DOI:10.20009/j.cnki.21-1106/TP.2022-0110
2023 年 10 月 第 10 期
Vol. 44 No. 10 2023

双通道知识蒸馏的节点分类方法

王新生,朱小飞,黄贤英

(重庆理工大学 计算机科学与工程学院,重庆 400054)
E-mail:zxf@cqut.edu.cn

摘 要: 近年来,基于教师-学生的知识蒸馏框架在图神经网络方面取得了较好的表现.然而这类知识蒸馏框架仍存在一些问题,如教师模型的知识信息不够全面,不能很好地指导学生模型;学生模型自身学习能力较差.为了解决这两方面的问题,本文提出了基于双通道知识蒸馏的节点分类方法.具体而言,该方法引入双教师模型,分别从拓扑结构和特征属性两个方面进行学习,保证了教师模型知识信息的多样性和全面性.学生模型采用参数化标签传播和邻居特征聚合两种预测机制,保证其具有更好的学习能力.最终,双教师模型分别从拓扑结构和特征属性两个方面对学生模型进行指导.在 5 个真实数据集上的实验结果表明该模型与最优基准模型相比具有更好的分类效果.

关 键 词: 知识蒸馏;图神经网络;节点分类;标签传播;注意力机制
中图分类号: TP391 **文献标识码:** A **文 章 编 号:** 1000-1220(2023)10-2284-07

Dual-channel Knowledge Distillation for Node Classification

WANG Xin-sheng,ZHU Xiao-fei,HUANG Xian-ying
(College of Computer Science and Engineering,Chongqing University of Technology,Chongqing 400054,China)

Abstract: In recent years,the teacher-student knowledge distillation framework has achieved encouraging performance in graph neural network. However,there are still some problems in this kind of knowledge distillation framework. For example,the knowledge information of teacher model is not comprehensive enough to guide student model well and student model has poor learning ability. To solve the two issues,this paper presents a node classification method based on dual channel knowledge distillation. Specifically,the method introduces a dual teacher model which learns from two perspectives,including topological structure and feature attribute,in order to maintain the diversity and comprehensiveness of the knowledge information in teacher model. The student model employs the parameterized label propagation and feature aggregation of neighbors to ensure a better learning capability. Finally,the dual teacher model is leveraged to guide the student model from the two perspectives. Experimental results on five real word datasets show that this model has better classification performance than the best baseline model.

Key words: knowledge distillation;graph neural network;node classification;label propagation;attention mechanism

1 引 言

图结构数据的半监督学习的目的是在给定网络结构、节点特征和标签节点子集的情况下对网络中的每个节点进行分类.在现实世界中,分类问题有十分广泛的应用,如节点分类[1]、推荐系统[2]和文本分类[3]等.研究表明,这些应用大多数都存在同质现象[4],即假设具有连边的节点往往具有相似的标签.在同质假设下,许多传统方法使用随机游走方式来传播标签[5,6],或者是正则化邻居节点之间的标签差异[7,8]来传播标签.

随着深度学习的兴起,知识蒸馏(Knowledge Distillation,KD)已经证明了它在神经网络方面的有效性.然而,知识蒸馏的研究大多集中在卷积神经网络(Convolutional Neural Networks,CNNs)[9]上,并且以常规数据作为输入实例,能够处理不规则数据的图神经网络(Graph Neural Networks,GNNs)[10]却很少受到关注.一个显著的差异是,GNNs[10]在跨网络层的特征嵌入更新中包含了拓扑信息,而大多数知识蒸馏的研究没有考虑到这一点,限制了它们对 GNNs[10]的潜在扩展.最近的知识蒸馏研究[11],提出通过转移局部结构将知识蒸馏与 GNNs[10]结合建模.实验结果表明,这种方法能够很好地解决上述问题.然而,该方法有两个主要缺点:1)知识蒸馏框架集中在单方面的知识指导,忽略了其他层面的知识,导致学生模型不能充分地从教师模型汲取知识信息;2)知识蒸馏框架未能很好地利用同质现象,忽略了原有数据集中的先验知识.

针对上述的 2 个问题,本文提出了一个双通道的知识蒸馏框架,将基于拓扑结构和特征属性的教师模型所学习到的知识注入到设计良好的学生模型中.教师模型分别为基于拓扑图的教师模型和基于特征图的教师模型,前者学习空间拓扑结构的知识信息,后者学习节点及邻居节点特征属性的知识信息,从两个不同的知识层面给予学生模型指导.学生模型

收稿日期:2022-02-16 收修改稿日期:2022-04-13 基金项目:国家自然科学基金项目(62141201)资助;重庆市自然科学基金面上项目(CSTB2022NSCQ-MSX1672)资助;重庆市教育委员会科学技术研究计划重大项目(KJZD-M202201102)资助. 作者简介:**王新生**,男,1997 年生,硕士研究生,CCF 会员,研究方向为图神经网络和自然语言处理;**朱小飞**(通讯作者),男,1979 年生,博士,教授,CCF 高级会员,研究方向为自然语言处理、数据挖掘与信息检索;**黄贤英**,女,1967 年生,硕士,教授,CCF 高级会员,研究方向为自然语言处理、信息检索和机器学习.

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

包括参数化标签传播模块和邻居特征聚合模块,二者根据同质现象进行设计. 前者根据邻居节点的标签表示分配不同权重,加强自身的标签表示,进而保留了基于结构的先验知识;后者采用图注意力网络(Graph Attention Network,GAT)[12],根据邻居特征表示分配不同权重,加强自身的特征表示,进而保留了基于特征的先验知识. 此外,采用多层融合模式,每一层都会将两个模块的知识进行融合,进一步提升学生模型自身的学习效果.

本文的主要贡献包括 3 个方面:

1)提出了一个双通道的知识蒸馏框架来提取 GNNs[10] 模型学习到的拓扑结构和特征属性方面的知识信息,并将其注入到学生模型中进行更有效的预测;

2)将学生模型设计为参数化标签传播模块和邻居特征聚合模块的可训练组合,使得学生模型保留了基于结构和基于特征的先验知识,同时保证了学生模型具有一定的学习能力;

3)在 5 个真实数据集上进行了实验,实验结果证明了本文所提出方法的有效性,且与最优的方法相比具有更好的分类效果.

2　相关工作

2.1　图神经网络

近几年,图神经网络在不规则数据的各种任务上取得了巨大成功,如节点分类、蛋白质属性预测等. 本文将简要介绍图神经网络中的 GAT[12].

图结构数据常常含有噪声,意味着节点与节点之间的边有时不是那么可靠,对于同一节点来说,邻居节点的相对重要性也有差异. 因此,为了解决邻居节点差异性这个问题,提出了 GAT[12]. 主要做法是在图算法中引入“注意力”机制. 通过计算当前节点与邻居的“注意力系数”,在聚合邻居的特征表示时进行加权,使得图神经网络能够更加关注重要的节点,以减少边噪声带来的影响,进而提升模型效果.

最近的一些研究表明,通过引入传统的预测机制,如标签传播,GNNs[10] 的性能可以得到进一步的提高. 例如,广义标签传播(Generalized Label Propagation,GLP)[13] 修改了图卷积滤波器以生成平滑特征,并对图的相似性进行编码. UniMP[14] 通过共享消息传递网络进而融合特征聚合和标签传播. GCN-LPA[15] 采用标签传播作为正则化,以帮助图卷积网络(Graph Convolution Neural Network,GCN)[16,17] 获得更好的性能. 值得注意的是,标签传播机制是用简单的基于结构的先验知识构建的. 以上的改进表明,这种先验知识在 GNNs[10] 中没有得到充分的探索.

2.2　知识蒸馏

知识蒸馏是 2015 年[18] 首次被提出的,其目标是将知识从典型的教师模型中提取出来,将其输入到参数量较少的学生模型,使学生模型的表现与教师模型相似. 这样,学生模型可以在不损失预测质量的情况下减少时间和空间复杂性. 知识蒸馏在计算机视觉领域得到了广泛的应用,例如将一个深度卷积神经网络(Convolution Neural Network,CNN)[9] 压缩成一个浅层神经网络来加速推理.

近年来,随着 GNN 在各种图任务上取得了巨大成功,将知识蒸馏和 GNN 相结合的研究也越来越多. 然而,他们的动机和模型架构和本文有很大的不同. Yang 等人[11] 在计算机视觉领域提出了一种利用局部结构保持模块,将具有较大特征映射的深层图卷积网络压缩为参数较少的浅层 GCN. Furlanello 等人[19] 通过可靠的数据蒸馏(Reliable Data Distillation,RDD)[20] 训练了多个具有相同架构的 GCN 学生模型,然后以类似于 BAN[19] 的方式将他们集成以获得更好的性能. 图马尔可夫神经网络(Graph Markov Neural Networks,GMNN)[21] 也可以看作是一种知识蒸馏方法,其采用两个参数大小不同的 GCN 相互学习.

与之相比,本文提出的知识蒸馏框架简单有效且非常全面. 两种教师模型包含图数据的全部信息,即拓扑结构和特征属性,使得学生模型能够全面汲取知识信息. 此外,在训练过程中,不需要在教师模型和学生模型之间进行整合或迭代提取,使得训练任务更加简单有效.

3　双通道知识蒸馏的节点分类

本章节将介绍半监督的节点分类任务的输入信息以及相关表示符号. 本文提出一个基于拓扑结构和特征属性的双通道知识蒸馏框架,如图 1 所示. 它由两种教师模型和学生模型

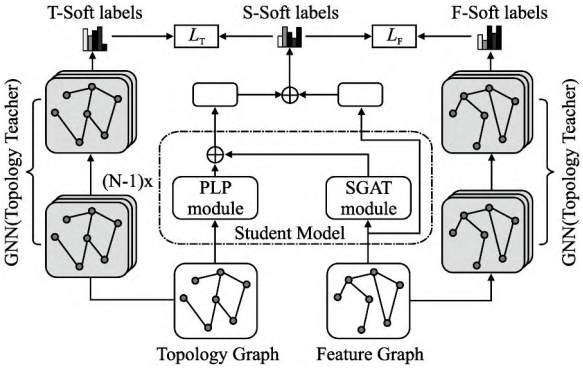

图 1　MPS 结构图

Fig. 1　Architecture of MPS

组成:基于拓扑图的教师模型和基于特征图的教师模型,学生模型由参数化标签传播模块和邻居特征聚合模块组成.

3.1　半监督节点分类

节点分类任务的原始信息为一个有 n 个节点的连通图 $G=(V,E,X)$,其中 $V=\{v_1,v_2,\cdots,v_n\}$ 记为 n 个节点的集合,$E=\{e_{ij}\}$ 记为边的集合,$X=\{x_1,x_2,\cdots,x_n\}\in\mathbb{R}^{n\times d}$ 记为所有节点的特征集合,x_i 是节点 v_i 的 d 维的特征向量. 邻接矩阵 $A=\{a_{ij}\}\in\mathbb{R}^{n\times n}$ 记为 G 的空间拓扑结构,如果节点 v_i 和节点 v_j 之间有边 e_{ij},那么 $a_{ij}>0$;否则 $a_{ij}=0$.

半监督节点分类任务的节点集合 V 将被划分为两部分:有标签的节点集合 $V_L\subset V$,其中 $X_L\subset X$;无标签的节点集合 $V_U=V\backslash V_L$,其中 $X_U=X\backslash X_L$.

3.2　教师模型

在知识蒸馏模型中,当教师模型的输入信息一定,那么该教师模型所学习到的知识就局限于输入的信息,给予学生模型的指导也只停留在单方面信息. 因此,本文提出双通道模型,通过不同的输入信息得到不同知识层面的教师模型,给予

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

学生模型不同知识层面的指导.

3.2.1 基于拓扑图的教师模型

本章节将对基于拓扑图的教师模型(Topology Teacher, Topology-Teacher)进行分析说明.

首先,将节点分类任务的原始信息,即 $G_T=(V,E,X)$,作为输入信息来通过图神经网络 GNN 模型,在本文中 GNN 模型使用的是 GAT[12]. 通过 GNN 模型得到一个预训练分类器,记为 $f_\theta:(\tilde{Y}_L,\tilde{Y}_U)\leftarrow f_\theta(X,A,Y_L)$,其中 θ 记为初始的模型参数,$\tilde{Y}_L$ 和 $\tilde{Y}_U$ 分别是节点集合 V_L 和 V_U 被预测的标签.

正常情况下,被预测的标签 $\tilde{Y}_L$ 要尽可能和真实标签 $\tilde{Y}_L$ 接近:

$$\theta^*=\arg\min_\theta distance(\tilde{Y}_L,Y_L)=\arg\min_\theta distance(f_{\theta\text{-}GNN}(G_T),Y_L) \tag{1}$$

其中,θ^* 记为训练得到的模型参数,$distance(\cdot)$ 表示计算两个标签集合之间的距离. 在这项工作中,使用的是欧式距离[1].

值得注意的是,知识蒸馏[22]中,学生模型训练时,通过拟合预训练教师模型的软标签来提升自身学习能力,最终超越教师模型. 因此,教师模型的最终预测结果将作为训练学生模型的指导信息.

3.2.2 基于特征图的教师模型

本节将对基于特征图的教师模型(Feature Teacher, Feature-Teacher)进行分析说明.

在这个模块,通过计算节点特征之间的余弦相似度,得到特征图. 即由节点特征的相似性来得到节点的高置信度邻居. 给定一对节点(v_i,v_j),它们的余弦相似度由如下公式计算得到:

$$S_{ij}=\frac{x_i^T x_j}{\|x_i\|\|x_j\|} \tag{2}$$

根据上述计算得到的每对节点的余弦相似度,为每个节点选择余弦相似度最大的 k 个节点作为邻居节点,得到新的边集合 E_F 和邻接矩阵 A_F. 最终得到特征图,记为 $G_F=(V,E_F,X)$.

同理,将 $G_F=(V,E_F,X)$ 输入 GNN 模型,得到另一个预训练分类器,即基于特征图的教师模型,记为 $f_\beta:(\tilde{Y}_L,\tilde{Y}_u)\leftarrow f_\beta(x,A_F,Y_L)$,其中 β 记为初始的模型参数.

最终得到训练后的模型参数 β^*:

$$\beta^*=\arg\min_\beta distance(\tilde{Y}_L,Y_L)=\arg\min_\beta distance(f_{\beta\text{-}GNN}(G_F),Y_L) \tag{3}$$

3.3 学生模型

学生模型包含两个模块:1)参数化标签传播模块,其能保留基于结构的先验知识;2)邻居特征聚合模块,其能保留基于特征的先验知识;最终将二者所学习到的知识进行融合.

3.3.1 参数化标签传播模块

在前面提到过同质现象,即具有连边的节点往往具有相似的标签. 针对这一现象,之前有许多研究工作都利用了同质现象来加强节点的表示. 早期的标签传播[23]方法在传播过程中平等地对待节点的所有邻居,即对于所有邻居赋予相同的权重. 然而,不同的邻居节点对节点的重要性是不同的. 因此,参数化标签传播对于不同邻居节点得到不同的参数置信分数,根据更大的参数置信分数赋予邻居节点更大的权重[24].

首先,本文用 f_{PLP} 表示参数化标签传播模块(Parameterized Label Propagation, PLP)的最终预测结果,f_{PLP}^m 表示 m 轮后 PLP 的预测结果. 接下来,将初始化每个节点 v 的一个预测结果表示:如果节点 v 是一个有标签的节点,那么就用一个 one-hot 向量来表示;否则,将对所有无标签的节点用一个统一的标签向量来表示. f_{PLP} 的初始化如下:

$$f_{PLP}^0(v)=\begin{cases}(0,\cdots,1,\cdots,0)\in\mathbb{R}^{|Y|}, & \forall_v\in V_L\\ \left(\frac{1}{|Y|},\cdots,\frac{1}{|Y|},\cdots,\frac{1}{|Y|}\right)\in\mathbb{R}^{|Y|}, & \forall_v\in V_U\end{cases} \tag{4}$$

$f_{PLP}^m(v)$ 表示节点 v 经过 m 次迭代之后的预测概率分布,$|Y|$ 表示节点的类别个数. 在第 $m+1$ 次迭代前,PLP 将会更新每个无标签节点 $v\in V_U$ 的预测表示:

$$f_{PLP}^{m+1}(v)=\sum_{u\in N_v\cup|v|}w_{uv}f_{PLP}^m(u) \tag{5}$$

其中 N_v 表示节点 v 的邻居节点集合,w_{uv} 表示节点 u 和节点 v 通过 *softmax* 函数计算出的边的权重:

$$w_{uv}=\frac{\exp(c_v)}{\sum_{u'\in N_v\cup|v|}\exp(c_{u'})} \tag{6}$$

在这里,参数置信分数 c_v 作为可选操作对于 *inductive* 和 *transductive* 任务:

$$c_v=z^T x_v \tag{7}$$

其中 $z\in$ 是一个可学习的参数.

3.3.2 邻居特征聚合模块

图结构数据主要包括结构知识和特征知识,在上一小节已经通过 PLP 模块学习到基于结构的先验知识,在这一小节将引入单层图注意力网络模块(Single Graph Attention network, SGAT),使得学生模型进一步学习基于特征的先验知识. 特征图 $G_F=(V,E_F,X)$ 是根据节点特征相似性得到的,并且是挑选每个节点最相似的 k 个节点作为其邻居,因此可以得到,每个节点及其一阶邻居是按照节点特征预先分好的簇,每个节点是该簇的中心. 进一步,将特征图 $G_F=(V,E_F,X)$ 输入 SGAT 模块,通过每个节点的邻居节点特征加强自身表示,最终得到每一个节点的特征表示:

$$h'_i=\sigma\left(\sum_{j\in N_i}\alpha_{ij}\mathbf{W}_h h_j\right) \tag{8}$$

其中 $\mathbf{W}_h$ 是一个可学习的参数矩阵,N_i 表示节点 i 的邻居集合,h_j 为节点 j 的特征表示,h_j 初始化表示为 $x_j\in X$,h'_i 表示节点 i 聚合邻居特征后的最终表示. α_{ij} 表示节点 i 与其邻居节点 j 的注意力分数:

$$\alpha_{ij}=\text{softmax}_j(e_{ij})=\frac{\exp(e_{ij})}{\sum_{k\in N_i}\exp(e_{ik})} \tag{9}$$

$$e_{ij}=f_{FNN}([\mathbf{W}_i h_i\,\|\,\mathbf{W}_j h_j]) \tag{10}$$

其中 $\mathbf{W}_i$ 和 $\mathbf{W}_j$ 是可学习的参数矩阵,f_{FNN} 表示单层前馈神经网络函数.

值得注意的是,SGAT 模块为了减少聚合节点邻居特征的噪声,只聚合每个节点的一阶邻居. 这样的做法充分利用了特征图的特点,只聚合每个节点最相似的 k 个节点的特征,保证了提取的特征知识的有效性.

[1]本文尝试用最小化 KL 散度或最大化交叉熵作为替代方案,但最终发现欧式距离表现最好.

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

3.3.3　融合节点表示模块

在这个模块,通过融合 PLP 模块和 SGAT 模块(merge PLP and SGAT,MPS)的节点表示,进一步提高节点特征的表征能力.具体来说,通过学习一个可训练的参数 $a_v \in [0,1]$,进而去平衡 PLP 模块和 SGAT 模块的节点表示:

$$f_{MPS}^{m+1}(v) = \alpha_v \sum_{u \in N_v \cup \{v\}} w_{uv} f_{MPS}^{m}(u) + (1-\alpha_v) h'_v \quad (11)$$

其中 w_{uv} 和 $f_{MPS}^{0}(v)$ 的初始状态和 PLP 模块初始状态一样,即 $f_{MPS}^{0}(v) = f_{PLP}^{0}(v)$.

注意,针对于 *inductive* 任务将使用公式(6)得到参数化置信分数,而对于 *transductive* 任务将为每个节点训练可学习参数来作为置信分数.

3.4　损失函数

之前的知识蒸馏模型,往往只针对单个教师模型进行知识蒸馏,给予学生模型的指导也只停留在单方面信息.因此,本文引入双通道模型,分别基于拓扑结构和特征属性来对学生模型进行指导.

假设学生模型有 m 层,那么基于拓扑结构的知识蒸馏公式为:

$$\mathcal{L}_T = \sum_{v \in V_U} \| f_{\theta^*\text{-}GNN}(v) - f_{MPS}^{m}(v) \|_2 \quad (12)$$

基于特征属性的知识蒸馏公式为:

$$\mathcal{L}_F = \sum_{v \in V_U} \| f_{\beta^*\text{-}GNN}(v) - f_{MPS}^{m}(v) \|_2 \quad (13)$$

其中 $\| \cdot \|_2$ 表示计算欧式距离.

最终的损失函数为:

$$\mathcal{L} = \lambda \mathcal{L}_T + (1-\lambda) \mathcal{L}_F \quad (14)$$

其中,λ 表示超参数.

4　实　验

4.1　数据集

本文使用 5 个公共基准数据集进行实验,数据集的统计信息如表1所示.与之前工作[25-27]一样,本文只考虑最大连通分量,并将连边视为无向的.这是因为针对图数据的研究,主要是探究图结构的作用,即图中各个节点之间关系的相互作用,因此主要考虑图数据的最大连通分量.其次,本文是基于同质图的研究,Cora 数据集、Citerseer 数据集和 Pubmed 数据集的图的连边表示论文之间的相互引用关系,为了统一边的类别,本文将“引用”和“被引用”视为同一种关系,即论文间存在引用关系;A-computer 数据集和 A-Photo 数据集表示商品共同购买的关系,本身就是一个无向图,因此连边是无向的.数据集的详细信息如下:

表 1　数据集统计

Table 1　Statistics of the datasets

数据集	节点数	连边数	特征数	标签数
Cora	2485	5069	143333	7
Citeseer	2110	3668	3703	6
Pubmed	19717	44324	500	3
A-Computers	13381	245778	767	10
A-Photo	7487	119043	745	8

1)Cora 数据集[28].Cora 数据集是一个由机器学习论文组成的基准引文数据集,其中每个节点代表一个文档,具有稀疏的词袋特征向量.连边表示文档之间的引用,标签表示每篇论文的研究领域.

2)Citeseer 数据集[28].Citeseer 数据集是另一个计算机科学出版物引文基准数据集,与 Cora 数据集有类似的属性.

3)Pubmed 数据集[29].Pubmed 数据集也是一个引文数据集,包含了 Pubmed 数据库中与糖尿病相关的文章.节点特征为 TF/IDF 加权词频,标签表示糖尿病类型.

4)A-Computers 数据集和 A-Photo 数据集[26].A-Computers 数据集和 A-Photo 数据集从 Amazon 的共购图中提取,其中节点表示产品,边表示两种产品是否经常同时被购买,特征表示用词袋模型编码的产品评论,标签是预定义的产品类别.

上述的 5 个数据集都是来自现实世界的实际应用,说明在 5 个数据集上的实验具有实际意义.其次,这 5 个数据集充分考虑了图的大小、稀疏性、节点特征维度的差异.如 Pubmed 数据集的节点数是 Cora 数据集的接近 8 倍,A-computers 数据集连边数与节点数的比值是 Citeseer 数据集的接近 11 倍,Cora 数据集特征维度是 Pubmed 数据集的接近 287 倍.因此,在这 5 个数据集上的实验可以验证模型的有效性以及普适性.

4.2　基准方法

本文将与基准方法进行比较,基准方法的简要描述如下.

1)Topology-Teacher,非知识蒸馏方法,只包含教师模型方法,即基于拓扑图的教师模型.

2)Feature-Teacher,非知识蒸馏方法,只包含教师模型方法,即基于特征图的教师模型.

3)CPF-ind/tra[30],知识蒸馏方法,其中 CPF-ind 针对于 *inductive* 任务,CPF-tra 针对于 *transductive* 任务.

4.3　实验设置

教师模型:使用 GAT 模型作为知识蒸馏框架中的教师模型.

学生模型:PLP 模块,具有参数化标签传播机制;SGAT 模块,聚合一阶邻居的图神经网络模型;MPS-ind,用于 *inductive* 任务的 MPS 模型;MPS-tra,用于 *transductive* 任务的 MPS 模型.

其他设置:随机初始化模型参数,采用早停法,早停值设置为 50.如果验证集的分类准确率在 50 次迭代内没有提高,将停止训练.对于超参数优化,采用启发式搜索,其中学生模型网络层数 $m \in \{1,3,5,7,9,11\}$,特征图的 $k \in \{1,2,5,10,20,30,50,100\}$,dropout 比率 $dr \in \{0.2,0.5,0.8\}$,Adam 优化器的学习率 $lr \in \{0.001,0005,0.01\}$,权重衰减 $wd \in \{0.005,0.001,0.01\}$.

4.4　节点分类

4.4.1　模型对比

表 2 和表 3 展示了本文提出的方法在 5 个数据集上的实验结果.其中 Topology-Teacher 和 Feature-Teacher 是非知识蒸馏方法,仅包含教师模型方法,分别为基于拓扑结构的教师模型和基于特征属性的教师模型;CPF-ind/tra[30]是最优的基准方法,是基于教师模型和学生模型的知识蒸馏框架;MPS-ind/tra 是本文提出的一种新型的双通道知识蒸馏框架.通过观察分析有以下结果:

1)知识蒸馏方法与非知识蒸馏方法相比具有更好的分

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

类效果. CPF-ind/tra[30] 和 MPS-ind/tra 均为知识蒸馏方法, Topology-Teacher 和 Feature-Teacher 为非知识蒸馏方法,从表 2 和表 3 可以看出,知识蒸馏方法在 5 个数据集上的分类效果均优于非知识蒸馏方法. 例如,在 Pubmed 数据集上,相比 Topology-Teacher 和 Feature-Teacher 非知识蒸馏方法, CPF-ind[30] 知识蒸馏方法的分类效果分别提升了 4.01%,1.21%; CPF-tra[30] 知识蒸馏方法的分类效果分别提升了 4.39%, 1.58%; MPS-ind 知识蒸馏方法的分类效果分别提升了 6.27%,3.41%; MPS-tra 知识蒸馏方法的分类效果分别提升了 3.41%,3.08%.

表 2 节点分类任务分类准确率

Table 2 Test accuracy on node classification

Model	Cora	Citeseer	Pubmed	A-Computers	A-Photo
Topology-Teacher	0.8389	0.7276	0.7702	0.8107	0.8987
Feature-Teacher	0.7035	0.6807	0.7915	0.7771	0.8754
CPF-ind	**0.8576**	0.7657	0.8011	0.8190	0.9221
MPS-ind(ours)	0.8515	**0.7768**	**0.8185**	**0.8461**	**0.9295**

注:黑体表示所有模型方法中最高的预测准确率

表 3 节点分类任务分类准确率

Table 3 Test accuracy on node classification

Model	Cora	Citeseer	Pubmed	A-Computers	A-Photo
Topology-Teacher	0.8389	0.7276	0.7702	0.8107	0.8987
Feature-Teacher	0.7035	0.6807	0.7915	0.7771	0.8754
CPF-tra	**0.8590**	0.7691	0.8040	0.8148	0.9199
MPS-tra(ours)	0.8529	**0.7746**	**0.8159**	**0.8437**	**0.9358**

注:黑体表示所有模型方法中最高的预测准确率

2)本文提出的 MPS-ind/tra 知识蒸馏方法相比最优的基准知识蒸馏方法 CPF-ind[30],在大部分数据集上具有更好的分类效果. 其中在 *inductive* 任务上,相比最优基准方法 CPF-ind,本文提出的方法 MPS-ind 在 4 个数据集 Citeseer, Pubmed,A-Computers 和 A-Photo 上取得更好的性能:在 Citeseer 数据集上分类准确率提升了 1.45%,在 Pubmed 数据集上提升了 2.17%,在 A-Computers 数据集上提升了 3.31%, 在 A-Photo 数据集上提升了 0.80%. 其中在 *transductive* 任务上,相比最优基准方法 CPF-tra,本文提出的方法 MPS-tra 取得了与 *inductive* 任务上相似的表现,即在 4 个数据集 Citeseer,Pubmed,A-Computers and A-Photo 上取得了比 CPF-tra 更好的性能:在 Citeseer 数据集上分类准确率提升了 0.72%,在 Pubmed 数据集上提升了 1.48%,在 A-Computers 数据集上提升了 3.55%,在 A-Photo 数据集上提升了 1.73%. 在 *inductive* 任务和 *transductive* 任务上,本文提出的方法在 Cora 数据集上的性能略低于最优基准方法,这是因为 Cora 的特征数在十万量级,远远大于其他数据集,在生成特征图时,计算节点特征之间的余弦相似度会引入噪声,因此降低了实验效果.

4.4.2 消融实验

为了进一步分析模型 MPS-ind/tra 中各个模块的作用,通过对不同教师模型和学生模型进行组合的实验,进而验证本文提出的方法的有效性. 因此,本文设置了以下消融实验:

1)MPS-ind/tra(Topology). 该方法是基于拓扑图教师模型和学生模型的单通道知识蒸馏框架,即基于拓扑图的教师模型对学生模型进行指导,学生模型由 PLP 模块和 SGAT 模块组成.

2)MPS-ind/tra(Feature). 该方法是基于特征图教师模型和学生模型的单通道知识蒸馏框架,即基于特征图的教师模型对学生模型进行指导,学生模型由 PLP 模块和 SGAT 模块组成.

3)MPS-ind/tra(Dual-MLP). 该方法是基于拓扑图教师模型、特征图教师模型和学生模型的双通道知识蒸馏框架. 此外,本文参考 CPF-ind/tra[30],用两层 MLP 来作为学生模型中的邻居特征聚合模块,即用两层 MLP 替代 SGAT 模块.

4)MPS-ind/tra(ours). 该方法是基于拓扑图教师模型、特征图教师模型和学生模型的双通道知识蒸馏框架,学生模型由 PLP 模块和 SGAT 模块组成.

表 4 MPS-ind 在 5 个数据集上的消融实验

Table 4 Ablation study of MPS-ind on 5 datasets

Model	Cora	Citeseer	Pubmed	A-Computers	A-Photo
MPS-ind(Topology)	0.8416	0.7657	0.8037	0.8170	0.9201
MPS-ind(Feature)	0.7081	0.7105	0.7967	0.7771	0.8754
MPS-ind(Dual-MLP)	0.8487	0.7702	0.8124	0.8356	0.9241
MPS-ind(ours)	**0.8515**	**0.7768**	**0.8185**	**0.8461**	**0.9295**

注:黑体表示所有模型方法中最高的预测准确

表 5 MPS-tra 在 5 个数据集上的消融实验

Table 5 Ablation study of MPS-tra on 5 datasets

Model	Cora	Citeseer	Pubmed	A-Computers	A-Photo
MPS-tra(Topology)	0.8454	0.7674	0.8040	0.8284	0.9199
MPS-tra(Feature)	0.6791	0.7717	0.7848	0.7895	0.8686
MPS-tra(Dual-MLP)	0.8501	0.7702	0.8082	0.8349	0.9210
MPS-tra(ours)	**0.8529**	**0.7746**	**0.8159**	**0.8437**	**0.9358**

注:黑体表示所有模型方法中最高的预测准确率

实验结果如表 4 和表 5 所示. 从实验结果来看,在大部分数据集上,MPS-ind/tra(Topology)和基于特征图 MPS-ind/tra (Feature)的效果相差不大. 其中,效果相差最大的是在 Cora 数据集上,前面提到过,由于 Cora 数据集的特征数太大,生成特征图时会引入大量噪声,导致实验效果变差. 随后,引入双教师模型的知识蒸馏框架,使得分类效果进一步提升,验证了双教师模型的有效性. 此外,本文参考 CPF-ind/tra[30],对 MPS-ind/tra(ours)的学生模型进行修改,用两层 MLP 替代 SGNN 模块,得到 MPS-ind/tra(Dual-MLP)方法,发现分类效果在 5 个数据集上均变差. 实验结果表明,本文提出的 MPS-ind/tra 各个子模块都是有益于整个模型的.

4.4.3 特征图中 k 不同取值的分析

为了分析 Feature-Teacher 中选取邻居节点个数对模型的效果影响,本文分别选取不同的 k 值($k \in \{1,2,5,10,20,30,50,100\}$)在 5 个数据集上进行实验. 从图 2 的实验结果可以看出,刚开始随着 k 值的增大,模型分类效果逐渐提高,但是随着 k 值的不断增大,模型分类效果降低了. 这是因为,刚开始引入邻居信息会加强节点表示,然而随着引入的邻居信息越来越多,会导致信息冗杂,噪声越来越多,导致模型效果变差.

4.4.4 标签样本数量的分析

为了进一步证明模型的有效性,本文将在训练集中选取不同数量的有标签样本进行实验. 如图 3 和图 4 所示,以 Citeseer 数据集为例,本文从各个类别中分别选取 n 个节点(n

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

∈{5,10,20})作为有标签节点,依次进行实验. 从实验结果可以看出,在*inductive*任务和*transductive*任务上,随着标签

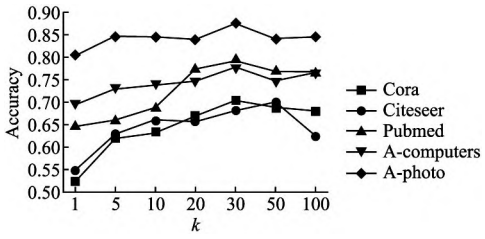

图 2　参数 k 的分析

Fig. 2　Analysis of parameter k

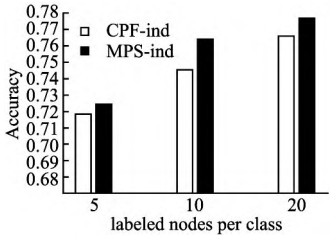

图 3　*inductive* 任务中标签样本数量分析

Fig. 3　Analysis on the number of labeled samples in the *inductive* task

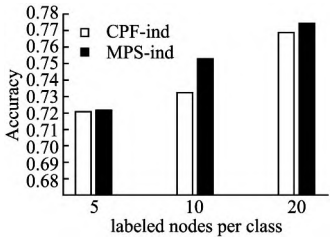

图 4　*transductive* 任务中标签样本数量分析

Fig. 4　Analysis on the number of labeled samples in the *transductive* task

节点数量增多,两种方法的分类准确率逐渐升高. 其中,本文提出的 MPS-ind/tra 方法在不同数量的标签节点上的分类效果皆优于最优的基准方法 CPF-ind/tra[30].

4.4.5　学生模型中网络层数的分析

在本小节中,将研究 MPS-ind/tra 体系结构中的一个关键超参数的影响,即学生模型网络层的数量 m. 实际上, GAT

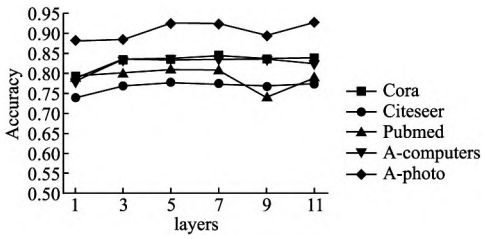

图 5　*inductive* 任务中参数 m 的分析

Fig. 5　Analysis of parameter m in the *inductive* task

等流行的 GNN 模型对网络层数非常敏感,过多的网络层会导致过度平滑问题,严重的还将损害模型性能. 为此,本文在 5 个数据集上进行实验,进一步分析该超参数.

图 5 和图 6 分别表示具有不同网络层数 $m \in \{1,3,5,7,9,11\}$ 学生模型的 MPS-ind 和 MPS-tra 的分类效果. 可以看出,在大多数数据集上具有不同网络层数m的分类效果差距

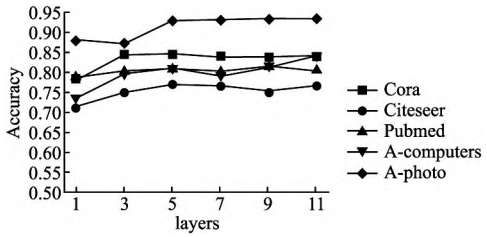

图 6　*transductive* 任务中参数 m 的分析

Fig. 6　Analysis of parameter m in the *transductive* task

相对较小. 对于每个数据集,本文计算其相应最佳和最差表现准确率之间的差距,MPS-ind 和 MPS-tra 的最大差距在 1% 左右,大部分数据集的最大差距不到 0.5%. 可以看出,本文提出的模型分类效果较为稳定,不会随着网络层数的变化,性能骤降.

5　总结与展望

本文提出了一个基于拓扑结构和特征属性的双通道知识蒸馏框架,即学习得到两个具有不同层面知识的预训练教师模型,来指导学生模型. 其中学生模型由两种预测机制组合而成:参数化标签传播模块和邻居特征聚合模块,根据同质现象,分别保留基于结构和基于特征的先验知识. 两种教师模型和学生模型组成知识蒸馏框架,使得学生模型能够利用先验知识,并且汲取两种教师模型的知识,从而达到更好的分类效果. 在 5 个真实数据集上的实验结果表明,本文提出的知识蒸馏框架能够使得学生模型很好地汲取教师模型的知识,且在 4 个真实数据集上的分类效果优于最优的基准方法.

在未来的工作中,将探讨在双通道模型的基础上,改进学生模型. 例如:构建多通道多学生的知识蒸馏框架,在教师模型知识多样性的基础上,打造学生模型知识的多样性. 另一个方向是实现教师模型和学生模型在训练中不断交互学习以获得更好的效果,进一步完善本文提出的模型框架.

References:

[1] Tang Jian, Qu Meng, Mei Qiao-zhu. Pte: predictive text embedding through large-scale heterogeneous text networks[C]//Proc of the 21st ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining, New York, USA, 2015: 1165-1174.

[2] Lei Tang, Huan Liu. Scalable learning of collective behavior based on sparse social dimensions[C]//Proceedings of the 18th ACM Conference on Information, New York, USA, 2009: 1107-1116.

[3] Charu C Aggarwal, Zhai Cheng-xiang. Mining text data[M]. New York, USA, 2012: 163-222.

[4] Miller McPherson, Lynn Smith-lovin, James M Cook. Birds of a feather: homophily in social networks[J]. Annual Review of Sociology, 2001, 27(1): 415-444.

[5] Martin Szummer, Tommi Jaakkola. Partially labeled classification with Markov random walks[C]//Advances in Neural Information Processing Systems, Cambridge, UK: MIT Press, 2002: 945-952.

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

[6] Zhu Xiao-jin, Zoubin Ghahramani, Lafferty J. Semi-supervised learning using Gaussian fields and harmonic functions[C]//International Conference on Machine Learning, New York, USA: ACM, 2003: 912-919.

[7] Thorsten Joachims. Transductive learning via spectral graph partitioning[C]//International Conference on Machine Learning, New York, USA: ACM, 2003: 290-297.

[8] Zhou Deng-yong, Olivier Bousquet, Thomas N Lal, et al. Learning with local and global consistency[C]//Advances in Neural Information Processing Systems, Cambridge, UK: MIT Press, 2003: 321-328.

[9] Krizhevsky A, Ilya Sutskever, Geoffrey E Hinton. ImageNet classification with deep convolutional neural networks[C]//Advances in Neural Information Processing Systems, Cambridge, UK: MIT Press, 2012: 1106-1114.

[10] Keyulu Xu, Weihua Hu, Jure Leskovec, et al. How powerful are graph neural networks? [C]//International Conference on Learning Representations, New Orleans, USA: OpenReview. net, 2019.

[11] Yang Yi-ding, Qiu Jia-yan, Song Ming-li, et al. Distilling knowledge from graph convolutional networks[C]//IEEE Conference on Computer Vision and Pattern Recognition, New York, USA, 2020: 7072-7081.

[12] Petar Velickovic, Guillem Cucurull, Casanova A. Graph attention networks[C]//International Conference on Learning Representations, New Orleans, USA: OpenReview. net, 2017.

[13] Imai Li, Wu Xiao-ming, Liu Han, et al. Label efficient semi-supervised learning via graph filtering[C]//IEEE Conference on Computer Vision and Pattern Recognition, New York, USA, 2019: 9574-9583.

[14] Shi Yun-sheng, Huang Z, Wang Wen-jin, et al. Masked label prediction: unified massage passing model for semi-supervised classification[J]. arXiv preprint, 2020, arXiv: 2009.03509.

[15] Wang Hong-wei, Leskovec J. Unifying graph convolutional neural networks and label propagation[J]. arXiv preprint, 2020, arXiv: 2002.06755.

[16] Kipf, Thomas, Welling M. Semi-supervised classification with graph convolutional networks[C]//International Conference on Learning Representations, New Orleans, USA: OpenReview. net, 2017.

[17] Jing Zhuang-wei, Guan Hai-yan, Peng Dai-feng, et al. Survey of research in image segmentation based on deep neural network[J]. Journal of Computer Engineering, 2020, 46(10): 1-17.

[18] Geoffrey E Hinton, Oriol Vinyals, Dean J. Distilling the knowledge in a neural network[J]. arXiv preprint, 2015, arXiv: 1503.02531.

[19] Furlanello, Tommaso, Zachary Chase Lipton, et al. Born again neural networks[C]//International Conference on Machine Learning, New York, USA: ACM, 2018: 1602-1611.

[20] Zhang Wen-tao, Miao Xu-peng, Shao Ying-xia, et al. Reliable data distillation on graph convolutional network[C]//SIGMOD International Conference on Management of Data, Portland, USA: ACM, 2020: 1399-1414.

[21] Qu Meng, Yoshua Bengio, Jian Tang. GMNN: graph Markov neural networks[C]//International Conference on Machine Learning, New York, USA: ACM, 2019: 5241-5250.

[22] Hinton, Geoffrey E, Oriol Vinyals, et al. Distilling the knowledge in a neural network[J]. arXiv preprint, 2015, arXiv: 1503.025311.

[23] Chen Bin, Li Jin-long. Attention-based network representation learning model using multi-neighboring information[J]. Journal of Chinese Computer Systems, 2021, 42(4): 761-765.

[24] Zhu Xiao-jin, Zoubin Ghahramani. Learning from labeled and unlabeled data with label propagation[R]. Pittsburgh, USA: Carnegie Mellon University, Technical Report: CMU-CALD-02-107, 2002, doi: 10.1109/ijcnn.2002.1007592.

[25] Prithviraj Sen, Galileo Namata, Mustafa Bilgic, et al. Tina eliassi-rad: collective classification in network data[J]. AI Magazine, 2008, 29(3): 93, doi: 10.1609/aimag.v29i3.2157.

[26] Oleksandr Shchur, Maximilian Mumme, Aleksandar Bojchevski, et al. Pitfalls of graph neural network evaluation[J]. arXiv preprint, 2018, arXiv: 1811.05868.

[27] Johannes Klicpera, Stefan Weissenberger, Stephan Günnemann. Diffusion improves graph learning[C]//Advances in Neural Information Processing Systems, Cambridge, USA: MIT Press, 2019: 13333-13345.

[28] Ke Sun, Zhu Zhan-xin, Lin Zhou-chen. Multi-stage self-supervised learning for graph convolutional networks[C]//AAAI Conference on Artificial Intelligence, Palo Alto, USA, 2020: 5892-5899.

[29] Galileo Namata, Ben London, Lise Getoor, et al. Query-driven active surveying for collective classification[C]//10th International Workshop on Mining and Learning with Graphs, 2012.

[30] Cheng Yang, Liu Jia-wei, Shi C. Extract the knowledge of graph neural networks and go beyond it: an effective knowledge distillation framework[C]//International World Wide Web Conferences, New York, USA, 2021: 1227-1237.

附中文参考文献：

[17] 景庄伟，管海燕，彭代锋，等. 基于深度神经网络的图像语义分割研究综述[J]. 计算机工程，2020，46(10)：1-17.

[23] 陈 斌，李金龙. 基于注意力机制融合多阶邻域信息的网络表示学习模型[J]. 小型微型计算机系统，2021，42(4)：761-765.

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net