(\mathbf{x}_i^{(l)}|\mathbf{h}_i^{(l)}\right),\mathbf{W}_h\cdot\left(\mathbf{x}_j^{(l)}|\mathbf{h}_j^{(l)}\right)\right)$.

    $\mathbf{A}_{\mathrm{SL}}^{(l)}=\bigcup_{i,j\in V^l}\left\{a_{i,j}^{(l)}\sim\mathrm{Ber}\left(\pi_{i,j}\right)\right\}$.

    $\mathbf{A}_{\mathrm{IB}}^{(l)}=\lambda\mathbf{A}_{\mathrm{SL}}^{(l)}+(1-\lambda)\mathbf{A}^{(l)},\ \mathbf{X}_{\mathrm{IB}}^{(l)}=\mathbf{X}^{(l)}$. ▷ 优化第 l 层图结构

    $Z_{\mathrm{IB}}^{(l)}=\mathrm{GNN}(\mathbf{X}_{\mathrm{IB}}^{(l)},\mathbf{A}_{\mathrm{IB}}^{(l)})$. ▷ 学习第 l 层的图表征

  end for

  $\mathcal{L}_{\mathrm{SNIB}}=\sum_{l=0}^{L}\left\{\mathbb{E}_{p(Y,Z_{\mathrm{IB}}^{(l)})}\left[\log q_\theta\left(Y|G_{\mathrm{IB}}^{(l)}\right)\right]-\beta\hat{\mathcal{I}}(Z_{\mathrm{IB}}^{(l)};G)^2\right\}$.

end for

代表帖子, 边代表两个帖子被同一人评论, 图标签是帖子对应的社区分类. COLLAB 数据集是作者协作网络的一个子集, 节点代表作者, 边表示作者之间的协作, 节点特征是通过平均作者发表的论文的词的表征, 边的特征包括年份和该年发表的合著论文数量. MUTAG 数据集由分子图组成, 根据化合物对细菌是否有诱变作用来对图进行标记. PROTEINS 数据集由蛋白质图组成, 节点表示蛋白质的二级结构, 边表示两个二级结构在氨基酸序列中或者三维结构中是邻居, 图标签表示该蛋白质为酶或者非酶. 对于图回归任务来说, 本文使用了两个聚合物数据集 (plym-melting, plym-galss) 和三个分子数据集 (ogbg-mollipo, ogbg-molesol, ogbg-molfreesolv) [53]. 其中, plym-melting 数据集为每个聚合物样本预测其熔融温度, plym-glass 数据集中为每个聚合物样本预测其玻璃转化温度, ogbg-mollipo 数据集中为每个分子样本预测其亲脂性, ogbg-molesol 数据集中为每个分子样本预测其水溶性 (每升摩尔的对数溶解度), ogbg-molfreesolv 数据集中为每个分子样本预测其水中分子的水合自由能.

基线方法. 本文将提出的 NIB-HGSL与两个专门面向图层级的结构学习方法 (HGP-SL [27] 和 VIB-GSL [23]) 以及其他一些可以扩展到图层级的结构学习基线方法 (NeuralSparse [11], SIB [35], IDGL [9]) 进行比较. (1) NeuralSparse [11]: 通过去除图中多余的与任务无关的边, 从而更清晰地划分类别边界, 提高模型泛化能力. (2) SIB [35]: 通过子图信息瓶颈原则识别图数据中具有预测性的关键子图, 结合互信息估计器设计了一种双层优化方案. (3) IDGL [9]: 将图结构学习看作节点特征的度量学习, 通过交替迭代与优化同时学习图结构与节点表征. (4) HGP-SL [27]: 首先通过节点选择进行池化, 再通过结构学习捕捉节点间关联. (5) VIB-GSL [23]: 使用变分信息瓶颈指导, 通过节点间的全局相似性学习一个单一尺度图结构.

实验设置. 本文将常见的图神经网络作为基础编码器嵌入 NIB-GSL 框架中 (包括经典的图卷积

https://www.sciengine.com/doi/10.1360/SSI-2024-0049

神经网络 GCN [54]、图注意力网络 GAT [55]、图同构网络 GIN [14], 以及近年提出的各向同性高效图卷积网络 EGC [56]), 以观察 NIB-HGSL 是否能提高这些基础编码器在图分类任务上的性能. 为了进行公平的比较, 全部方法表征维度为 16, 使用同样的双层感知机作为分类器. 对于层次结构学习中的池化聚类计算, 实际训练中由于相近轮次节点嵌入更新差别较小, 因此为了提高训练效率间隔固定轮次进行更新. NIB-HGSL 设置温度参数 t 为 0.1, $c=1$, 并对 α, β, λ 进行超参数搜索. 所有模型的参数都经过仔细的调整优化, 并在单个 NVIDIA V100 32GB GPU 上进行训练和测试.

5.2 图分类实验结果与分析

首先验证 NIB-HGSL 改善基础图神经网络图分类的能力, 评价指标为图分类的准确率 (accuracy), 分类准确率越高代表方法效果越好. 实验中进行了十折交叉验证, 所有方法的平均准确率和标准差如表 1, 每个方法组的最佳结果用粗体表示, 效果次优的用下划线表示.

从图分类的结果可以看出, 所提出的 NIB-HGSL 方法在所有数据集上的图分类准确率都超过了基线方法. NIB-HGSL 相对于基础图神经网络的性能优势意味着它能够学习与任务更密切相关的图结构, 进而提高图神经网络学到的图表征质量. 一般来说, 图稀疏化方法 (即 NeuralSparse [11] 和 SIB [35]) 在提升基础图神经网络准确性方面只有很小的改善, 甚至对性能有负面影响 (例如在 COLLAB 上), 这是因为它们受制于观察到的结构, 而没有挖掘节点间潜在的关联关系. 相当于全局结构学习的方法 (IDGL [9] 和 VIB-GSL [23]), 我们的方法 NIB-HGSL由于能够重新学习图的层次结构关系, 从而保留不同层级的图性质, 因此更适合于图层级的任务. 而相对于层次结构池化的方法 HGP-SL [27], 我们的方法 NIB-HGSL 由于引入了信息瓶颈理论去指导图结构优化, 可以更好地发现在每个层次上关键的结构信息, 从而自适应地保留相应的结构性质, 因此取得了更好的表现.

5.3 图回归实验结果与分析

其次验证 NIB-HGSL 在图回归任务上的效果, 评价指标为均方绝对误差 (mean absolute error, MAE), 均方绝对误差越低代表方法效果越好. 所有方法的平均 MAE 和标准差如表 2, 每个方法组的最佳结果用粗体表示, 效果次优的用下划线表示. 从图回归的结果可以看出, 所提出的 NIB-HGSL 方法在所有数据集上的回归均方绝对误差都小于基线方法. 例如, 在 plym-melting 数据集上, NIB-HGSL 相对于次优的基线方法, 在不同基础图神经网络编码器的配置下图回归均方绝对误差分别降低了 5.25, 3.06, 2.47, 3.77, 说明了本文所提方法具有更强的提取图关键信息的能力.

5.4 鲁棒性分析

NIB-HGSL 在信息瓶颈原则指导下进行结构优化, 旨在压缩噪声信息、保留有效信息. 本节分别针对两类典型的噪声场景: 随机噪声和对抗攻击噪声, 开展对 NIB-HGSL 鲁棒性的分析验证.

随机噪声鲁棒性分析. 为评估 NIB-HGSL 在噪声图数据上的鲁棒性, 本节通过扰动 REDDIT-B 数据集中图样本的边来生成一个带噪声的合成数据集, 并用其进行图去噪任务验证. 具体来说, 对数据集中的每个样本, 随机地翻转 25%, 50%, 75% 的边 (如果不存在边则增加, 如果存在边则删除). 重复五次实验, 在 IMDB-B、REDDIT-B 和 MUTAG 数据集上的结果如图 2 所示. 可以看出, GCN 方法的分类准确率随着边翻转比例的增大呈现出剧烈下降, 这表明图神经网络对结构噪声较为敏感. 图结构学习基线方法 HGP-SL 和 VIB-GSL 相对于 GCN 来说准确率下降趋势较缓, 这说明了结构学习对于增强方法鲁棒性的作用. 本文方法 NIB-HGSL同时受益于信息瓶颈理论的指导和层次化的结构学习, 可以去除结构中的噪声信息, 因此相对于 HGP-SL 和 VIB-GSL 来说具有更好的鲁棒性.

https://www.sciengine.com/doi/10.1360/SSI-2024-0049

表 1　图分类任务上方法准确率和标准差

Table 1　Accuracy (± standard deviation, %) of the graph classification task

Backbone	Method	Dataset					
		IMDB-B	**IMDB-M**	**REDDIT-B**	**COLLAB**	**MUTAG**	**PROTEINS**
GCN	Vanilla	62.10 ± 9.04	46.00 ± 4.33	79.05 ± 5.76	68.86 ± 6.18	70.27 ± 6.35	71.45 ± 6.35
	NeuralSparse	70.60 ± 3.50	45.73 ± 3.97	83.15 ± 2.70	69.64 ± 2.14	74.27 ± 7.20	73.67 ± 3.40
	SIB	67.50 ± 3.54	46.27 ± 3.05	83.35 ± 4.82	69.04 ± 1.45	76.16 ± 8.84	74.71 ± 3.24
	IDGL	68.80 ± 7.61	46.15 ± 3.45	84.65 ± 4.31	68.18 ± 1.90	75.24 ± 6.86	73.06 ± 3.88
	HGP-SL	69.80 ± 3.58	46.28 ± 4.90	82.10 ± 3.57	68.20 ± 1.52	73.84 ± 6.61	72.99 ± 3.44
	VIB-GSL	70.80 ± 4.49	46.00 ± 3.80	83.20 ± 2.57	67.52 ± 1.22	75.41 ± 6.64	73.41 ± 4.31
	NIB-HGSL	**70.90 ± 4.41**	**46.66 ± 3.78**	**85.80 ± 1.53**	**71.60 ± 1.12**	**77.12 ± 7.42**	**75.31 ± 3.15**
GAT	Vanilla	68.40 ± 5.56	44.80 ± 5.36	76.85 ± 4.83	70.60 ± 2.08	72.34 ± 6.43	73.06 ± 3.99
	NeuralSparse	68.60 ± 4.33	47.48 ± 4.84	78.41 ± 5.35	71.60 ± 1.40	81.92 ± 7.72	73.06 ± 2.18
	SIB	68.60 ± 6.10	47.26 ± 2.79	78.05 ± 6.07	68.83 ± 1.19	83.51 ± 5.86	73.69 ± 2.48
	IDGL	67.50 ± 5.38	46.62 ± 3.58	78.30 ± 3.85	69.32 ± 6.50	82.86 ± 4.93	74.86 ± 3.83
	HGP-SL	68.05 ± 3.40	45.73 ± 4.11	76.25 ± 2.73	69.22 ± 2.51	83.23 ± 6.33	73.28 ± 2.73
	VIB-GSL	68.60 ± 5.80	46.25 ± 3.55	76.75 ± 2.73	68.04 ± 1.52	81.52 ± 4.51	73.28 ± 3.16
	NIB-HGSL	**69.00 ± 4.50**	**47.53 ± 3.43**	**78.95 ± 6.91**	**71.70 ± 1.79**	**84.68 ± 9.42**	**76.21 ± 4.51**
GIN	Vanilla	68.00 ± 4.94	42.99 ± 7.32	83.40 ± 3.27	69.84 ± 6.45	80.66 ± 8.53	72.75 ± 3.53
	NeuralSparse	68.10 ± 6.95	44.53 ± 4.12	86.70 ± 4.18	68.61 ± 2.73	84.55 ± 3.07	72.32 ± 4.31
	SIB	68.90 ± 3.54	45.81 ± 2.85	86.55 ± 3.72	67.32 ± 2.17	85.09 ± 4.99	72.80 ± 2.72
	IDGL	68.00 ± 3.27	45.74 ± 5.62	86.20 ± 3.02	70.06 ± 4.97	85.52 ± 4.83	72.79 ± 5.63
	HGP-SL	68.80 ± 3.22	45.52 ± 4.41	86.35 ± 1.93	69.10 ± 1.54	84.49 ± 5.69	70.64 ± 4.60
	VIB-GSL	67.70 ± 5.06	45.32 ± 3.45	85.80 ± 2.25	68.58 ± 2.41	84.58 ± 6.37	72.97 ± 2.83
	NIB-HGSL	**69.70 ± 3.33**	**46.29 ± 3.68**	**88.35 ± 2.49**	**70.56 ± 1.68**	**87.81 ± 8.57**	**73.25 ± 2.30**
EGC	Vanilla	68.30 ± 3.02	46.20 ± 4.15	80.75 ± 6.06	69.48 ± 1.84	77.70 ± 6.76	72.52 ± 4.21
	NeuralSparse	66.70 ± 4.16	46.61 ± 4.45	82.40 ± 5.88	68.24 ± 2.52	79.07 ± 8.68	74.32 ± 2.40
	SIB	71.00 ± 4.27	46.74 ± 3.50	81.25 ± 4.14	69.72 ± 1.82	81.42 ± 5.56	74.72 ± 2.78
	IDGL	71.60 ± 4.30	47.56 ± 5.22	83.85 ± 4.55	69.96 ± 1.46	80.83 ± 7.80	73.24 ± 4.11
	HGP-SL	69.10 ± 4.43	46.39 ± 3.06	83.50 ± 3.52	68.12 ± 1.94	81.44 ± 7.02	71.90 ± 3.72
	VIB-GSL	71.10 ± 3.35	46.81 ± 2.97	82.45 ± 4.53	68.62 ± 1.11	77.64 ± 7.46	73.51 ± 3.00
	NIB-HGSL	**72.20 ± 3.77**	**48.00 ± 4.07**	**87.90 ± 2.88**	**70.44 ± 1.96**	**83.56 ± 7.93**	**75.49 ± 3.66**

对抗攻击鲁棒性分析.　为评估 NIB-HGSL 面对图对抗攻击的鲁棒性, 本节采用一种典型的图层级攻击方法 PRBCD [57] 对每个图样本进行逃逸攻击 (evasion attack). 具体来说, PRBCD 方法通过投影随机块下降寻找对性能影响最大的边进行翻转, 从而实现攻击, 本文采用 2.5%, 5.0%, 7.5% 的攻击预算来模仿不同强度的对抗攻击. 在不同攻击强度下重复五次实验, 在 IMDB-B, REDDIT-B 和 MUTAG 三个数据集上的结果如图 3 所示. 从图中可以看出, 各方法随着攻击力度的增大准确率都有一定程度的下降, 其中 HGP-SL 的下降程度最为剧烈, 对对抗攻击噪声极为敏感. 而另一个基于信息瓶颈的结构学习方法 VIB-GSL 的准确率下降趋势稍缓, 表现出更好的鲁棒性, 这说明了信息瓶颈原则可以有效提高图神经网络对于对抗攻击的防御能力. 而本文方法 NIB-HGSL同时受益于信息瓶颈原则对噪声信

https://www.sciengine.com/doi/10.1360/SSI-2024-0049

表 2 图回归任务上方法均方绝对误差和标准差

Table 2 MAE (± standard deviation, %) of the graph regression task

Backbone	Method	Dataset				
		plym-melting	**plym-glass**	**ogbg-mollipo**	**ogbg-molesol**	**ogbg-molfreesolv**
GCN	Vanilla	65.67 ± 9.88	49.43 ± 5.49	0.86 ± 0.09	1.03 ± 0.06	2.26 ± 0.20
	NeuralSparse	62.17 ± 6.16	43.72 ± 2.68	0.87 ± 0.06	0.91 ± 0.06	1.94 ± 0.43
	SIB	64.58 ± 4.22	41.81 ± 2.34	0.84 ± 0.06	0.75 ± 0.02	1.90 ± 0.28
	IDGL	64.71 ± 2.82	48.82 ± 0.87	0.91 ± 0.05	0.93 ± 0.29	1.97 ± 0.20
	HGP-SL	66.31 ± 5.59	48.67 ± 3.89	0.91 ± 0.03	0.98 ± 0.23	2.14 ± 0.37
	VIB-GSL	60.64 ± 4.68	49.58 ± 1.10	0.93 ± 0.03	0.99 ± 0.12	2.04 ± 0.35
	NIB-HGSL	**55.39 ± 1.96**	**37.33 ± 2.25**	**0.80 ± 0.03**	**0.68 ± 0.06**	**1.65 ± 0.34**
GAT	Vanilla	73.96 ± 9.88	50.44 ± 3.54	0.83 ± 0.07	1.00 ± 0.36	2.17 ± 0.28
	NeuralSparse	69.08 ± 8.84	46.53 ± 3.71	0.83 ± 0.13	0.93 ± 0.12	1.91 ± 0.20
	SIB	66.92 ± 6.35	43.38 ± 3.31	0.84 ± 0.05	0.74 ± 0.06	1.94 ± 0.40
	IDGL	70.22 ± 1.90	48.25 ± 3.51	0.83 ± 0.04	0.83 ± 0.10	1.88 ± 0.27
	HGP-SL	73.22 ± 1.85	50.83 ± 3.65	0.92 ± 0.05	0.91 ± 0.21	2.06 ± 0.30
	VIB-GSL	66.34 ± 5.05	49.05 ± 2.70	0.91 ± 0.07	0.84 ± 0.16	2.09 ± 0.25
	NIB-HGSL	**63.28 ± 4.61**	**39.73 ± 1.90**	**0.80 ± 0.04**	**0.71 ± 0.08**	**1.81 ± 0.21**
GIN	Vanilla	66.61 ± 7.60	55.63 ± 6.00	0.95 ± 0.04	1.12 ± 0.32	2.74 ± 0.37
	NeuralSparse	65.96 ± 7.91	53.21 ± 6.14	0.97 ± 0.04	1.06 ± 0.27	2.13 ± 0.68
	SIB	61.48 ± 4.94	45.15 ± 2.33	0.94 ± 0.06	1.05 ± 0.13	2.63 ± 0.20
	IDGL	64.07 ± 4.60	45.52 ± 2.92	0.89 ± 0.06	0.95 ± 0.18	2.33 ± 0.58
	HGP-SL	65.72 ± 3.14	53.47 ± 5.19	0.95 ± 0.03	1.09 ± 0.12	2.68 ± 0.27
	VIB-GSL	64.47 ± 4.26	53.57 ± 4.13	0.99 ± 0.09	0.97 ± 0.13	2.53 ± 0.67
	NIB-HGSL	**59.01 ± 3.94**	**36.10 ± 2.19**	**0.84 ± 0.05**	**0.86 ± 0.06**	**1.99 ± 0.44**
EGC	Vanilla	69.81 ± 3.70	45.93 ± 2.63	0.75 ± 0.06	0.76 ± 0.04	2.12 ± 0.28
	NeuralSparse	69.28 ± 3.11	43.29 ± 2.53	0.79 ± 0.11	0.84 ± 0.12	1.81 ± 0.85
	SIB	65.38 ± 8.11	40.15 ± 1.34	0.74 ± 0.05	0.72 ± 0.09	1.81 ± 0.39
	IDGL	65.38 ± 8.11	41.70 ± 3.85	0.79 ± 0.03	0.83 ± 0.14	1.78 ± 0.16
	HGP-SL	68.05 ± 6.02	45.95 ± 5.97	0.87 ± 0.04	0.80 ± 0.17	2.35 ± 0.43
	VIB-GSL	65.00 ± 6.09	45.68 ± 3.82	0.89 ± 0.04	0.85 ± 0.12	2.11 ± 0.62
	NIB-HGSL	**61.23 ± 3.08**	**38.15 ± 0.71**	**0.72 ± 0.06**	**0.68 ± 0.12**	**1.73 ± 0.38**

息的压缩, 以及层次结构学习中对于子结构语义的增强, 展现出最好的鲁棒性.

5.5 NIB-HGSL分析

为验证本文方法 NIB-HGSL 中信息瓶颈原则与层次结构学习的有效性, 本节首先对其进行消融分析实验. 此外, 本文提出方法 NIB-HGSL 通过同时对有效信息保留和噪声信息压缩来帮助获得优化的图结构, 其中信息压缩程度决定了方法的效果. 因此, 本节对拉格朗日 (Lagrange) 乘子 β 的参数敏感性分析和对压缩过程中的信息平面进行进一步分析.

https://www.sciengine.com/doi/10.1360/SSI-2024-0049

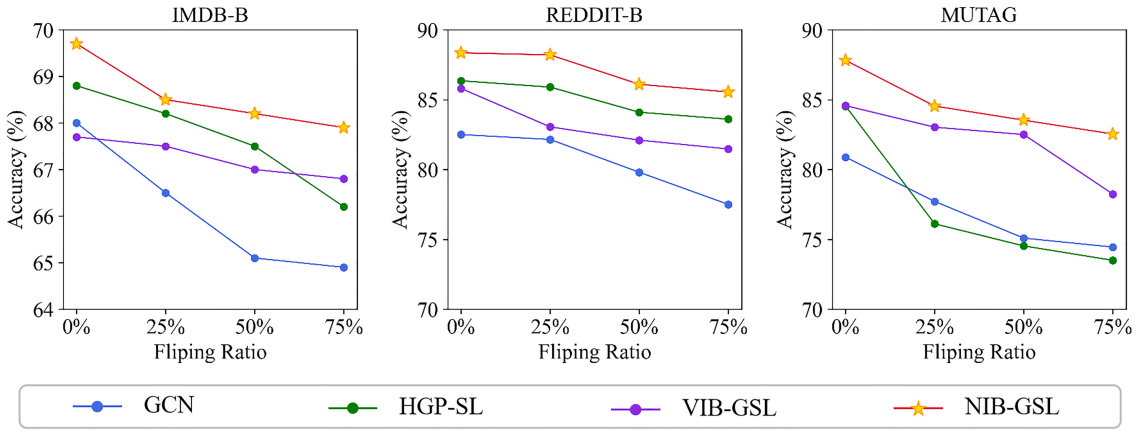


图 2　(网络版彩图) NIB-HGSL 在随机噪声图上的鲁棒性分析

Figure 2　(Color online) Robustness analysis of NIB-HGSL on graphs with random noise

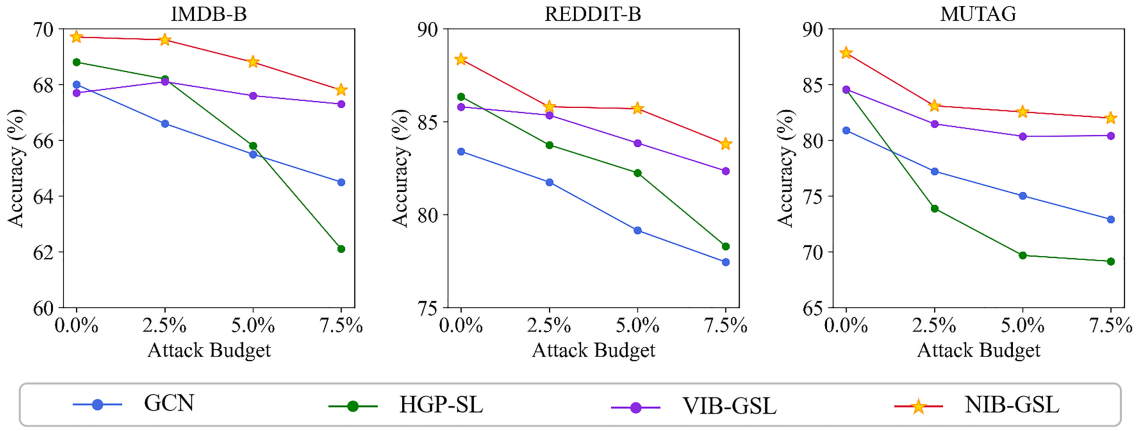


图 3　(网络版彩图) NIB-HGSL 在对抗攻击图上的鲁棒性分析

Figure 3　(Color online) Robustness analysis of NIB-HGSL on graphs with adversarial attacks

消融实验.　为验证信息瓶颈原则和层次结构学习的有效性, 分别构造了两个 NIB-HGSL的变种方法 NIB-HGSL(w/o IB) 和 NIB-HGSL(w/o HGSL). 其中 NIB-HGSL(w/o IB) 为去掉信息瓶颈的指导, 即去掉优化函数中的信息压缩项, 仅通过分类损失函数指导整体框架的优化; NIB-HGSL(w/o HGSL) 为去掉层次结构学习, 即为每个图样本学习一个全局的图结构. 消融实验的结果如图 4 所示, 可以看出在五个数据集上, 本文提出的信息瓶颈原则的指导以及层次化的图结构学习方法都能带来一致性的效果提升, 其中非线性信息瓶颈原则的引入在大部分数据集上带来的效果提升更为明显.

拉格朗日乘子 β 参数敏感性.　β 在公式 (2) 和 (25) 中起到对有效信息保留项 $\mathcal{I}(G_{\mathrm{IB}};Y)$ 和噪声信息压缩项 $\mathcal{I}(G_{\mathrm{IB}};G)$ 之间进行权衡的作用. 值得注意的是, 在信息瓶颈原则中, 增加和减少信息瓶颈维度 K 之间存在着关联关系 [18]. 随着 β 变化, NIB-HGSL 在 REDDIT-B 数据集上图分类准确性的变化趋势如图 5 所示. 可以观察到, NIB-HGSL 的图分类准确率随着 β 的增加都呈现驼峰型. 在其他数据集的 β 敏感性分析实验中也呈现出同样趋势, 在此没有一一展示. 分类准确率首先随着 β 的增大而增加, 表明去除与任务不相关的噪声信息确实可以有效增强图表征的质量; 然后, 准确率逐渐

https://www.sciengine.com/doi/10.1360/SSI-2024-0049

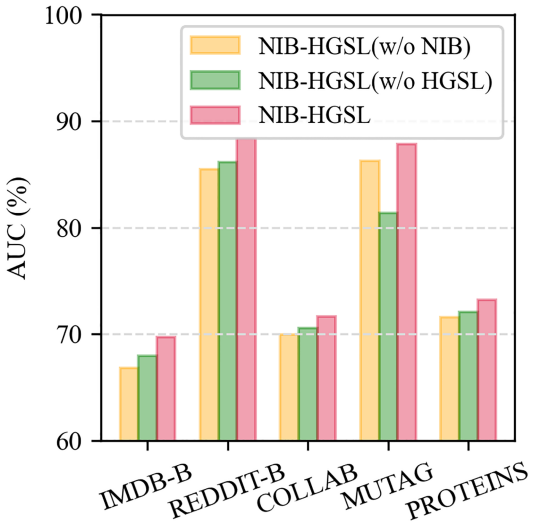


图 4 (网络版彩图) 消融实验分析

Figure 4 (Color online) Ablation study

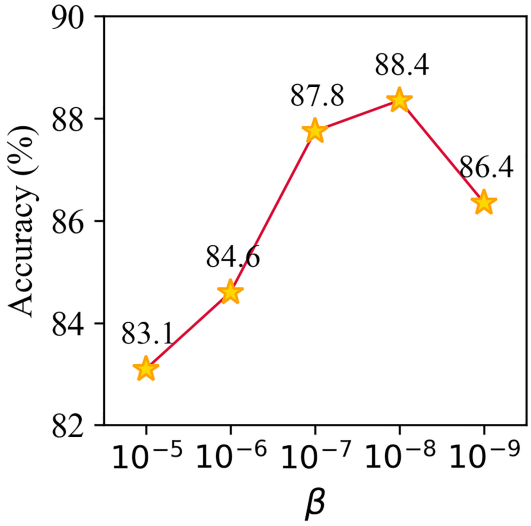


图 5 (网络版彩图) β 参数敏感性

Figure 5 (Color online) Sensitivity of β

下降并达到较低的值, 这表明过度的信息压缩会损失部分有效信息. 不同的数据集由不同比例的任务无关信息组成, 因此需要不同程度的信息压缩, 因此适当的 β 值可以提高图分类模型的性能.

可视化分析. 我们通过信息平面 [34] 来分析 NIB-HGSL在学习过程中的表现, 该方法被广泛应用于分析训练过程中输入、潜在表征和输出之间互信息的变化. NIB-HGSL在 REDDIT-B 数据集上训练的信息平面如图 6 所示, 图中 X 轴代表压缩项 $\mathcal{I}(G_{\mathrm{IB}};G)$, Y 轴代表预测项 $\mathcal{I}(G_{\mathrm{IB}};Y)$, 散点代表模型在训练过程中每个轮次所处的信息压缩状态. 可以看出, 训练过程中互信息变化分为两个阶段. 在第一个阶段, $\mathcal{I}(G_{\mathrm{IB}};G)$ 和 $\mathcal{I}(G_{\mathrm{IB}};Y)$ 都增加, 这表明学习到的表征正在提取关于输入数据和标签的互信息, 并且坐标向右上角移动. 在第二阶段, $\mathcal{I}(G_{\mathrm{IB}};G)$ 开始下降, 而 $\mathcal{I}(G_{\mathrm{IB}};Y)$ 的增长速度放缓, 并收敛到左上角, 这表明我们的 NIB-HGSL开始生效, 导致表征向量中开始压缩无关信息, 仅保留与下游任务相关的最小充分信息.

模型效率分析. 首先, 对本文所提出方法 NIB-HGSL的复杂度进行分析. NIB-HGSL为一个层次化结构学习方法, 过图池化技术递归地聚合节点集以形成超节点, 生成不同结构粒度的层次图, 再通过结构学习对多层次的图结构进行优化. 此处设一共有 L 层, 对于每层分别保留 $\{N_0, N_1, \cdots, N_{L-1}\}$ 个超节点, 其中 $N_0 = |V|$, 则对于第 l 层的结构学习来说, 复杂度为 $\mathcal{O}(N_l^2)$. 因此, 整体方法的复杂度为 $\mathcal{O}(|V|^2)$, 和其他结构学习的基线方法 (NeuralSparse, SIB, IDGL, HGP-SL, VIB-GSL) 相同. 其次, 图 7 对 NIB-HGSL和其他基线方法的效率进行了对比, 其中, 横轴为平均每个轮次的训练时间 (s), 纵轴为图分类任务的准确率 (%). 从图 7 可以看出, NIB-HGSL在准确率达到最高的同时, 其训练效率优于等 IDGL、HSP-SL、SIB 最新基线方法. 虽然 NIB-HGSL的训练效率不如 GCN、VIB-GSL、NeuralSparse, 但其准确率能够具有较大优势.

6 总结

本文工作针对图数据中存在的结构噪声问题, 从信息论的视角出发, 提出非线性信息瓶颈指导的图结构学习方法 NIB-HGSL. 该方法通过变分近似将信息瓶颈原则扩展到图数据的层次结构学习上,

https://www.sciengine.com/doi/10.1360/SSI-2024-0049

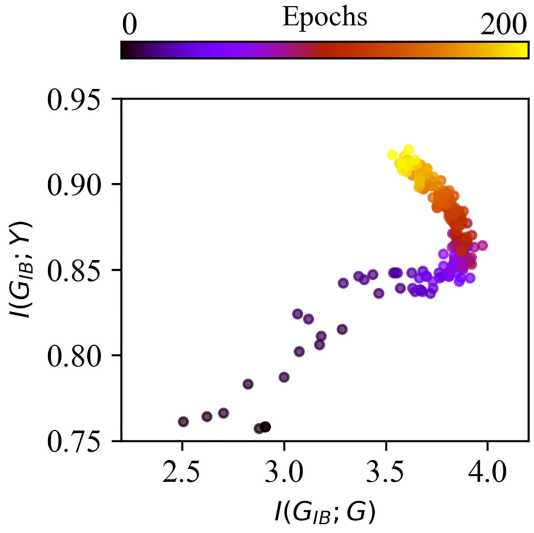


图 6 (网络版彩图) 信息平面

Figure 6 (Color online) Information plane

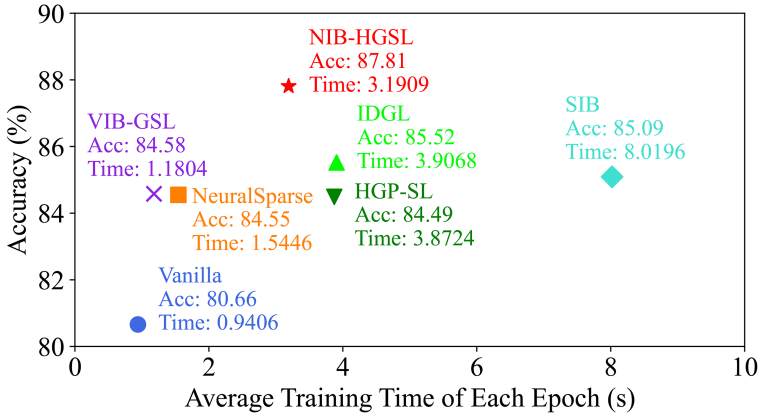


图 7 (网络版彩图) NIB-HGSL 训练效率分析

Figure 7 (Color online) Analysis of training efficiency of NIB-HGSL

通过联合优化层次图结构和图表征来去除结构中的噪声信息, 并增强各层结构中的有效信息表达, 获得更鲁棒的图表征. 在图分类数据集、图回归数据集和结构噪声及对抗攻击数据集上的实验结果表明, NIB-HGSL可以提高图表征学习的准确性和鲁棒性.

参考文献

1 Zhang D, Yin J, Zhu X, et al. Network representation learning: a survey. IEEE Trans Big Data, 2018, 6: 3–28

2 Liu Z Y, Zhang L, Tu C C, et al. Statistical and semantic analysis of rumors in Chinese social media. Sci Sin Inform, 2015, 45: 1536–1546 [刘知远, 张乐, 涂存超, 等. 中文社交媒体谣言统计语义分析. 中国科学: 信息科学, 2015, 45: 1536–1546]

3 Zheng Z M, Lyu J H, Wei W, et al. Refined intelligence theory: artificial intelligence regarding complex dynamic objects. Sci Sin Inform, 2021, 51: 678–690 [郑志明, 吕金虎, 韦卫, 等. 精准智能理论: 面向复杂动态对象的人工智能. 中国科学: 信息科学, 2021, 51: 678–690]

https://www.sciengine.com/doi/10.1360/SSI-2024-0049

4 Yan Z, Xiang X G, Li Z C. Item correlation modeling in interaction sequence for graph convolutional session recommendation. Sci Sin Inform, 2022, 52: 1069–1082 [闫昭, 项欣光, 李泽超. 基于交互序列商品相关性建模的图卷积会话推荐. 中国科学: 信息科学, 2022, 52: 1069–1082]
5 Sun L C, Dou Y T, Yang C, et al. Adversarial attack and defense on graph data: A survey. IEEE T Knowl Data En, 2018, 35: 7693–7711
6 Zügner D, Akbarnejad A, Günnemann S. Adversarial attacks on neural networks for graph data. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, London: ACM, 2018. 2847–2856
7 Zhou Z Y, Zhou S, Mao B C, et al. Opengsl: A comprehensive benchmark for graph structure learning. In: Advances in Neural Information Processing Systems 36, New Orleans, 2023
8 Zhu Y Q, Xu W Z, Zhang J H, et al. Deep graph structure learning for robust representations: A survey. 2021. ArXiv:2103.03036
9 Chen Y, Wu L F, Zaki M. Iterative deep graph learning for graph neural networks: Better and robust node embeddings. In: Advances in Neural Information Processing Systems 33, 2020. 19314–19326
10 Jin W, Ma Y, Liu X R, et al. Graph structure learning for robust graph neural networks. In: Proceedings of the 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, California: ACM, 2020. 66–74
11 Zheng C, Zong B, Cheng W, et al. Robust graph representation learning via neural sparsification. In: Proceedings of the 37th International Conference on Machine Learning, Vienna: PMLR, 2020. 11458–11468
12 Zhang G, Wu J, Yang J, et al. Fraudre: Fraud detection dual-resistant to graph inconsistency and imbalance. In: Proceedings of IEEE International Conference on Data Mining, Auckland: IEEE, 2021. 867–876
13 Sun M Z, Zhang X, Zheng J Q, et al. Ddgcn: Dual dynamic graph convolutional networks for rumor detection on social media. In: Proceedings of the 36th AAAI Conference on Artificial Intelligence, Palo Alto: AAAI Press, 2022. 4611–4619
14 Xu K Y L, Hu W H, Leskovec J, et al. How powerful are graph neural networks? In: Proceedings of the 7th International Conference on Learning Representations, New Orleans, 2019
15 Liu Z, Zhou P, Li Z, et al. Think like a graph: real-time traffic estimation at city-scale. IEEE Trans Mobile Comput, 2018, 18: 2446–2459
16 Ying R, You J X, Morris C, et al. Hierarchical graph representation learning with differentiable pooling. In: Advances in Neural Information Processing Systems 31, Montreal, 2018. 4805–4815
17 Tishby N, Pereira F C, Bialek W. The information bottleneck method. 2000. ArXiv:0004057
18 Shamir O, Sabato S, Tishby N. Learning and generalization with the information bottleneck. Theor Comput Sci, 2010, 411: 2696–2711
19 Sinha S, Bharadhwaj H, Goyal A, et al. Diversity inducing information bottleneck in model ensembles. In: Proceedings of the 35th AAAI Conference on Artificial Intelligence, Palo Alto: AAAI Press, 2020. 9666–9674
20 Kolchinsky A, Tracey B D, Wolpert D H. Nonlinear Information Bottleneck. Entropy, 2019, 21: 1181
21 Luo D S, Cheng W, Yu W C, et al. Learning to drop: Robust graph neural network via topological denoising. In: Proceedings of the 14 ACM International Conference on Web Search and Data Mining, Israel: ACM, 2021. 779–787
22 Franceschi L, Niepert M, Pontil M, et al. Learning discrete structures for graph neural networks. In: Proceedings of the 36th International Conference on Machine Learning, California: PMLR, 2019. 1972–1982
23 Sun Q Y, Li J X, Peng H, et al. Graph structure learning with variational information bottleneck. In: Proceedings of the 36th AAAI Conference on Artificial Intelligence, Palo Alto: AAAI Press, 2022. 4165–4174
24 Yang L, Kang Z S, Cao X C, et al. Topology optimization based graph convolutional network. In: Proceedings of the 28th International Joint Conference on Artificial Intelligence, Macao, 2019. 4054–4061
25 Yu J C, Xu T Y, Rong Y, et al. Graph information bottleneck for subgraph recognition. In: Proceedings of 9th International Conference on Learning Representations, Austria, 2021
26 Wang R J, Mou S, Wang X, et al. Graph structure estimation neural networks. In: Proceedings of the Web Conference 2021, Ljubljana: ACM, 2021. 342–353
27 Zhang Z, Bu J J, Ester M, et al. Hierarchical graph pooling with structure learning. IEEE T Knowl Data En, 2023, 35: 545–559
28 Shannon C E. A mathematical theory of communication. Bell Syst Technical J, 1948, 27: 379–423

https://www.sciengine.com/doi/10.1360/SSI-2024-0049

29 Petz D. Entropy, von neumann and the von neumann entropy: Dedicated to the memory of alfred wehrl. John von Neumann and the Foundations of Quantum Physics. Netherlands: Springer, 2001. 83–96
30 Li A, Pan Y. Structural information and fynamical complexity of networks. IEEE Trans Inform Theor, 2016, 62: 3290–3339
31 Belghazi M I, Baratin A, Sai Rajeshwar, et al. Mutual information neural estimation. In: Proceedings of the 35th International Conference on Machine Learning, Stockholm: PMLR, 2018. 531–540
32 Poole B, Ozair S, Oord A V D, et al. On variational bounds of mutual information. In: Proceedings of the 36th International Conference on Machine Learning, Long Beach: PMLR, 2019. 5171–5180
33 Wu T L, Ren H Y, Li P, et al. Graph information bottleneck. In: Advances in Neural Information Processing Systems 33, 2020. 20437–20448
34 Shwartz-Ziv R, Tishby N. Opening the black box of deep neural networks via information. 2017. ArXiv:1703.00810
35 Yu J C, Xu T Y, Rong R, et al. Recognizing predictive substructures with subgraph information bottleneck. IEEE T Pattern Anal, 2021, 46: 1650–1663
36 Yang L, Wu F, Zheng Z C, et al. Heterogeneous graph information bottleneck. In: Proceedings of the 30th International Joint Conference on Artificial Intelligence, Montreal, 2021. 1638–1645
37 Cover T M, Thomas J A. Elements of Information Theory. John Wiley & Sons, 2001. 1–527
38 Schlichtkrull M, Kipf T N, Bloem P, et al. Modeling relational data with graph convolutional networks. In: Proceedings of the Semantic Web-15th International Conference, Heraklion: Springer, 2018. 593–607
39 Busbridge D, Sherburn D, Cavallo P, et al. Relational graph attention networks. 2019. ArXiv:1904.05811
40 Achille A, Soatto S. Emergence of invariance and disentanglement in deep representations. J Mach Learn Res, 2018, 19: 1947–1980
41 Kingma D P, Welling M. Auto-encoding variational bayes. In: Proceedings of 2nd International Conference on Learning Representations, Canada, 2016
42 Elinas P, Bonilla E V, Tiao L. Variational inference for graph convolutional networks in the absence of graph data and adversarial settings. In: Advances in Neural Information Processing Systems 33, 2020. 18648–18660
43 Kazi A, Cosmo L, Ahmadi S A, et al. Differentiable graph module (DGM) for graph convolutional networks. IEEE Trans Pattern Anal Mach Intell, 2022, 45: 1606–1617
44 Jang E, Gu S X, Poole B. Categorical reparameterization with gumbel-softmax. In: Proceedings of 5th International Conference on Learning Representations, 2017
45 Paninski L. Estimation of entropy and mutual information. Neural Computation, 2003, 15: 1191–1253
46 Tian Y L, Sun C, Poole B, et al. What makes for good views for contrastive learning? In: Advances in Neural Information Processing Systems 33, 2020. 6827–6839
47 Yu Y C, Cao J, He R. Improving subgraph recognition with variational graph information bottleneck. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans: IEEE, 2022. 19374–19383
48 Kolchinsky A, Tracey B B, Kuyk S V. Caveats for information bottleneck in deterministic scenarios. In: Proceedings of 7th International Conference on Learning Representations, New Orleans, 2019
49 Li Z X, Wang L, Sun X, et al. Gslb: The graph structure learning benchmark. In: Advances in Neural Information Processing Systems 36, New Orleans, 2023
50 Rossi R, Ahmed N. The network data repository with interactive graph analytics and visualization. In: Proceedings of the 29th AAAI Conference on Artificial Intelligence, Austin: AAAI Press, 2015. 4292–4293
51 Debnath A K, Lopez de Compadre R L, Debnath G, et al. Structure-activity relationship of mutagenic aromatic and heteroaromatic nitro compounds. Correlation with molecular orbital energies and hydrophobicity. J Med Chem, 1991, 34: 786–797
52 Borgwardt K M, Ong C S, Schonauer S, et al. Protein function prediction via graph kernels. Bioinformatics, 2005, 21: i47–i56
53 Liu G, Inae E, Zhao T, et al. Data-centric learning from unlabeled graphs with diffusion model. In: Advances in Neural Information Processing Systems 36, New Orleans, 2023
54 Kipf T K, Welling M. Semi-supervised classification with graph convolutional networks. In: Proceedings of 5th International Conference on Learning Representations, Toulon, 2017
55 Veličković P, Cucurull G, Casanova A, et al. Graph attention networks. In: Proceedings of 6th International Conference

https://www.sciengine.com/doi/10.1360/SSI-2024-0049

on Learning Representations, Vancouver, 2018
56 Tailor S A, Opolka F L, Lio P, et al. Do we need anisotropic graph neural networks? In: Proceedings of the 10th International Conference on Learning Representations, 2022
57 Geisler S, Schmidt T, Şirin H, et al. Robustness of graph neural networks at scale. In: Advances in Neural Information Processing Systems 34, 2021. 7637–7649

Hierarchical graph structure learning with nonlinear information bottleneck

Qingyun SUN[1,2], Jiayi LUO[1,2], Beining YANG[1,2] & Jianxin LI[1,2*]

1. *School of Computer Science and Engineering, Beihang University, Beijing* 100191*, China*;
2. *Advanced Innovation Center for Big Data and Brain Computing, Beijing* 100191*, China*
* Corresponding author. E-mail: lijx@buaa.edu.cn

Abstract In recent years, graph neural networks (GNNs) have become extremely popular due to their powerful expressive capabilities and widespread availability. They have been successful in various real-world applications. GNNs work by learning the node representations through message passing along with the structure. Most empirical studies of GNNs assume that the observed graph structure perfectly depicts the accurate and complete relations between nodes. However, in real-world scenarios, graphs are often noisy, incomplete, or manipulated by adversaries. This noise and interference can affect the quality of graph representations during information aggregation. In this paper, we propose a method called NIB-HGSL, which is a hierarchical graph structure learning method based on the nonlinear information bottleneck principle. NIB-HGSL aims to learn robust graph representations without noisy information, obtaining the minimum sufficient structure for downstream tasks through a balanced optimization of relevant information preservation and noisy information compression. Our comprehensive empirical evaluations demonstrate the effectiveness and robustness of NIB-HGSL, enhancing the power of GNNs for in-the-wild extrapolation.

Keywords graph representation learning, information bottleneck, graph structure learning, graph neural networks, graph classification

https://www.sciengine.com/doi/10.1360/SSI-2024-0049