



小型微型计算机系统

Journal of Chinese Computer Systems

ISSN 1000-1220, CN 21-1106/TP

《小型微型计算机系统》网络首发论文

题目：基于图粗化的层次图池化方法研究
作者：陈洁，薛远远，曹京晶，赵姝，张燕平
收稿日期：2021-07-09
网络首发日期：2021-12-14
引用格式：陈洁，薛远远，曹京晶，赵姝，张燕平. 基于图粗化的层次图池化方法研究[J/OL]. 小型微型计算机系统.
<https://kns.cnki.net/kcms/detail/21.1106.tp.20211213.1525.024.html>



网络首发：在编辑部工作流程中，稿件从录用到出版要经历录用定稿、排版定稿、整期汇编定稿等阶段。录用定稿指内容已经确定，且通过同行评议、主编终审同意刊用的稿件。排版定稿指录用定稿按照期刊特定版式（包括网络呈现版式）排版后的稿件，可暂不确定出版年、卷、期和页码。整期汇编定稿指出版年、卷、期、页码均已确定的印刷或数字出版的整期汇编稿件。录用定稿网络首发稿件内容必须符合《出版管理条例》和《期刊出版管理规定》的有关规定；学术研究成果具有创新性、科学性和先进性，符合编辑部对刊文的录用要求，不存在学术不端行为及其他侵权行为；稿件内容应基本符合国家有关书刊编辑、出版的技术标准，正确使用和统一规范语言文字、符号、数字、外文字母、法定计量单位及地图标注等。为确保录用定稿网络首发的严肃性，录用定稿一经发布，不得修改论文题目、作者、机构名称和学术内容，只可基于编辑规范进行少量文字的修改。

出版确认：纸质期刊编辑部通过与《中国学术期刊（光盘版）》电子杂志社有限公司签约，在《中国学术期刊（网络版）》出版传播平台上创办与纸质期刊内容一致的网络版，以单篇或整期出版形式，在印刷出版之前刊发论文的录用定稿、排版定稿、整期汇编定稿。因为《中国学术期刊（网络版）》是国家新闻出版广电总局批准的网络连续型出版物（ISSN 2096-4188，CN 11-6037/Z），所以签约期刊的网络版上网络首发论文视为正式出版。

基于图粗化的层次图池化方法研究

陈洁^{1,2,3}, 薛远远^{1,2,3}, 曹京晶⁴, 赵姝^{1,2,3,+}, 张燕平^{1,2,3}

¹ (计算智能与信号处理教育部重点实验室, 合肥 230601)

² (安徽大学计算机科学与技术学院, 合肥 230601)

³ (安徽省信息材料与智能传感重点实验室, 合肥 230601)

⁴ (科学技术部科技人才交流开发服务中心, 北京 100045)

+ 通讯作者 E-mail: zhaoshuzs2002@hotmail.com

摘要: 图神经网络 (Graph Neural Networks, GNNs) 已被证明能有效对图结构数据进行建模, 池化机制在使用 GNN 模型提取图层次特征过程中至关重要, 近年来已经引起了越来越多研究者的关注。现有基于聚类的层次图池化方法要么需要增加额外的神经网络层以实现特征图的粗化; 要么不能从全局角度捕获节点在图中的重要性大小。针对以上问题, 本文提出一种基于图粗化的层次图池化方法 (Hierarchical Graph Pooling Based on Graph Coarsening, HGP-GC), 用于学习图的层次特征表示。该方法主要包括图结构粗化和图属性粗化两个部分。利用结构粗化实现特征图尺寸的缩减; 利用属性粗化突显图中重要节点对图级表示的关键作用。通过将 HGP-GC 池化策略与现有神经网络相结合, 在不同规模公共数据集上的图分类实验结果证明了 HGP-GC 的有效性。

关键词: 图神经网络; 图池化; 层次图表示学习; 节点重要性; 图分类

中图分类号: TP391

文献标识码: A

Research on Hierarchical Graph Pooling Based on Graph Coarsening

CHEN Jie^{1,2,3}, XUE Yuan-yuan^{1,2,3}, CAO Jing-jing⁴, ZHAO Shu^{1,2,3,+}, ZHANG Yan-ping^{1,2,3}

¹ (Key Laboratory of Computational Intelligence and Signal Processing, Ministry of Education, Hefei 230601, China)

² (College of Computer Science and Technology, Anhui University, Hefei 230601, China)

³ (Key Laboratory of Information Materials and Intelligent Sensing of Anhui Province, Hefei 230601, China)

⁴ (Science and Technology Talent Exchange Development Service Center of Ministry of Science and Technology, Beijing 100045, China)

Abstract: Graph Neural Networks have been proved to be effective in modeling graph-structured data. The pooling mechanism is very important in the process of extracting graph hierarchical features using GNN models, which has attracted more and more researchers' attention in recent years. The existing hierarchical graph pooling methods based on clustering either need to add an extra neural network layer to achieve the coarsening of the feature graph, Or you can't capture the importance of the node in the graph from a global perspective. To solve the above problems, a Hierarchical Graph Pooling Based on Graph Coarsening is proposed for learning the hierarchical feature representation of graphs. This method mainly includes two parts: graph structure coarsening and graph attribute coarsening. Structure coarsening is used to reduce the size of the feature map. Attribute coarsening is used to highlight the key role of important nodes in the graph to the graph-level representation. By combining the HGP-GC pooling strategy with the existing neural network, the effectiveness of HGP-GC is proved by the experimental results of graph classification on common datasets of different sizes.

Key words: graph neural network ; graph pooling; hierarchical graph representation learning; node importance; graph classification

收稿日期: 2021-07-09 收修稿日期: 2021-08-09 基金项目: 国家重点研发计划子课题项目 (2017YFB1401903) 资助; 国家自然科学基金项目 (61876001) 资助。作者简介: 陈洁, 女, 1982 年生, 博士, 副教授, CCF 会员, 研究方向为机器学习、情感计算、三支决策; 薛远远, 男, 1996 年生, 硕士研究生, 研究方向为机器学习、网络表示学习、图神经网络; 曹京晶, 女, 1987 年生, 学位, 助理研究员, 研究方向为科技人才信息、科技政策、人才评价; 赵姝, 女, 1979 年生, 博士, 教授, CCF 会员, 研究方向为机器学习、社交网络、粒计算; 张燕平, 女, 1962 年生, 博士, 教授, CCF 会员, 研究方向为计算智能与商空间理论, 机器学习方法与应用, 人工神经网络与智能信息处理。

1 引言

受卷积神经网络 (Convolutional Neural Networks, CNNs) 在图像目标识别、语音识别和自然语言处理等领域^[1-3] 成功

应用的启发,许多研究人员试图将图卷积运算扩展到图结构数据中,例如社交网络^[4]、引文网络^[5]和生物网络^[6,7]等。近些年来,研究学者们提出了各种 GNN 模型,它们在图表示学习领域大都取得了显著成功,尤其是在节点级表示学习任务,如节点分类^[8,9]和链路预测^[10,11]等。然而,如果缺少池化机制,几乎现有的 GNN 模型都会缺乏学习图层次表示的能力,这使它们在图分类任务^[12]中的推广与应用受到很

大阻碍。在图数据挖掘领域中,图分类是一个重要的研究方向,许多实际任务中都有相关应用。例如,根据已经存在的化合物的信息预测未知化合物是否有毒以及判断 DNA 蛋白质是否发生突变等。目前如何将池化操作扩展到图数据吸引了越来越多研究者的关注。

早期一些工作通过全局汇总图中所有节点表示执行图的池化操作。然而采用这种方式生成的图级表示本质上依旧是“平坦的”。它们不能以分层的方式聚合节点表示,也就导致无法有效提取图潜在层次结构中包含的丰富信息。例如 Zhang 等人^[13]提出的排序图池化方法 SortPool,该方法首先根据图中节点的角色将节点重新排列为有意义的顺序,然后再进行池化操作。

最近有一些工作专注于 GNNs 中的分层池化过程,旨在学习图的层次表示,获得更完备的图级表示。本文根据模型建模方式将这些方法分为两大类:基于采样的层次图池化方法和基于聚类的层次图池化方法。

基于采样的层次图池化方法^[14-16],例如 SAGPool 算法^[14]和 Graph U-Nets 算法^[15],它们主要通过从原始图中选择前 k 个重要的节点生成更粗粒度的图作为下一卷积层的输入图。然而,采样图中重要节点生成的粗化图可能无法保留关键的局部结构信息,并且会破坏图拓扑结构的完整性。例如,在图中两个不直接相连但共享许多邻居的节点在生成的粗图中可能会变得彼此不可达,这将会阻碍节点之间消息的传递过程。

基于聚类的层次图池化方法^[17-20],它们通常将图的池化操作看做是节点的聚类问题,图中的节点被逐层分组合并以实现图的粗化。然而, Ying 等人^[17]提出的 Diffpool 算法需要额外的神经网络层计算存储分配矩阵,复杂度较高。另外,该方法采用加和的方式求解粗图中超节点的特征表示不能捕获节点在图中的地位和作用。Ekagra 等人^[18]提出的自适应感知池 ASAP 仅在每个子图中区分节点的重要性,并且该方法对图分层聚类后生成的子图之间存在重叠节点。

综上所述,目前基于聚类的层次图池化方法主要存在以下问题:它们或者需要增加额外的神经网络层实现图的粗化,或者没有从全局的角度区分节点的重要性大小。为了解决这些问题,本文提出一种基于图粗化的层次图池化方法 HGP-GC。该方法对特征图执行池化过程中主要包括图结构粗化和图属性粗化两大部分。图结构粗化部分并不需要额外

的神经网络层即可实现特征图尺寸的缩减。图属性粗化部分利用一个图注意力网络从全局角度捕获节点的重要性。实验证明, HGP-GC 在多个基准数据集上的图分类性能相比于基线方法都有不同程度的提升。简而言之,本文的主要贡献可归纳如下:

- 1) 本文提出一种基于图粗化的池化方法 HGP-GC 学习图的层次表示。该方法同时考虑图拓扑结构信息和节点特征信息捕获节点在全图上的重要性大小。
- 2) 对图执行下采样过程中, HGP-GC 不需要引入额外的神经网络层计算存储分类矩阵,而是直接通过定义稀疏位置矩阵记录节点之间的合并关系。
- 3) 本文在 5 个公共数据集上进行图分类实验,证明了 HGP-GC 方法与一系列最先进的图池化方法相比在图分类任务中的优越性。

2 相关工作

2.1 图神经网络

近些年来将卷积操作推广到图数据上的一些研究工作,它们大致可以分为两大类:谱方法和空间方法^[21]。谱方法通过图傅里叶变换和卷积定理定义图卷积运算,主要挑战在于谱域中定义的卷积滤波器未局限于节点域中。例如,谱卷积神经网络 (Spectral CNN)^[22]是最早提出在图结构数据上构建卷积神经网络的方法,该方法基于卷积定理通过学习卷积核实现图卷积操作,以完成节点之间的信息聚合。小波神经网络 (GWNN)^[23]用小波变换替代傅里叶变换实现卷积定理,不仅让图卷积神经网络满足了局部性,而且也大大降低了计算的复杂度。为了使图卷积神经网络在半监督学习领域发挥作用, Kipf 等人^[8]对切比雪夫网络进行简化并提出一阶图卷积神经网络模型。

与谱方法不同,空间方法在节点域中定义卷积,其卷积定义为位于目标节点附近的所有邻居节点的加权平均函数,主要挑战在于邻域的大小在每个节点之间变化很大。例如,图注意力网络 (GAT)^[9]通过注意力机制定义聚合函数,将邻居节点的表示以加权的形式聚合到自身。但节点之间的权重计算依赖于节点的特征表示,因此计算过程中需要加载整个图的节点特征,这阻碍了它在大规模图上的应用。与 GAT 需要考虑图中全部节点不同, Hamilton 等人^[24]提出图采样聚合网络 (GraphSAGE),该方法对邻居节点做随机采样,仅把采样的节点作为相关节点,这就避免了整个图数据的加载。Liang 等人^[25]从邻域采样和邻域聚合两个角度提出的 HGCN 模型,该模型根据关系紧密度衡量不同邻居对当前节点的重要程度,选择关系紧密度高的邻居节点进行采样,旨在减少随机采样带来的不确定性。最近,基于置信度的图卷积网络 (ConfGCN)^[26]认为节点是以一定的置信度为一个标签,因此, CongGCN 为图中每个节点学习置信度函数,并将其作用在节点相关性上用于修正聚合函数。

虽然如此,上文中提到的各种 GNN 模型大多都是用来学习有意义的节点级表示,这些模型由于缺少池化机制而无法提取图的层次结构信息,这严重阻碍了它们在图级表示学习任务中的推广与应用。

2.2 全局图池化方法

为了将图神经网络应用于图级表示学习相关的任务中,

需要一种方法汇总学习到的节点表示以生成图的表示。早期的一些方法采用全局池化方式,通过将图中所有节点表示进行融合得到图的表示。它们通常采用的融合方式包括求和 (Sum)、最大值 (Max) 和平均值 (Mean)。如公式 (1) 所示, Henaff 等人^[12] 证明,在模型的开始部分执行简单的操作可以减小图的尺寸,降低问题的求解复杂度。其中, k 表示卷积层的层数, H_i 表示第 i 个节点的嵌入向量。

$$Z_G = \text{sum/max/mean}(H_1^k, H_2^k, \dots, H_N^k) \quad (1)$$

例如 Vinyals 等人^[27] 提出 SET2SET 模型通过 LSTM 汇总图中所有节点表示实现全局池化操作。Zhang 等人^[13] 提出的排序图池化方法 SortPool 根据节点在图中的角色将节点重新排列为有意义的顺序,对图执行池化操作后,利用特征图值的最后一个通道将节点表示合并汇总。然而,这些方法不能以分层的方式提取图的特征信息,忽略了图中可能存在的层次结构,不利于研究人员为图级预测任务构建模型。

2.3 层次图池化方法

最近有一些研究工作使用分层池化的方法学习图的特征表示,它们可以分为以下两大类:基于采样的分层池化方法和基于聚类的分层池化方法。

SAGPool^[14] 和 Graph U-Nets^[15] 两种算法思想相似,它们均是根据图结构信息和节点的属性信息为图中每个节点学习一个标量,以此标量表征节点在图上的重要性程度,对此标量进行排序后选择前 k 个节点用于生成粗粒度图,从而学习图的层次表示。然而,这种方式尽管很有效,却可能会破坏图结构信息的完整性。

与采样的层次池化方法不同, Ying 等人^[17] 提出可微图池 Diffpool, 该方法使用图神经网络学习节点的低维度嵌入向量,然后根据节点的向量表示将图中相似的节点映射到一组簇中,把每一个簇看做超节点生成粗粒度的图作为下一神经网络层的输入图,从而实现以分层的方式推断和聚合节点的信息。然而, Diffpool 需要增加额外的神经网络层学习分配矩阵,软簇分配需要存储每一层中的分配矩阵,这导致该方法内存需求较大,限制了其在大规模图上的应用。用于学习图层次表示的自适应结构感知池 ASAP, 该模型首先对原始图分层聚类,聚类得到的每个簇是由节点及其一阶邻居组成。然后使用注意力机制计算簇中节点重要性得分,依据每个节点的注意力分数选择分数高的簇生成粗化图。然而, ASAP 池化方法仅在每个聚类生成的簇,即节点自身及其一阶邻居中区分节点的重要性程度,我们认为这种方式在一定程度上限制了模型对空间信息相关性的捕获能力。

为解决上述挑战,本文提出一种基于图粗化的层次图池化方法 HGP-GC。对图执行池化操作时, HGP-GC 不仅能够从全图的角度捕获节点的重要性大小,而且不需要增加额外的神经网络层用于分配矩阵的计算和存储。

3 预备知识

本节中,我们对图分类的数学定义以及 HGP-GC 池化策略实现过程中使用的相关技术进行简单介绍。

3.1 问题定义

对于任意无向图 $G = (V, \mathcal{E}, X)$, 其中 V 表示图 G 的节点集合, $|V| = N$ 表示图中共有 N 个节点, \mathcal{E} 表示边的集合。节点 $v_i \in V$ 具有由 x_i 表示的 d 维特征表示。 $X \in \mathcal{R}^{N \times d}$

表示图节点特征矩阵, $A \in \mathcal{R}^{N \times d}$ 表示图的邻接矩阵,用于定义节点之间的相互连接关系。 $G^p = (V^p, E^p, X^p)$ 为图 G 经过一次池化操作后的符号表示。 X_{coar} 和 A_{coar} 分别表示图 G^p 的节点特征矩阵和邻接矩阵。

给定数据集 $D = \{(G_1, y_1), (G_2, y_2), \dots\}$, 图分类的目标是学习一个映射 $f: \mathcal{G} \rightarrow \mathcal{Y}$, 其中 \mathcal{G} 表示输入图的集合, \mathcal{Y} 是与图关联的标签集合。

3.2 图卷积神经网络

图卷积神经网络 (GCN)^[9] 在各种具有挑战性的任务中都表现出了非常高效的性能。因此,本研究中使用的 GCN 模型提取用于图分类任务的节点特征信息。简要回顾其消息传递机制,对于 GCN 的第 l 层,它将图 G 的邻接矩阵 A 和节点特征矩阵 $H^{(l)}$ 作为输入,则第 $l+1$ 卷积层的输出节点特征表示为:

$$H^{(l+1)} = f(X_l, A_l) = \sigma(\tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}} H^{(l)} W^{(l)}) \quad (2)$$

其中, $\sigma(\cdot)$ 表示非线性激活函数。 \tilde{A} 等于 $A + I$, I 是单位矩阵。 \tilde{D} 是 \tilde{A} 的度矩阵, $\tilde{D}_{ii} = \sum_j \tilde{A}_{ij}$ 。 H 是当前层的节点特征矩阵,对于输入层, $H = X$, X 是输入图的属性矩阵。 $W^{(l)} \in \mathcal{R}^{d \times f}$ 是第 l 层中的可学习参数矩阵。

3.3 图傅里叶变换

传统傅里叶变换^[24] 将平方可积的函数 $f(t)$ 表示成复指数函数的积分或级数形式:

$$F(w) = F[f(t)] = \int_{-\infty}^{\infty} f(t) e^{-iwt} dt \quad (3)$$

由传统傅里叶变换可知,如果能够在图上找到一组基向量,就可以实现图上的傅里叶变换。

图上傅里叶变换的定义依赖于拉普拉斯矩阵的特征向量,以特征向量作为谱空间下的一组基底,则图上信号 $x \in \mathcal{R}^n$ 的傅里叶变换为:

$$\hat{x} = U^T x \quad (4)$$

其中, x 指信号在空间域的原始表示。 \hat{x} 指信号 x 变换到谱域后的表示。 U^T 表示特征向量矩阵的转置,用于傅里叶变换。因此,信号 x 的傅里叶逆变换可以表示为:

$$x = U \hat{x} \quad (5)$$

4 基于图粗化的层次图池化方法: HGP-GC

本研究旨在开发一个由卷积层和池化层组成的图神经网络模型,用于提取图结构数据的层次特征信息以应用于图分类任务。为了解决目前层次图池化模型中存在的一些问题,本文提出一种基于图粗化的池化方法 HGP-GC。该方法主要由两部分组成:图结构粗化和图属性粗化。

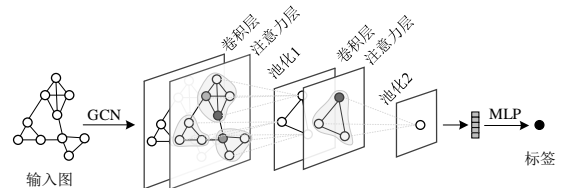


图1 结合 HGP-GC 池化方法的图神经网络

Fig. 1 Graph neural network combined with HGP-GC pooling method

图结构粗化主要解决目前已存在模型训练过程中需要增加额外的神经网络层问题。在利用谱聚类方法对节点聚类后, HGP-GC 模型直接通过定义一个位置信息矩阵用于确定超节点之间的连接关系, 实现特征图规模的缩小, 这使得以分层方式逐层提取图的特征信息成为可能。

图属性粗化利用图注意力网络计算节点在图中的重要性分数, 基于该分数更新节点的特征表示获得节点的增强特征表示。然后, 使用傅里叶变换原理实现属性粗化。

图 1 展示了结合 HGP-GC 方法的图神经网络模型体系结构。图中节点颜色的深浅表示该节点在图中的重要程度, 颜色越深表示节点在图中的地位越重要。给定一个输入图, 首先利用一个 GCN 卷积层学习节点的低维嵌入向量。然后注意力层综合考虑节点特征信息和图拓扑结构信息计算节点重要性得分更新节点的特征表示。最后, 利用位置矩阵实现图结构粗化, 使用傅里叶变换原理实现图属性粗化。该池化方法的具体实现过程如算法 1 所示:

算法 1: HGP-GC 图池化策略

输入: 图 $G = (V, \mathcal{E}, \mathbf{A})$ 和属性矩阵 \mathbf{X}

输出: 粗图 G^p 的邻接矩阵 \mathbf{A}_{coar} 和特征矩阵 \mathbf{X}_{coar}

```

1  节点特征表示矩阵:  $\mathbf{H}^{(l)} \leftarrow f(\mathbf{X}, \mathbf{A})$ 
2  子图集合:  $\mathcal{C} = \{G^1, G^2, \dots, G^k\} \leftarrow \text{谱聚类}$ 
3  // 图结构粗化
4  粗图邻接矩阵  $\mathbf{A}_{coar} \leftarrow \mathbf{S}_l^T \cdot \mathbf{A}_l \cdot \mathbf{S}_l$ 
5  // 图属性粗化
6  节点增强特征表示矩阵:  $\mathbf{X}' \in \mathcal{R}^{N \times d} \leftarrow \text{注意力层}$ 
7  傅里叶变换基底:  $\mathbf{U}_n \in \mathcal{R}^{N \times k} \leftarrow [\bar{\mathbf{u}}_n^1, \bar{\mathbf{u}}_n^2, \dots, \bar{\mathbf{u}}_n^k]$ 
8  粗图节点特征矩阵:
9   $\mathbf{X}_{coar} = \mathbf{U}_1^T \mathbf{X}' \parallel \mathbf{U}_2^T \mathbf{X}' \parallel \dots \parallel \mathbf{U}_n^T \mathbf{X}', n = 1, \dots, N_k$ 
return:  $\mathbf{A}_{coar}, \mathbf{X}_{coar}$ 

```

4.1 图结构粗化

图结构粗化是基于子图划分的, 目前已有多种方法可以将一个给定图划分成一组不含重叠节点的子图。本文采用谱聚类方法获取子图。图中节点经过聚类后形成一组子图, 每个子图看作一个超节点生成粗粒度的图。如何确定超节点之间的连接关系是结构粗化过程中面临的一大问题。

Diffpool 算法通过增加额外的神经网络层学习对应每一层的软分配矩阵确定超节点之间的关系, 这种方式不但复杂度高, 而且存储每次计算得到的分配矩阵导致模型的内存需求较高。不同于 Diffpool 模型, HGP-GC 通过定义每层的稀疏位置信息矩阵 \mathbf{S} 用于记录节点的合并关系。

给定的输入图 G , 矩阵 $\mathbf{A} \in \mathcal{R}^{N \times N}$ 和 $\mathbf{X} \in \mathcal{R}^{N \times d}$ 分别表示其邻接矩阵和节点特征矩阵。用 $\mathcal{C} = \{G^1, G^2, \dots, G^k\}$ 表示使用谱聚类方法对图中节点执行一次聚类操作后生成的子图集合。

定义第 l 层的位置矩阵 $\mathbf{S}_l \in \mathcal{R}^{N \times k}$ 用于记录图 G 中的节点在粗图 G^p 中的位置信息。位置矩阵 \mathbf{S}_l 的具体定义如公式 (6) 所示:

$$s_{ik} = \begin{cases} 1, & \text{若图 } G \text{ 中的节点 } v_i \text{ 聚类后属于子图 } G^k \\ 0, & \text{其它} \end{cases} \quad (6)$$

那么, 卷积神经网络第 l 层的输入图 G 经过结构粗化后生成的第 $l+1$ 层中的粗化图 G^p , 其邻接矩阵 \mathbf{A}_{coar} 可以被确定为:

$$\mathbf{A}_{coar} = \mathbf{S}_l^T \cdot \mathbf{A}_l \cdot \mathbf{S}_l \quad (7)$$

4.2 图属性粗化

网络中不同的节点起着大小不同的作用, 例如社交网络中有意见领袖的大 V , 交通网络中有至关重要的交通枢纽。本文对网络中节点重要性大小定义了一个指标, 用来定量地衡量每个节点在网络中重要性的大小。

对图执行池化操作过程中, 为了能够从全图角度捕获不同节点的重要性, HGP-GC 采用一种图注意力机制实现。注意力机制可以让模型更关注输入数据中和当前任务更相关的信息, 减少了对外部信息的依赖, 用来捕获节点特征的内部相关性。

给定输入图节点的特征矩阵 $\mathbf{X} \in \mathcal{R}^{N \times d}$, 利用公式 (8) 计算图中每个节点的重要性分数 $\mathbf{S}_c \in \mathcal{R}^{N \times 1}$ 。利用这种方式计算得到的节点重要性分数能够同时考虑图拓扑结构信息和节点的特征信息。

$$\mathbf{S}_c = \sigma \left(\tilde{\mathbf{D}}^{-\frac{1}{2}} \tilde{\mathbf{A}} \tilde{\mathbf{D}}^{-\frac{1}{2}} \mathbf{X}^{(l)} \boldsymbol{\Theta}^{(l)} \right) \quad (8)$$

其中, $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}$ 用于自连接, $\tilde{\mathbf{D}} \in \mathcal{R}^{N \times N}$ 是 $\tilde{\mathbf{A}}$ 的度对角矩阵, $\mathbf{X}^{(l)} \in \mathcal{R}^{N \times d}$ 表示第 l 层节点的特征矩阵, $\boldsymbol{\Theta} \in \mathcal{R}^{d \times 1}$ 是计算节点重要性分数所需的可学习参数, $\sigma(\cdot)$ 是非线性激活函数。

节点的重要性分数作为评价图中节点重要性指标, 基于该分数更新得到节点的增强特征表示。图注意力层的输入为所有节点特征集 $\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$, $\mathbf{x}_i \in \mathcal{R}^d$, 其中 N 为图中节点的个数, d 表示节点特征维度。定义输出新的节点特征集为 $\mathbf{x}' = \{\mathbf{x}'_1, \mathbf{x}'_2, \dots, \mathbf{x}'_N\}$, $\mathbf{x}'_i \in \mathcal{R}^d$ 。根据节点的重要性分数对节点自身的特征表示进行加权处理, 得到节点增强特征矩阵 $\mathbf{X}' \in \mathcal{R}^{N \times d}$ 。

$$\mathbf{x}'_i = (s_c)_i \mathbf{x}_i, \quad i = 1, 2, \dots, N \quad (9)$$

给定输入图 G , 利用谱聚类方法对图中节点进行聚类后生成的一组子图记为 G^1, G^2, \dots, G^k 。记 N_k 为子图 G^k 的节点数量, \mathcal{L}^k 表示子图 G^k 中的所有节点组成的列表, 即 $\mathcal{L}^k = [v_1, v_2, \dots, v_{N_k}]$ 。用 L_k 表示子图 G^k 的拉普拉斯矩阵, 其特征向量用 $\mathbf{u}_1^k, \mathbf{u}_2^k, \dots, \mathbf{u}_3^k$ 表示, 通过映射矩阵 $\mathbf{M}^k \in \mathcal{R}^{N \times N_k}$ 可以将子图 G^k 的特征向量 \mathbf{u}_i^k 映射到图 G 中, 并记为 $\bar{\mathbf{u}}_i^k$ 。

$$\mathbf{M}^k[i, j] = 1, \quad \text{当且仅当 } \mathcal{L}^k(j) = v_i \quad (10)$$

$$\bar{\mathbf{u}}_n^k = \mathbf{M}^k \mathbf{u}_n^k, \quad n = 1, 2, \dots, N_k \quad (11)$$

这里需要注意的是, 由于每个子图中的节点数量并不相同, 所以其拉普拉斯矩阵的特征向量的个数也不相同。故事先这样约定: 当一个子图中的节点数量时, 也就是子图不包含的情形, 此时令 $\mathbf{u}_n^k = \mathbf{0}$ 。实验中, 我们将参数 n 的值设置为 1。超参数 n 在图属性粗化过程中起到关键作用, 其不同取值对模型图分类性能影响详见实验分析部分。

记 $\mathbf{U}_n = [\bar{\mathbf{u}}_n^1, \bar{\mathbf{u}}_n^2, \dots, \bar{\mathbf{u}}_n^k]$, $\mathbf{U}_n \in \mathcal{R}^{N \times k}$ 作为傅里叶变换的一组基底。由于傅里叶变换将图信号变换到谱域过程中可

以同时考虑子图结构和节点特征信息,所以采用傅里叶变换对图属性信息进行粗化,可以进一步捕获不同子图结构中包含的丰富信息。则 \mathbf{X}_{coar} 可通过公式 (12) 求解得到。其中, \parallel 表示拼接操作。

$$\mathbf{X}_{coar} = \mathbf{U}_1^T \mathbf{X}' \parallel \mathbf{U}_2^T \mathbf{X}' \parallel \cdots \parallel \mathbf{U}_n^T \mathbf{X}', n = 1, \dots, N_k \quad (12)$$

5 实验与结果

为了验证 HGP-GC 池化方法的性能,本文在图分类任务上与目前相关的先进模型进行了比较。实验结果表明 HGP-GC 方法的图分类性能相比于基线方法具有更高的分类准确率。

5.1 数据集

本文在 5 个生物数据集上进行图分类实验,表 1 总结了实验数据集的详细统计信息。其中 G_{avg} , C_{avg} , V_{avg} 和 E_{avg} 分别代表每个数据集中图的数量、图的标签、节点数量以及边条数的平均值。下面简单介绍一下各个数据集的基本情况。

表 1 实验数据集

Table 1 Experimental datasets

	ENZYMES	PROTEINS	DD	NCI1	NCI109
G_{avg}	600	1113	1178	4110	4337
C_{avg}	6	2	2	2	2
V_{avg}	32.63	39.06	284.32	29.87	16.9
E_{avg}	62.14	72.82	715.66	32.3	17.88

ENZYMES^[29,30] 是一个蛋白质三级结构数据集,由 BRENDA 酶数据库中的 600 种酶组成,酶的分类是 6 个 EC 顶类中的一个。DD^[6,7] 是一个包含 1178 个蛋白质结构的数据集,每个图表示一个蛋白质结构,图中的节点代表氨基酸,图的标签表示该蛋白质是酶或非酶。PROTEINS^[7,29] 是一个包含 1113 个图的蛋白质数据集,图中的节点表示氨基酸序列的组成部分,图的标签表示该蛋白质是酶或非酶。NCI1 和 NCI109^[31,32] 数据集中的每个图表示一个化合物,图中的节点表示化合物的原子,边表示化学键,这两个数据集用于抗癌活性分类。

5.2 对比算法

本文实验部分对比了一些具有代表性和最新的具有不同池化方法的图神经网络模型,主要选择以下 6 种池化方法作为基线方法: SET2SET^[27]、SortPool^[13]、DiffPool^[17]、Graph U-Nets^[15]、SAGPool^[14] 和 ASAP^[18]。接下来对基线方法进行简单介绍。

SET2SET: 该方法建立在 GCN 的基础上,利用 Set2Set 方法获得整个图的表示,而不是对所有节点表示求平均值,本质上仍然是“平坦的”。

SortPool: 该基线方法提出一种端到端深度学习结构用于图分类。在将节点的特征输入传统的一维卷积之前,首先对节点特征进行排序,而不是对它们进行汇总,能够从全局图拓扑结构中学习图表示。该算法也是“扁平化”的,并不具备提取图层次特征的能力。

DiffPool: 该方法是一种可微的图池化方法,它能够以

端到端的方式适应各种图神经网络架构,以生成图的层次表示。该算法通过一个逐渐压缩信息的过程得到图的表示,而不是采用先得到所有节点的特征表示后进行全局池化操作一次性得到图的表示。

Graph U-Nets: 该方法基于池化层和上采样层构建一个作用在图数据上的编码器-解码器模型。池化层根据节点在可训练投影向量上的投影值对重要节点进行采样,上采样层则依据节点位置信息将图恢复至原始结构。该方法可应用于节点分类和图分类任务。

SAGPool: 该方法是一种基于自注意力机制的图池化方法,利用自注意力机制来区分应该保留和应该删除的节点,即采样重要节点构成粗粒度图。

ASAP: 该方法是一种稀疏可微的层次图池化方法,它对局部子图进行分层聚类,并利用自注意力机制使模型能够更好的捕捉子图内部节点的重要性,然后选择具有代表性的子图生成粗粒度图,实现对图的下采样操作。

5.3 实验设置及评价指标

实验过程中,每个数据集被随机拆分成 3 部分,即 80% 作为训练集,10% 作为验证集,10% 作为测试集。每次实验重复 10 次这种随机拆分,并记录 10 中不同拆分情形下的平均性能作为实验结果。

为了更直观的体现本文所提出模型性能的有效性,根据以往的工作,实验采用广泛使用的评价指标,即图分类的测试准确率来评估模型的性能。

5.4 图分类结果

本节将 HGP-GC 池化方法与上述 6 种基线方法在 5 个公用数据集上的图分类结果进行比较,分类结果详见表 2。这里需要特别说明: 除 ENZYMES 数据集,其余所有数据集上的图分类结果均引用于文献^[18]中。从该表中,我们可以得出以下结论:

- 1) 首先,从表中可以观察到一个普遍的结果,即 HGP-GC 池化方法在所有数据集上的图分类性能均优于其它基线方法。
- 2) 其次, HGP-GC 方法与基于采样的层次图池化方法 (Graph U-Nets 和 SAGPool 方法) 相比,其图分类性能在 5 个数据集上都有相对较高的提升。这是因为采样图中重要节点压缩特征图的方式可能会忽略图中关键局部结构中的信息,并且会破坏原始图拓扑结构的完整性。
- 3) 最后, HGP-GC 方法相比于 DiffPool, 在 DD, NCI1 和 NCI109 数据集上的图分类性能提高了约 10%。特别地,与 HGP-GC 方法较为类似的 ASAP 方法相比,在 ENZYMES 数据集上的分类精度提高了约 24.97%, 其他数据集上的结果也均有不同程度的提高。HGP-GC 从全图的角度捕获不同节点的重要性大小,而 ASAP 算法则只在每个子图中区分节点的重要性后更新节点的特征表示。

5.5 区分节点重要性对模型性能的影响分析

图中不同节点具有不同的地位和角色,重要节点对图级表示学习的作用更为关键,如何有效捕获图中不同节点的重要性大小是图数据挖掘领域的一大挑战。

表3 节点重要性对模型的影响

Table 3 The influence of node importance on the model

Dataset	ENZYMES	DD	NCI1
HGP-GC_noatten	51.67±1.67	76.18±1.65	72.87±1.68
HGP-GC	52.67±1.33	78.07±0.77	73.45±0.97

HGP-GC 方法属性粗化部分通过一个注意力层捕获节点在图中的重要性大小。为了验证注意力层对模型性能的影响，本文设计了 HGP-GC 方法的一种变体方法 HGP-GC_noatten，即利用 GCN 学习节点低维特征表示后，不进行节点重要性的区分，直接利用傅里叶变换对图执行属性粗化。该模型在 ENZYMES, DD 和 NCI1 数据集上进行图分类任务，实验结果如表 3 所示。从表中可以观察到，对图执行属性粗化之前捕获节点的重要性的图分类精确度都有不同程度的提升。

表2 HGP-GC 模型图分类实验结果

Table 2 Experimental results of HGP-GC model graph classification

Model	ENZYMES	DD	PROTEINS	NCI1	NCI109
SET2SET	25.83±1.08	71.60±0.87	72.16 ± 0.43	66.97 ± 0.74	61.04 ± 2.69
SortPool	51.94±1.34	71.87±0.96	73.91 ± 0.72	68.74 ± 1.07	68.59 ± 0.67
Diffpool	51.33±1.50	66.95±2.41	68.20 ± 2.02	62.32 ± 1.90	61.98 ± 1.98
Graph U-Nets	40.00±1.67	75.01±0.86	71.10 ± 0.90	67.02 ± 2.25	66.12 ± 1.60
SAGpool	23.33±1.33	76.45±0.97	71.86 ± 0.97	67.45 ± 1.11	67.86 ± 1.41
ASAP	27.70±0.86	76.87±0.70	74.19 ± 0.79	71.48 ± 0.42	70.07 ± 0.55
HGP-GC	52.67±1.33	78.07±0.77	75.59±1.20	73.45 ± 0.97	71.68 ± 0.59

5.6 超参数分析

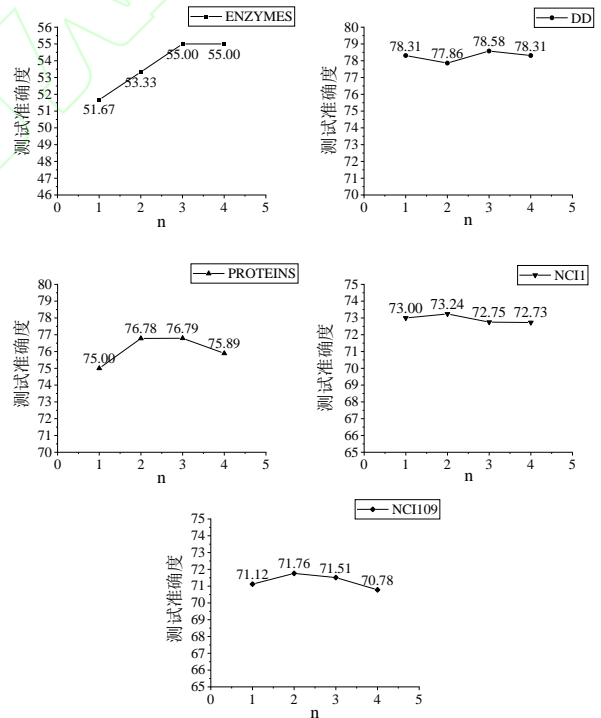
模型训练过程中，利用注意力层得到节点增强特征表示后，本文使用了傅里叶变换原理对图进行属性粗化。其中子图特征向量个数 n 为模型的一个关键超参数。为了分析其不同取值情况下对模型图分类准确度的影响，本文通过在不同尺度上改变该超参数来进一步研究分析。图 2 展示了超参数 n 取值分别为 1,2,3,4 情况下 HGP-GC 方法的图分类准确度。从该图中可以得出以下结论：

在数据集 DD, NCI1 和 NCI109 上，超参数 n 的不同取值对模型的图分类结果影响相对平缓，模型具有较好的鲁棒性。而在数据集 ENZYMES 和 PROTEINS 上，超参数 n 取值为 2 时，模型的图分类精度分别提高了约 1.66% 和 1.78%，出现这种情况的原因在于这两个数据集中的节点附带了属性信息，使得模型可以捕获更多的节点属性信息。另外，在 n 取值为 4 时，部分数据集上的分类准确度有所下降，我们分析出现这种情况的原因是神经网络模型过拟合问题造成的。

6 结论与展望

图神经网络是机器学习和数据挖掘的一个重要研究领域，利用图神经网络模型提取图潜在层次结构中包含的丰富信息对将其应用于图级任务具有重要意义。本文提出了一种基于图粗化的层次图池化方法 HGP-GC。该方法对特征图粗化过程中不需要增加额外的神经网络层即可以记录节点之间的合并关系，而且该方法综合考虑了节点的特征信息和图拓扑结构信息来捕获节点在全图上的重要性程度，旨在突出关键节点对图级表示的重要作用。本文将 HGP-GC 池化方法与现有的图神经网络相结合，并以端到端的方式提取图潜在层次结构中包含的丰富信息，学习图的层次表示。在 5 个公用基准数据集上的图分类实验结果证明了该池化方法的有效性。

使用谱聚类对网络中的节点进行聚类操作时涉及到图拉普拉斯矩阵的分解，时间和空间复杂度相对较高。接下来的工作可考虑寻找一种新的聚类方法降低粗化过程中的复杂度问题，并在更多的任务上评估模型的性能。

图2 超参数 n 对模型性能的影响Fig. 2 The influence of super parameter n on model performance

References:

- [1] He Kai-ming, Zhang Xiang-yu, Ren Shao-qing, et al. Deep Residual learning for image recognition[C] //IEEE Conference on Computer Vision and Pattern Recognition

- (CVPR), Las Vegas, USA, 2016:770-778.
- [2] Andrej Karpathy, George Toderici, Sanketh Shetty, et al. Large-scale video classification with convolutional neural networks[C] // IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 2014:1725-1732.
 - [3] Dzmitry Bahdanau, Kyunghyun Cho, Yoshua Bengio. Neural machine translation by jointly learning to align and translate[C] // International Conference on Learning Representations (ICLR). San Diego, CA, USA, 2015.
 - [4] Lars Backstrom, Jure Leskovec. Supervised random walks: predicting and recommending links in social networks[C] // 4th International Conference on Web Search and Web Data Mining (WSDM), Hong Kong, China, 2011:635-644.
 - [5] Jure Leskovec, Jon Kleinberg, Christos Faloutsos. Graphs over time : densification laws, shrinking diameters and possible explanations[C] // 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), Chicago, Illinois, USA, 2005:177-187.
 - [6] Nino Shervashidze, Pascal Schweitzer, Erik Jan van Leeuwen, et al. Weisfeiler-lehman graph kernels[J] Journal of Machine Learning Research, 2011, 12(77): 2539-2561.
 - [7] Paul D Dobson, Andrew J Doig. Distinguishing enzyme structures from non-enzymes without alignments[J] Journal of Molecular Biology, 2003, 330(4):771-783.
 - [8] Thomas N Kipf, Max Welling. Semi-supervised classification with graph convolutional networks[C] // International Conference on Learning Representations, Toulon, France (ICLR), 2017.
 - [9] Petar Veličković, Guillem Cucurull, Arantxa Casanova, et al. Graph attention networks[C] // International Conference on Learning Representations (ICLR), Vancouver, BC, Canada, 2018.
 - [10] Michael Sejr Schlichtkrull, Thomas N Kipf, Peter Bloem, et al. Modeling relational data with graph convolutional networks[C] // Semantic Web Fifteenth International Conference. Greece, 2018:593-607.
 - [11] Shikhar Vashishth, Soumya Sanyal, Vikram Nitin, et al. Composition-based multi-relational graph convolutional networks[C] // International Conference on Learning Representations (ICLR), Addis Ababa, Ethiopia, 2020.
 - [12] Mikael Henaff, Joan Bruna, Yann LeCun. Deep convolutional networks on graph-structured data[Z] ArXiv Preprint, 2015:1506.05163.
 - [13] Zhang Mu-han, Cui Zhi-cheng, Marion Neumann, et al. An end-to-end deep learning architecture for graph classification[C] // 32nd AAAI Conference on Artificial Intelligence (AAAI), New Orleans, Louisiana, USA, 2018:4438-4445.
 - [14] Junhyun Lee, Inyeop Lee, Jaewoo Kang. Self-attention graph pooling[C] // 36th International Conference on Machine Learning (ICML), Long Beach, California, USA, 2019:3734-3743.
 - [15] Gao Hong-yang, Ji Shui-wang. Graph u-nets[C] // 36th International Conference on Machine Learning (ICML), California, USA, 2019:2083-2092.
 - [16] Zhang Zhen, Bu Jia-jun, Martin Ester, et al. Hierarchical graph pooling with structure learning[Z]. ArXiv Preprint 2019:1911.05954.
 - [17] Ying Zhi-tao, You Jia-xuan, Christopher Morris, et al. Hierarchical graph representation learning with differentiable pooling[C] // 32nd International Conference on Neural Information Processing Systems (NIPS), Canada, 2018:4805-4815.
 - [18] Ekagra Ranjan, Soumya Sanyal, Partha Pratim Talukdar. ASAP: adaptive structure aware pooling for learning hierarchical graph representations[C] // AAAI Conference on Artificial Intelligence (AAAI), New York, USA, 2020:5470-5477.
 - [19] Ma Yao, Wang Sun-hang, Charu C. Aggarwal, et al. Graph convolutional networks with eigenPooling[C] // 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 2019: 723-731.
 - [20] Yuan Hao, Ji Shu-wang. StructPool: structured graph pooling via conditional random fields[C] // 8th International Conference on Learning Representations (ICLR), 2020.
 - [21] Xu Bing-bing, Cen Ke-ting, Huang Jun-jie, et al. A survey on graph convolutional neural network[J]. Chinese Journal of Computers, 2020, 43(5): 755-780.
 - [22] Joan Bruna, Wojciech Zaremba, Arthur Szlam, et al. Spectral networks and locally connected networks on graphs[C] // Secod International Conference on Learning Representations (ICLR), Canada, 2014.
 - [23] Xu Bing-bing, Shen Hua-wei, Cao Qi, et al. Graph wavelet neural network[C] // 7th International Conference on Learning Representations (ICLR), New Orleans, LA, USA, 2019.
 - [24] William L Hamilton, Ying Zhi-tao, Jure Leskovec. Inductive representation learning on large graphs[C] // Advances in Neural Information Processing Systems (NIPS), Long Beach, CA, USA, 2017:1024-1034.
 - [25] Liang Shun-pan, Tu Hao, Wang Rong-sheng, et al. Knowledge graph recommendation algorithm combining importance sampling and pooling aggregation[J]. Journal of Chinese Computer Systems, 2021, 42(5):967-971.
 - [26] Shikhar Vashishth, Prateek Yadav, Manik Bhandari. Confidence-based graph convolutional networks for semi-supervised learning[C] // 22nd International

- Conference on Artificial Intelligence and Statistics (AISTATS), Naha, Okinawa, Japan, 2019:1792-1801.
- [27] Oriol Vinyals, Samy Bengio, Manjunath Kudlur. Order matters: sequence to sequence for sets[C] //4th International Conference on Learning Representations (ICLR), San Juan, Puerto Rico, 2016.
- [28] Li She-lei, Yang Bo-xiong, Lu Jiao-jiao. Spectral fourier transform and spectral wavelet transform analysis[J]. Computer technology and development. 2021, 31(5): 85-89.
- [29] Karsten M Borgwardt, Cheng Soon Ong, Stefan Schöner, et al. Protein function prediction via graph kernels[C] //13th International Conference on Intelligent Systems for Molecular Biology, MI, USA, 2005:47-56.
- [30] Ida Schomburg, Antje Chang, Christian Ebeling, et al. BRENDA, the enzyme database: updates and major new developments[J]. Nucleic Acids Research, 2004, 32(90001):431-433.
- [31] Nikil Wale, George Karypis. Comparison of descriptor spaces for chemical compound retrieval and classification[C] //6th IEEE International Conference on Data Mining (ICDM), Hong Kong, China, 2006:678-689.
- [32] Nino Shervashidze, Pascal Schweitzer, Erik Jan van Leeuwen, et al. Weisfeiler-lehman graph kernels[J]. Journal of Machine Learning Research, 2011, 12(77):2539-2561.

附中文参考文献:

- [21] 徐冰冰, 岑科廷, 黄俊杰, 等. 图卷积神经网络综述[J]. 计算机学报, 2020, 43(5):755-780.
- [25] 梁顺攀, 涂浩, 王荣生, 等. 融合重要性采样和池化聚合的知识图推荐算法[J]. 小型微型计算机系统, 2021, 42(5):967-971.
- [28] 李社蕾, 杨博雄, 陆娇娇. 谱图傅里叶变换与谱图小波变基分析研究[J]. 计算机技术与发展, 2021, 31(5):85-89.