arXiv:2301.01150v2 [cs.LG] 4 Jan 2023

RELIANT：面向图神经网络的公平知识蒸馏

Yushun Dong* Binchi Zhang* Yiling Yuan† Na Zou‡ Qi Wang§ Jundong Li*

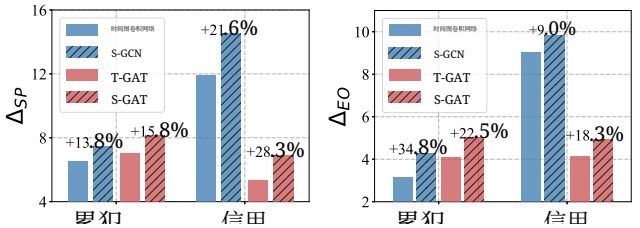
摘要

图神经网络（GNN）在各种图学习任务中展现出了令人满意的性能。为获得更强的拟合能力，大多数图神经网络都包含大量参数，因此计算成本高昂。这使得它们难以部署在计算资源有限的边缘设备上，例如手机和可穿戴智能设备。知识蒸馏（KD）是压缩图神经网络的一种常见方法：通过让轻量级模型（即学生模型）模仿计算成本高昂的图神经网络（即教师图神经网络模型）的行为，达到压缩目的。然而，大多数现有的基于图神经网络的知识蒸馏方法都没有考虑公平性。因此，学生模型通常会继承甚至放大教师图神经网络中的偏差。为解决这一问题，我们初步探索了面向图神经网络的公平知识蒸馏。具体而言，我们首先针对基于图神经网络的师生框架，提出了一个全新的公平知识蒸馏问题。随后，我们提出了一个原理明确的框架 RELIANT，以减轻学生模型表现出的偏差。值得注意的是，RELIANT 的设计与任何特定的教师或学生模型结构相互独立，因此能够轻松适配各种基于图神经网络的知识蒸馏框架。我们在多个真实世界数据集上进行了大量实验，结果证实，RELIANT 能够在保持较高预测效用的同时，实现偏差更小的图神经网络知识蒸馏。开源代码见 https://github.com/yushundong/RELIANT。

关键词： 图神经网络，算法公平性，知识蒸馏

1 引言

近年来，图神经网络（GNN）在大量现实应用中展现出令人满意的性能，例如医疗诊断 [27] 和信用风险评分 [30]，等。在实践中，图神经网络的深度和参数数量在很大程度上



(a) CPF 上的 Δ_{SP} 偏差。 (b) Δ_{EO} on AKD 下的偏差。

图 1：基于两种具有代表性的图神经网络知识蒸馏框架（CPF 和 GraphAKD），对比教师模型与学生模型表现出的偏差。“T”和“S”分别代表教师模型和学生模型。图神经网络的名称标示了相应的教师模型。

决定了其表达能力 [16]，，并直接影响其在各种图学习任务中的表现 [2]。通常，更深的 GNN 层使模型能够捕获距离任意节点多个跳数的信息 [21]，，而更多的可学习参数则使图神经网络能够拟合更复杂的底层数据模式 [2]。然而，在大多数情况下，深层结构或大量参数不可避免地会降低图神经网络的推理效率。效率下降自然会使这些图神经网络无法部署在计算资源有限的边缘设备（例如手机）上 [16, 19]。

鉴于上述问题，有必要压缩这些计算成本高昂的图神经网络，以便部署在边缘设备上。知识蒸馏（KD）是压缩图神经网络的常见方法，同时仍能保持相近的预测性能 [34, 16, 19]。知识蒸馏的基本思想是，让轻量级学生模型（作为压缩后的图神经网络）模仿教师模型（通常是计算成本高昂的图神经网络）的行为（例如输出 logit）。然而，现有的大多数知识蒸馏方法并未考虑不同人口统计学子群体之间的公平性，优化后的学生模型往往会保留甚至加剧教师图神经网络表现出的偏差。因此，当压缩后的模型部署于现实应用场景时，可能会歧视特定人群。在此，我们基于两个具有代表性的图神经网络知识蒸馏框架，即 CPF [34]和 GraphAKD [16]，进行了初步分析。具体而言，我们测量了其表现出的

*弗吉尼亚大学，电子邮箱：{yd6eb, epb6gw, jundong}@virginia.edu
†北京邮电大学，电子邮箱：yuanyiling@bupt.edu.cn
‡得克萨斯农工大学，电子邮箱：nzou1@tamu.edu§东北大学，电子邮箱：q.wang@northeastern.edu

版权所有 © 2023年 SIAM 未经授权，禁止复制本文

偏差，研究对象是在两个真实世界数据集上广泛研究的节点分类任务。此处，Recidivism 是由被告组成的网络 [18, 1]，而 Credit 是银行客户之间构成的网络 [35, 1]。我们采用两种传统指标，即 Δ_{SP}（衡量统计均等 [10]下的偏差程度）和 Δ_{EO}（衡量机会均等 [15]下的偏差程度），来测量图神经网络预测表现出的偏差。图 1 展示了教师模型与学生模型表现出的偏差对比。实证结果表明，与教师图神经网络模型相比，学生模型往往会产生偏差更大的结果，这可能归因于训练过程中教师图神经网络提供的偏差指导。值得注意的是，在大多数情况下，直接重新训练教师图神经网络以去偏并不可行，因为重新训练参数量庞大的教师图神经网络计算成本高昂。因此，消除学生模型的偏差已迫在眉睫。

尽管有必要缓解学生模型的偏差，但现有研究仍然十分有限。本文旨在迈出初步一步，开发一种能够轻松适配各种现有基于图神经网络的知识蒸馏方法的去偏框架。然而，这项任务并非易事，主要面临以下三项挑战：**(1)迈向公平知识的差距：**对于大多数旨在压缩图神经网络的知识蒸馏框架，教师图神经网络模型通常是学生模型训练过程中唯一的监督信号来源。因此，如果教师图神经网络存在任何偏差，这种有偏知识往往会被学生模型继承。因此，如何在教师图神经网络提供有偏监督的情况下学习公平的学生模型，是我们的首要挑战。**(2) 迈向端到端学习的差距：**现有知识蒸馏模型的一项关键优势是端到端学习范式，这使蒸馏知识能够针对特定下游任务进行调整。在这种端到端学习过程中，广泛采用高效的基于梯度的优化技术。然而，常用的公平性概念（例如统计均等和机会均等）是根据预测标签定义的。因此，相应的偏差度量相对于学生模型参数自然不可微。如何开发适用于端到端学习范式中基于梯度的优化技术的去偏框架，是我们的第二项挑战。**(3)迈向泛化的差距：**人们已提出各种知识蒸馏模型，以压缩图神经网络并满足不同的应用场景。事实上，大多数知识蒸馏模型都是基于特定的学生模型设计开发的。如何开发一种与学生模型无关、且能轻松适配不同知识蒸馏模型的框架，是我们的第三项挑战。

为应对上述挑战，本文提出一种名为 RELIANT（图神经网络公平知识蒸馏）的新框架，以减轻学生模型学到的偏差。具体而言，我们首先提出一个全新的研究问题：面向基于图神经网络的教师—学生框架的公平知识蒸馏。为应对第一个挑战，我们为学生模型引入一种可学习的偏差代理。这样一来，尽管知识（来自教师图神经网络）存在偏差，通过对该代理进行适当操作，学生模型仍能做出偏见较少的预测。为应对第二个挑战，我们提出近似学生模型偏差程度的方法，该近似过程关于学生模型参数可微。由此得以保留高效的端到端学习范式，并继续适用基于梯度的优化技术。为应对第三个挑战，我们以与学生模型无关的方式设计了所提出的 RELIANT 框架。换言之，学生模型的去偏不依赖于针对学生模型结构定制的特定设计。因此，RELIANT 可轻松适配不同的基于图神经网络的知识蒸馏方法。本文的主要贡献总结如下。

- **问题形式化**。我们对基于图神经网络的教师—学生框架中的公平知识蒸馏这一新颖研究问题进行了形式化，并开展了初步研究。
- **算法设计**。我们提出了一个名为 RELIANT 的原则性框架，在知识蒸馏过程中学习学生模型的偏差代理。RELIANT 通过在推理过程中操纵该代理，实现了与学生模型无关的去偏。
- **实验评估**。我们在多个真实世界数据集上进行了全面实验，以验证所提出的 RELIANT 框架在学习偏差较小的学生模型方面的有效性。

2 问题定义

符号。我们分别用粗体大写字母（例如，$\mathbf{X}$）、粗体小写字母（例如，$\mathbf{x}$）和常规小写字母（例如，x）表示矩阵、向量和标量。对于任意矩阵，例如 $\mathbf{X}$，我们用 $\mathbf{X}_{i,\ j}$ 表示第 i 行、第 j 列的元素。

预备知识。我们用 $\mathcal{G} = \{\mathcal{V}, \mathcal{E}, \mathcal{X}\}$ 表示一个属性网络（图）。其中，$\mathcal{V} = \{v_1, ..., v_n\}$ 是节点集合，$\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ 是边集合，$\mathcal{X} = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$ $(\mathbf{x}_i \in \mathbb{R}^d,\ 1 \leq i \leq n)$ 是节点属性向量集合。我们用 $\mathbf{A} \in \{0,1\}^{n \times n}$ 表示图的邻接矩阵。如果存在一条从第 i 个节点指向第 j 个节点的边，则 $\mathbf{A}_{i,\ j} = 1$；否则 $\mathbf{A}_{i,j} = 0$。此外，我们将知识蒸馏框架中的预训练教师图神经网络模型表示为 $f_{\boldsymbol{\theta}}$，其参数由 $\boldsymbol{\theta}$ 表示。其中，$\boldsymbol{\theta}$ 表示预训练教师模型经过优化的 $\boldsymbol{\theta}$。类似地，

版权 © 2023年 SIAM。未经授权，禁止转载本文。

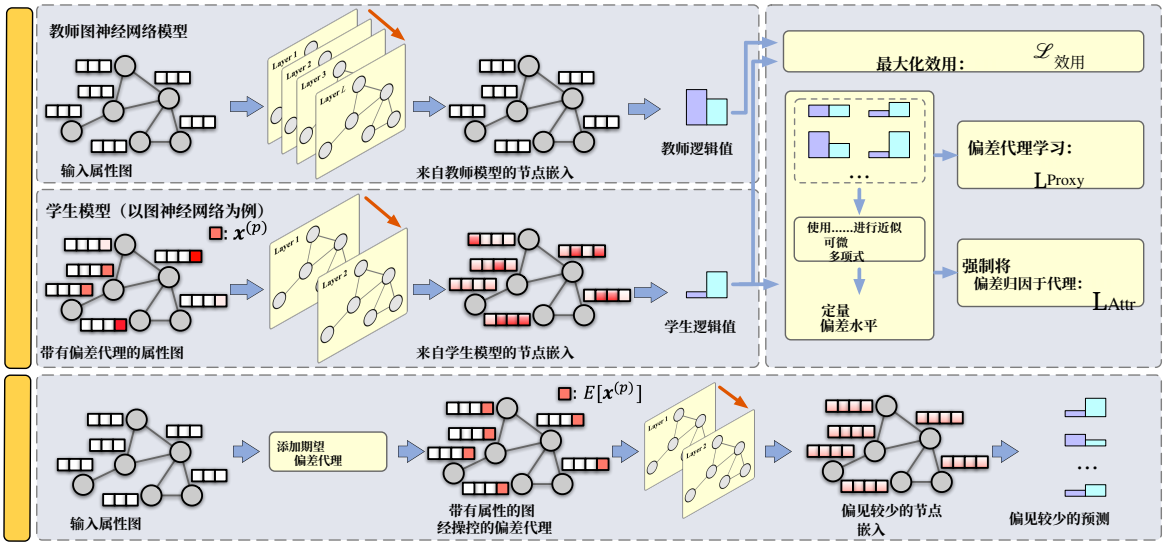


图 2：所提出的 RELIANT 框架概览，包括训练和推理阶段。

我们将学生模型表示为 g_ϕ，其参数由ϕ表示。我们将学生模型训练后优化得到的 ϕ 表示为 ϕ。不失一般性，我们考虑研究最广泛的下游任务——节点分类。对于教师模型 $f_{\hat{\theta}}(v)$，我们将结果 logits 集合（即对应于每个节点的连续输出向量）表示为 $\hat{\mathcal{Y}}^{(t)} = \{\hat{\mathbf{y}}_1^{(t)}, \hat{\mathbf{y}}_2^{(t)}, ..., \hat{\mathbf{y}}_n^{(t)}\}$，其中 $\hat{\mathbf{y}}_i^{(t)} \in \mathbb{R}^c$。这里，$c$ 是节点类别总数。相应地，我们将学生模型 $g_{\boldsymbol{\phi}}(v)$ 的结果 logits 集合表示为 $\hat{\mathcal{Y}}^{(s)} = \{\hat{\mathbf{y}}_1^{(s)}, \hat{\mathbf{y}}_2^{(s)}, ..., \hat{\mathbf{y}}_n^{(s)}\}$。对于任意节点 v_i，学生模型给出的预测标签（记作第 i 个节点的 $\hat{Y}_i^{(s)}$）由 $\hat{\mathbf{y}}_i^{(s)}$ 中所有 c 个维度上的最大值决定。

基于上述定义，我们将面向基于 *GNN* 的教师-学生框架的公平知识蒸馏问题表述如下。

问题 1. 面向基于 ***GNN*** 的教师-学生框架的公平知识蒸馏。给定一个属性网络 $\mathcal{G}$ 以及一个基于图神经网络的教师 - 学生框架，其中包括一个经过训练的教师图神经网络 $f_{\hat{\theta}}$ 和一个待训练的学生模型 g_ϕ，我们的目标是针对节点分类任务，获得一个比 $f_{\hat{\theta}}$ 更加公平且预测效用相近的学生模型。

3 方法

本节首先概述所提出的框架 RELIANT，随后介绍目标函数的构建和优化策略。

3.1 RELIANT 的工作流程 我们首先介绍所提出框架 RELIANT 的工作流程。总体而言，我们介绍所提出框架 RELIANT 涉及的三项主要功能，即最大化效用、偏差代理学习，以及强制将偏差归因于代理。图 2 展示了 RELIANT 的概览。具体而言，为应对第一个挑战（公平知识方面的差距），我们提出首先将偏差代理作为额外的输入属性进行学习，使学生模型能够考虑到（训练节点上）表现出的偏差；然后在测试时，使用操控后的伪代理排除代理信息。为应对第二个挑战（端到端学习方面的差距），我们以可微分的方式（相对于学生模型的参数）制定去偏目标。为应对第三个挑战（泛化方面的差距），我们以与学生模型无关的方式实现去偏。换言之，所提出的 RELIANT 框架无需依赖任何特定的学生模型结构即可实现去偏。下面将详细阐述。

最大化效用。 一般而言，现有的基于图神经网络的知识蒸馏框架将计算成本较高的图神经网络作为教师模型，其目标是学习一个计算成本较低、但预测效用相近的学生模型（例如节点分类任务中的准确率）。为保持教师模型的效用，需要利用教师模型的知识作为学生模型训练的监督信号。具体而言，一种常见做法是将教师模型的输出分类逻辑值用作监督信号，本文也以此为例。具体来说，我们最小化学生模型与教师模型输出的 logit 之间的距离。我们将优化目标形式化为

$$\min_{\boldsymbol{\phi}} \sum_{v_i \in \mathcal{V}} \gamma_{\mathrm{d}} \left(\hat{\mathbf{y}}_i^{(t)}, \hat{\mathbf{y}}_i^{(s)} \right), \tag{3.1}$$

其中 $\hat{\mathbf{y}}_i^{(s)}$ 和 $\hat{\mathbf{y}}_i^{(t)}$ 分别是学生

版权 © 2023年 SIAM。未经授权，禁止复制本文。

模型 $g_{\boldsymbol{\phi}}(v_i)$ 和教师模型 $f_{\boldsymbol{\theta}}(v_i)$ 的输出 logit。函数 $\gamma_{\mathrm{d}}(.,.)$ 用于衡量两个 logit 向量之间的距离。可以采用不同的方法来衡量距离，例如余弦距离和欧几里得距离。因此，为了最大化预测效用，我们最小化目标函数

$$\mathscr{L}_{\mathrm{Utility}}(\boldsymbol{\phi}) = \sum_{v_i \in \mathcal{V}} \gamma_{\mathrm{d}} \left(\hat{\mathbf{y}}_i^{(t)}, \hat{\mathbf{y}}_i^{(s)} \right). \tag{3.2}$$

偏差代理学习。值得注意的是，即使从输入数据中移除敏感属性，学生模型在其预测中仍可能表现出偏差。主要原因在于，这些敏感属性与非敏感属性之间可能存在依赖关系。此外，有关敏感属性的信息也可能被编码在输入网络结构 [6] 中。因此，很难阻止学生模型利用敏感属性的信息。为处理此类问题，我们提出学习偏差代理 $\mathbf{x}_i^{(p)}$，将其作为每个节点的额外输入属性 v_i。这里的基本思路是：如果有关偏差的大量信息来自所学习的代理，而不是来自非敏感属性或网络结构中编码的信息，那么在测试期间不使用此类代理中的信息，就能够实现偏见较少的预测。因此，这种偏差代理应尽可能反映学生模型表现出的偏差。换言之，表现出的偏差应主要归因于偏差代理，而非网络数据中编码的敏感信息。更具体地说，为了促使偏差代理对学生模型中表现出的偏差作出贡献，我们提出在将这些代理与其他属性及网络结构一同作为学生模型的输入时，最大化表现出的偏差。我们将目标形式化表述为

$$\max_{\mathbf{X}^{(p)}} J_{\mathrm{Bias}}(\{g_{\boldsymbol{\phi}}(\gamma(v_i, \mathbf{X}^{(p)})) : i \in \mathcal{V}\}), \tag{3.3}$$

其中，$\gamma(.,.)$ 是一个以节点和偏差代理矩阵为输入，并输出带有拼接节点属性向量的节点 $[\mathbf{x}_i, \mathbf{x}_i^{(p)}]$ 的函数。这里，$\mathbf{x}_i^{(p)}$ 是 $\mathbf{X}^{(p)}$ 的第 i 行。$J_{\mathrm{Bias}}(.)$ 是一个以学生模型输出的 logit 集合为输入，并输出表示表现出的偏差程度的值的函数。然而，在统计均等和机会均等等传统公平性概念下，该计算不可微。这里，我们提出利用正交多项式（例如勒让德多项式 [4]）来近似传统公平性概念下的偏差程度，这些多项式相对于输出 logit 可微。这样便使其相对于 J_{Bias} 可微。

可学习参数 $\boldsymbol{\phi}$。相应地，我们正式给出实现上述目标的目标函数：

$$\mathscr{L}_{\mathrm{Proxy}}(\mathbf{X}^{(p)}) = -J_{\mathrm{Bias}}(\{g_{\boldsymbol{\phi}}(\gamma(v_i, \mathbf{X}^{(p)})) : i \in \mathcal{V}\}). \tag{3.4}$$

强制将偏差归因于代理。仅实现式（3.3）还不足以确保代理在很大程度上解释学生模型表现出的偏差。这是因为原始节点属性仍可能导致表现出的偏差。具体而言，我们将 $P(\hat{Y}(s))$ 表示为：对于任意特定节点，学生模型给出某一分类预测的概率。我们假设存在潜在的无偏节点属性 $\mathbf{X}^{(u)}$和有偏节点属性 $\mathbf{X}^{(b)}$。当式（3.3）实现时，显然 $P(\hat{Y}(s)|\mathbf{X}^{(u)}, \mathbf{X}^{(b)}, \mathbf{X}^{(p)})$，即 $P(\hat{Y}(s)|\mathbf{X}, \mathbf{X}^{(p)})$，是有偏的。然而，$\mathbf{X}^{(b)}$ 和 $\mathbf{X}^{(p)}$ 都可能是表现出的偏差的来源。值得注意的是，我们的目标是学习代理 $\mathbf{X}^{(p)}$，以尽可能多地解释表现出的偏差。因此，为确保代理有效，必须确保将表现出的偏差归因于来自 $\mathbf{X}^{(p)}$ 的偏差信息，而非 $\mathbf{X}^{(b)}$。换言之，我们需要强制使 $P(\hat{Y}(s)|\mathbf{X}^{(u)}, \mathbf{X}^{(b)})$ 的偏差减小，从而确保 $\mathbf{X}^{(p)}$ 尽可能多地解释表现出的偏差。然而，考虑到学生模型的输入维度数量固定，$P(\hat{Y}(s)|\mathbf{X}^{(u)}, \mathbf{X}^{(b)})$ 是难以处理的。因此，我们提出一种替代方法。分别用 $\mathbf{x}^{(p)}$ 和 s 表示学得的偏差代理以及任意节点的潜在敏感属性向量。我们提出使用向量 $\mathbb{E}[\mathbf{x}^{(p)}]$ 替换 $\mathbf{X}^{(p)}$ 中的每一行，作为操控后的伪代理 $\tilde{\mathbf{X}}^{(p)}$。这样，$\tilde{\mathbf{X}}^{(p)}$ 中的各行便独立于 s，即清除了 $\mathbf{X}^{(p)}$中编码的敏感属性信息。为强制将偏差归因于代理 $\mathbf{X}^{(p)}$，移除有关 $\mathbf{X}^{(p)}$ 的信息后，预测应尽可能公平。因此，我们将最后一个优化目标表述为

$$\min_{\boldsymbol{\phi}} J_{\mathrm{Bias}}(\{g_{\boldsymbol{\phi}}(\tilde{\gamma}(v_i, \tilde{\mathbf{X}}^{(p)})) : i \in \mathcal{V}\}), \tag{3.5}$$

其中 $\tilde{\gamma}(.,.)$ 是一个函数，它将节点和矩阵 $\tilde{\mathbf{X}}^{(p)}$ 作为输入，并返回该输入节点及拼接后的节点属性向量 $[\mathbf{x}_i, \tilde{\mathbf{x}}_i^{(p)}]$。这里，$\tilde{\mathbf{x}}_i^{(p)}$ 是矩阵 $\tilde{\mathbf{X}}^{(p)}$的第 i 行。我们将相应的目标函数正式表示为

$$\mathscr{L}_{\mathrm{Attr}}(\boldsymbol{\phi}) = J_{\mathrm{Bias}}(\{g_{\boldsymbol{\phi}}(\tilde{\gamma}(v_i, \tilde{\mathbf{X}}^{(p)})) : i \in \mathcal{V}\}). \tag{3.6}$$

使用学生模型进行推理。为了实现偏差更小的推理，理想情况下应使用 $P(\hat{Y}(s)|\mathbf{X}^{(u)})$ 进行预测。然而，很难显式地从 $\mathbf{X}$中提取$\mathbf{X}^{(u)}$。相反，我们认为

版权所有 © 2023年 SIAM。未经授权，禁止转载本文

$P(\hat{Y}_{(s)}|\mathbf{X}^{(u)}, \mathbf{X}^{(b)}, \tilde{\mathbf{X}}^{(p)})$ 与 $P(\hat{Y}_{(s)}|\mathbf{X}^{(u)})$ 相比，表现出相近程度的偏差。这是因为：(1) 在应用公式（3.3）和公式（3.5）后，$P(\hat{Y}_{(s)}|\mathbf{X}^{(u)}, \mathbf{X}^{(b)}, \tilde{\mathbf{X}}^{(p)})$ 所表现出的偏差在很大程度上不依赖于 $\mathbf{X}^{(b)}$；(2) $\tilde{\mathbf{X}}^{(p)}$ 中未编码有关敏感属性的更多信息（如上所述）。因此，我们建议在推理阶段利用 $g_{\phi}(\tilde{\gamma}(v_i, \tilde{\mathbf{X}}^{(p)}))$，以实现对节点 v_i 偏差更小的预测。

4 优化目标与策略

本节将介绍 RELIANT 的优化目标，随后说明其训练策略。

优化目标。基于上述讨论，我们在此总结所提出的 RELIANT 的优化目标。首先，为了优化参数 ϕ，我们将统一的目标函数正式表述为

$$\mathscr{L}_{\phi} = \mathscr{L}_{\text{Utility}}(\phi) + \lambda \cdot \mathscr{L}_{\text{Attr}}(\phi). \tag{4.7}$$

这里，λ 是用于控制学生模型去偏效果的超参数。其次，为了优化可学习偏差代理 $\mathbf{X}^{(p)}$，我们将目标函数正式表述为

$$\mathscr{L}_{\mathbf{X}^{(\mathrm{P})}} = \mathscr{L}_{\text{Proxy}}(\mathbf{X}^{(p)}). \tag{4.8}$$

优化策略。为了训练所提出的框架 RELIANT，我们采用交替方式优化参数 ϕ 和可学习偏差代理 $\mathbf{X}^{(p)}$。我们在算法 1 中给出了 RELIANT 的算法流程。

算法 1 面向图神经网络的公平知识蒸馏

输入： $\mathcal{G}$：图数据；$f_{\hat{\theta}}$：训练完成的教师图神经网络模型；g_{ϕ}：学生模型；
输出： $g_{\hat{\phi}}$：优化后的学生模型；$\mathbf{X}^{(p)}$：偏差的代理矩阵；

```
随机初始化X^(p);
当停止训练的条件未满足时执行
    根据公式（4.7）计算 L_φ；
    用 ∂L_φ/∂φ 更新 φ；
    根据公式（4.8）计算 L_X^(p)；
    更新X^(p)，使用 ∂L_X^(p)/∂X^(p)；
结束 while 循环
返回 g_φ̂ 和 X^(p);
```

5 实验评估

本节首先介绍下游学习任务和所采用的真实世界数据集，接着介绍骨干模型、基线方法和评估指标。然后，我们介绍模型的实现细节。最后，讨论所提出的 RELIANT 的评估结果。具体而言，我们旨在通过实验回答以下研究问题：

实验：**研究问题1：** 与其他基线方法相比，RELIANT 在平衡学生模型的效用与公平性方面表现如何？ **RQ2：** RELIANT 的各个组成部分在多大程度上有助于整体去偏性能？ **RQ3：** 超参数 λ 的选择将如何影响 RELIANT 的性能?

5.1 实验设置 这里介绍我们实验评估所采用的设置。

下游任务与真实世界数据集。 本文采用广泛研究的节点分类作为下游任务。实验评估采用四个真实世界数据集，包括两个广泛使用的网络数据集（累犯 [18, 1] 和信用违约者 [35, 1]），以及两个基于真实世界数据新构建的数据集（DBLP 和 DBLP-L）。在累犯数据集中，节点表示获准保释的被告，边表示根据被告过往犯罪记录计算出的被告之间的联系。这里的敏感特征是种族，我们旨在判断某个被告保释后是否不太可能犯罪。在信用违约者数据集中，节点表示信用卡用户，边表示这些用户之间的联系。这里我们将用户的年龄段视为敏感特征，并旨在预测信用卡还款的未来违约情况。此外，我们还基于 AMiner 网络 [29], 构建了两个合著网络，即 DBLP 和 DBLP-L。AMiner 网络是从计算机科学文献目录收集的合著网络。具体而言，我们首先筛除 AMiner 网络中信息不完整的节点。随后，我们采用两种不同的方法，从筛选后的数据集中采样得到连通网络：DBLP 是通过随机游走采样得到的子图，而 DBLP-L 是筛选后的 AMiner 网络的最大连通分量。在这两个数据集中，节点代表不同领域的研究人员，边表示研究人员之间的共同署名关系。敏感属性是每位研究人员所属机构所在的大洲，我们旨在预测每位研究人员的主要研究领域。四个数据集的详细统计信息见表 1。**知识蒸馏框架骨干与教师图神经网络**。 为评估 RELIANT 对不同基于图神经网络的知识蒸馏骨干的泛化能力，我们采用了两个专为压缩图神经网络设计的代表性知识蒸馏框架，即 CPF [34] 和 GraphAKD [16]。通常，CPF 最小化教师模型与学生模型输出的 logit 之间的分布距离，为学生模型提供监督信息；而 GraphAKD 则利用对抗训练为学生模型实现知识蒸馏。CPF 和 GraphAKD 的学生模型是 PLP

[34] 和 SGC [32], resp地 y。对于每种知识蒸馏框架中，我们采用两种类型的图神经网络（包括 GCN [21] 和 GraphSAGE [14]）作为教师图神经网络。

版权所有 © 2023年 SIAM。未经授权，不得复制本文。

表 1：用于实验评估的真实世界数据集的基本信息。Sens. 表示敏感属性的语义含义。

数据集	累犯	信用违约者	DBLP	DBLP-L
节点数	18,876	30,000	39,424	129,726
# 边	321,308	1,436,858	52,460	591,039
# 属性	18	13	5,693	5,693
平均度	34.0	95.8	1.3	4.6
敏感性	Race	Age	所属洲	所属洲
标签	保释决定	未来违约	研究领域	研究领域

表 2：基于节点分类准确率和 Δ_{SP}的实验结果。我们使用后缀“(T)”和“(S)”分别表示教师模型和学生模型。这里，Vanilla(S) 表示使用原始知识蒸馏框架训练的学生模型；One-Hot(S) 表示使用独热偏差代理训练的学生模型；RELIANT(S) 表示使用我们提出的模型训练的学生模型。 ↑ 表示越大越好；而↓ 表示相反。所有定量结果均以百分比表示。最佳结果以**粗体**标出。

			DBLP	DBLP-L	信用	累犯
CPF +GCN	准确率 (↑)	GCN(T)	92.37 ± 0.06	94.20 ± 0.09	76.39 ± 0.48	**93.68 ± 0.21**
		Vanilla(S)	**93.14 ± 0.10**	**94.30 ± 0.04**	**77.85 ± 0.10**	89.41 ± 0.12
		One-Hot(S)	93.04 ± 0.34	94.16 ± 0.02	77.65 ± 0.10	89.15 ± 0.37
		RELIANT(S)	92.70 ± 0.40	94.07 ± 0.18	77.82 ± 0.45	88.88 ± 0.57
	Δ_{SP} (↓)	GCN(T)	7.66 ± 0.26	7.33 ±0.44	15.81 ±0.40	6.10 ±0.05
		Vanilla(S)	8.55 ± 0.50	7.16 ± 0.16	14.90 ± 0.89	6.85 ± 0.05
		One-Hot(S)	7.97 ± 0.63	7.46 ± 0.24	13.80 ± 0.32	6.78 ± 0.51
		RELIANT(S)	**2.27 ± 1.00**	**3.09 ± 0.36**	**10.28 ± 1.86**	**4.06 ± 0.64**
CPF +SAGE	准确率（↑）	SAGE(T)	92.57 ± 0.28	94.10 ± 0.25	77.88 ± 0.06	**89.71 ± 0.14**
		Vanilla(S)	**93.25 ± 0.15**	**94.97 ± 0.10**	77.97 ± 0.26	89.20 ± 0.11
		One-Hot(S)	93.07 ± 0.10	94.32 ± 0.07	78.01 ± 0.23	89.11 ± 0.29
		RELIANT(S)	92.91 ± 0.51	94.17 ± 0.93	**78.28 ± 0.36**	88.85 ± 0.27
	Δ_{SP} (↓)	SAGE(T)	8.32 ±0.24	7.81 ±0.08	14.08 ± 1.37	6.50 ±0.39
		Vanilla(S)	8.29 ± 0.85	7.02 ± 0.13	13.44 ± 5.23	4.41 ± 0.43
		One-Hot(S)	8.01 ± 0.25	7.52 ± 0.32	16.86 ± 3.86	6.62 ± 0.38
		RELIANT(S)	**2.01 ± 1.21**	**2.97 ± 0.61**	**10.06 ± 1.70**	**3.94 ± 0.60**
AKD +GCN	准确率（↑）	GCN(T)	**92.37 ± 0.06**	**94.20 ± 0.09**	**76.39 ± 0.48**	**93.68 ± 0.21**
		Vanilla(S)	92.06 ± 0.16	94.07 ± 0.11	76.35 ± 0.31	92.08 ± 0.29
		One-Hot(S)	91.55 ± 0.40	94.07 ± 0.04	75.65 ± 0.75	92.07 ± 0.03
		RELIANT(S)	91.39 ± 0.24	93.98 ± 0.08	75.64 ± 0.06	91.21 ± 0.14
	Δ_{SP} (↓)	GCN(T)	7.66 ±0.26	7.33 ±0.44	15.81±0.40	6.10 ±0.05
		Vanilla(S)	7.87 ± 0.25	6.79 ± 0.10	13.61 ± 2.00	6.54 ± 0.17
		One-Hot(S)	7.39 ± 0.35	6.72 ± 0.19	14.30 ± 0.24	6.44 ± 0.32
		RELIANT(S)	**3.66 ± 1.09**	**5.18 ± 0.16**	**8.47 ± 1.92**	**5.70 ± 0.18**
AKD +SAGE	准确率（↑）	SAGE(T)	**92.57 ± 0.28**	94.10 ± 0.25	77.88 ± 0.06	**89.71 ± 0.14**
		Vanilla(S)	92.23 ± 0.07	94.45 ± 0.03	78.10 ± 0.24	89.67 ± 0.07
		One-Hot(S)	92.31 ± 0.06	**94.52 ± 0.11**	78.24 ± 0.45	89.60 ± 0.12
		RELIANT(S)	92.07 ± 0.07	94.28 ± 0.06	**78.60 ± 0.33**	88.87 ± 0.31
	Δ_{SP} (↓)	SAGE(T)	8.32 ±0.24	7.81 ±0.08	14.08 ± 1.37	6.50 ±0.39
		Vanilla(S)	7.53 ± 0.29	7.34 ± 0.41	14.41 ± 0.15	6.24 ± 0.20
		One-Hot(S)	7.72 ± 0.44	7.26 ± 0.36	11.69 ± 0.93	6.18 ± 0.30
		RELIANT(S)	**4.91 ± 0.64**	**4.05 ± 0.14**	**5.00 ± 1.63**	**6.06 ± 0.26**

基线。 据我们所知，这是首篇研究如何缓解图神经网络知识蒸馏框架中所呈现偏差的工作。在实验中，我们采用原始知识蒸馏框架生成的学生模型作为第一个基线。对于第二个基线，我们将偏差的可学习代理替换为知识蒸馏框架输入中的朴素代理。具体而言，在训练期间，我们使用独热向量作为不同人口统计学子群体的朴素代理，其中独热向量标记不同节点所属的群体。在推理阶段，我们将所有代理向量替换为所有实例的平均代理向量。其依据是这是因为更具区分性的属性更容易被深度学习模型在训练过程中学习，而这些独热向量则充当了偏差信息的“更容易”指标。这样一来，如果这些独热代理能够解释学生模型训练后表现出的偏差，那么在推理过程中也可以减轻这种偏差，因为此时这类信息会被清除。

评估指标。我们从效用和公平性两个方面评估压缩图神经网络模型（即知识蒸馏框架输出的学生模型）的性能。具体而言，在效用方面，我们采用节点分类准确率作为相应的指标；

版权所有 © 2023年 SIAM 未经授权，禁止复制本文。

在公平性方面，我们采用两种传统指标 Δ_{SP} 和 Δ_{EO}。其中，Δ_{SP} 衡量统计均等这一公平性概念下（预测）的偏差程度，而 Δ_{EO} 衡量机会均等这一公平性概念下的偏差程度。由于篇幅有限，其他结果请参见本文在线版本的附录。

实现细节。RELIANT 使用 PyTorch [25] 实现，并采用 Adam 优化器 [20] 进行优化。在我们的实验中，学习率选取为$\{10^{-2}, 10^{-3}, 10^{-4}\}$，CPF 的训练轮次设为 1,000，GraphAKD 的训练轮次设为 600。实验在 Nvidia RTX A6000 上进行，报告的数值结果为三次不同运行的平均值。更多细节见附录。

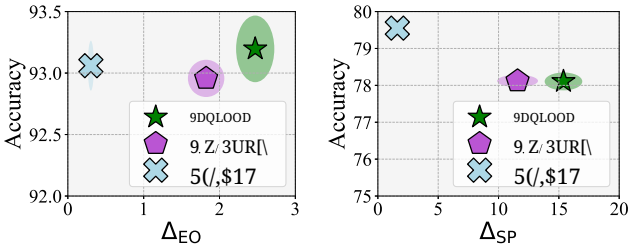
5.2 RELIANT 的有效性这里我们旨在回答**研究问题1**。具体而言，我们在两种知识蒸馏骨干网络 CPF 和 GraphAKD 上评估所提出的框架 RELIANT。对于每种知识蒸馏骨干网络，我们采用两种不同的图神经网络（GCN 和 GraphSAGE），以评估所提出框架对不同图神经网络的泛化能力。我们比较教师图神经网络模型与采用三种不同框架训练的学生模型的相应表现（即原始知识蒸馏框架、采用偏差的独热代理的知识蒸馏框架，以及所提出的 RELIANT）。我们在表 2 中给出了节点分类准确率和 Δ_{SP} 的定量结果。此外，我们还基于机会均等进行了实验（见附录），观察结果一致。我们从表 2 中得出以下观察。

•从预测效用的角度来看，采用三种知识蒸馏框架训练的学生模型，其表现均与教师模型相当。这表明，三种知识蒸馏框架都能够有效地实现知识蒸馏。

• 从偏差缓解的角度来看，在所有情况下，采用原始知识蒸馏框架训练的学生模型都会继承教师图神经网络模型所表现出的偏差，甚至使其加剧。使用独热代理训练学生模型在大多数情况下可以缓解偏差。与采用原始知识蒸馏框架和独热代理训练的学生模型相比，RELIANT 在统计均等方面始终表现出更小的偏差。

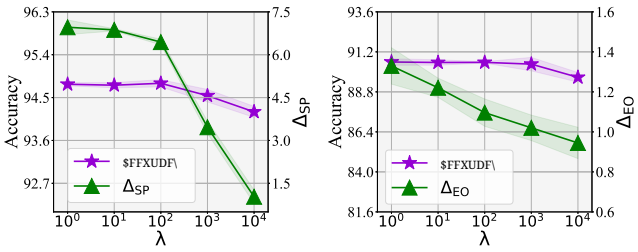
• 根据 RELIANT 在这两个方面的表现，RELIANT 能够有效地对学生模型进行去偏，同时仍能使其模型效用与教师模型相当。因此，我们

我们认为，RELIANT 在去偏与维持效用之间实现了令人满意的平衡。



(a) 基于信用的 GraphAKD。 (b) DBLP 上的 CPF。

图 3：RELIANT 的消融研究。“原始方法”表示使用原始知识蒸馏框架训练的学生模型，而“原始方法 + 代理”表示仅学习偏差代理、并在知识蒸馏框架下训练的学生模型。



(a) DBLP-L 上的 CPF。 (b) 累犯数据集上的 GraphAKD。

图 4：基于两种不同知识蒸馏骨干网络，在两个真实世界数据集上对 λ 进行参数敏感性分析。我们在其他数据集上也观察到类似结果。

5.3 消融研究 本小节旨在回答**RQ2** 。具体而言，对于每个框架，我们评估 RELIANT 的两个模块（包括偏差代理学习 和强制将偏差归因于代理）对学生模型性能的贡献程度。结果如图 3 所示。其中，图 3a 展示了基于 DBLP-L 数据集的 CPF 中准确率与 Δ_{SP} 的关系，图 3b 展示了基于再犯数据集的 GraphAKD 中准确率与 Δ_{EO} 的关系。值得注意的是，在其他设置下，我们也观察到类似的结果。我们得出以下结论。

•从预测效用的角度来看，我们观察到这三种情况下的预测效用相当。这进一步证实，两个模块对节点分类准确率的影响都有限。

• 从偏差缓解的角度来看，在原始知识蒸馏框架中加入偏差代理学习 模块，对偏差缓解的作用有限。这是因为偏差也可能来自非敏感节点属性（如第3.1节所述）。同时加入强制将偏差归因于代理 模块和偏差代理学习模块后，RELIANT便能在偏差缓解方面取得令人满意的表现。

版权所有 © 2023年 SIAM。未经授权，禁止转载本文。

5.4 参数敏感性我们通过研究模型效用和表现出的偏差随超参数 λ变化的趋势来回答**RQ3** 。此处，λ 控制 $\mathscr{L}_{\text{Attr}}$的影响。具体而言，我们将 λ的取值设为{$10^0, 10^1, 10^2, 10^3, 10^4$}，并在图4中展示了采用RELIANT训练的学生模型的节点分类准确率和表现出的偏差的相应趋势。其中，图4a基于GraphAKD下的Credit数据集，图4b基于CPF下的DBLP数据集。我们在其他数据集上也观察到类似现象。根据图4，我们得出以下观察结果。

- 从预测效用的角度来看，当 λ 的值从 1 增加到 10_4时，两个数据集和知识蒸馏骨干网络上的节点分类准确率均未出现明显下降。这验证了预测效用对 λ 并不敏感。
- 从偏差缓解的角度来看，随着 λ 从 1 增加到 10_4，学生模型表现出的偏差会减少。具体而言，当 λ 相对较小（例如 1）时，学习到的偏差代理只能部分解释表现出的偏差；当λ 的值增加时，更多偏差便会归因于学习到的代理。综合考虑模型效用与偏差缓解，建议将 λ 的范围设为 10_2 到 10_3。

6 相关工作

图神经网络中的算法公平性。 现有大多数旨在提升图神经网络算法公平性的研究，关注的要么是群体公平性 [10]，要么是个体公平 [36]。具体而言，群体公平性是基于一组预先定义的敏感属性（例如性别和种族）来界定的。这些敏感属性将整个群体划分为不同的人口统计学子群体。群体公平性要求每个子群体都应根据图神经网络的输出预测公平地获得其应得份额 [23]。研究者已开展多种探索，以提升图神经网络的群体公平性 [7]。通过对抗学习使输出预测与敏感属性脱钩，是现有研究中最受欢迎的方法之一 [31, 3]。其他常见策略包括通过公平性正则化重新构造目标函数 [11, 24],、重新平衡两个不同人口统计学子群体之间的组内边数量 [6, 22], ，以及删除对表现出的偏差影响最大的节点或边 [8, 9],等。另一方面，个体公平不依赖任何敏感属性。相反，个体公平要求相似的节点（在输入空间中）应受到相似的对待（在输出空间中） [10]。为实现图神经网络中的个体公平，将公平性感知正则化项添加到优化目标中是最广泛采用的方法 [5, 28]。

知识蒸馏。 近年来，知识蒸馏已被证明能够有效压缩模型，同时仍保持相近的模型预测性能 [12]。因此，它已广泛应用于众多领域，包括视觉识别 [33], 自然语言处理 [13, 17],等。知识蒸馏的核心思想是将计算成本高昂的教师模型中的知识迁移到轻量级学生模型中，从而使学生模型能够在计算资源有限的平台上运行 [16, 19]。值得注意的是，这种策略也已被证明能够有效压缩图神经网络 [34, 16, 19]。因此，越来越多的研究开始关注利用知识蒸馏压缩图神经网络，以提高推理效率。例如，已有研究证明，通过正则化鼓励学生模型生成与教师图神经网络相似的输出是有效的 [16]。此外，对抗学习也是一种常用技术，可用于获得轻量且准确的学生模型 [16]。然而，这些面向图神经网络的大多数框架都未考虑公平性。因此，学生模型往往会受到教师图神经网络中有偏知识的影响。与现有研究不同，我们提出了一种通用的知识蒸馏框架，明确考虑图神经网络中的公平性，同时保持图神经网络预测的效用。

7 结论

尽管知识蒸馏（KD）在压缩图神经网络方面取得了成功，但现有大多数研究并未考虑公平性。因此，使用知识蒸馏框架训练的学生模型往往会继承甚至放大教师图神经网络中的偏差。本文初步探索了如何为基于图神经网络的知识蒸馏框架学习偏差更小的学生模型。具体而言，我们首先针对基于 GNN 的教师-学生框架提出了公平知识蒸馏这一新问题，随后提出名为 RELIANT 的框架，以实现偏差更小的学生模型。值得注意的是，RELIANT 的设计与教师模型和学生模型的具体结构无关。因此，该框架可以轻松适配不同的知识蒸馏方法，用于去偏。大量实验表明，RELIANT 在通过知识蒸馏实现图神经网络压缩的公平性方面具有良好效果。

8 致谢

本研究得到美国国家科学基金会资助，资助项目编号为 IIS-2006844、IIS-2144209、IIS-2223768、IIS-2223769、CNS-2154962、CMMI-2125326、BCS-2228533 和 BCS-2228534；同时还获得摩根大通教师研究奖和思科教师研究奖的支持。我们谨向匿名评审人致谢，感谢他们提出的建设性意见。

版权所有 © 2023年 SIAM。未经授权，禁止复制本文。

参考文献

[1]Agarwal, C., Lakkaraju, H., 和 Zitnik, M. 面向公平且稳定的图表示学习统一框架。载于 *UAI* （2021年）。[2] Chen, M., Wei, Z., Huang, Z., Ding, B., 和 Li, Y. 简单而深层的图卷积网络。载于 *ICML* （2020年）。[3] Dai, E., 和 Wang, S. 向歧视说不：利用有限的敏感属性信息学习公平图神经网络。载于 *WSDM* （2021年）。[4] Dattoli, G., Ricci, P. E., 和 Cesarano, C. 关于勒让德多项式的一则说明。《国际非线性科学与数值模拟杂志》*2*，4（2001年），365–370。[5] Dong, Y., Kang, J., Tong, H., 和 Li, J. 图神经网络的个体公平：一种基于排序的方法。载于 *SIGKDD* （2021年）。[6] Dong, Y., Liu, N., Jalaian, B., 和 Li, J. EDITS：为图神经网络建模并缓解数据偏差。载于*WWW* （2022年）。[7] Dong, Y., Ma, J., Chen, C., 和 Li, J. 图挖掘中的公平性：综述。*arXiv* 预印本 *arXiv:2204.09888* （2022年）。[8] Dong, Y., Wang, S., Ma, J., Liu, N., 和 Li, J. 通过训练节点归因解释图神经网络中的不公平现象。载于 *AAAI* （2023年）。[9] Dong, Y., Wang, S., Wang, Y., Derr, T., 和 Li, J. 关于图神经网络中偏差的结构性解释。载于 *SIGKDD* （2022年）。[10] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., 和 Zemel, R. 通过认知实现公平。载于 *ITCS* （2012年）。[11] Fan, W., Liu, K., Xie, R., Liu, H., Xiong, H., 和 Fu, Y. 基于瓦瑟斯坦距离的公平图自编码器，用于学习无偏图表示。载于 *ICDM* （2021年）。[12] Gou, J., Yu, B., Maybank, S. J., 和 Tao, D. 知识蒸馏：综述。*International Journal of Computer Vision 129*，6（2021年），1789–1819。[13] Haidar, M., Rezagholizadeh, M., 等. Textkd-gan：利用知识蒸馏和生成对抗网络进行文本生成。载于 加拿大人工智能会议 （2019年），施普林格，第107–118页。[14] Hamilton, W., Ying, Z., 和 Leskovec, J. 大型图上的归纳式表示学习。载于 *NeurIPS* （2017年）。[15] Hardt, M., Price, E., 和 Srebro, N. 监督学习中的机会平等。载于 *NeurIPS* （2016年）。[16] He, H., Wang, J., Zhang, Z., 和 Wu, F. 通过对抗式知识蒸馏压缩深层图神经网络。*arXiv* 预印本 *arXiv:2205.11678* （2022年）。[17] Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., Wang, F., 和 Liu, Q. TinyBERT：为自然语言理解蒸馏 BERT。载于 *EMNLP* （2020年）。[18] Jordan, K. L., 和 Freiburger, T. L. 种族/族裔对量刑的影响：考察刑罚类型、拘留时间和监禁时间。*J Crim Justice* （2015年）。

[19] Joshi, C. K., Liu, F., Xun, X., Lin, J., 和 Foo, C.-S. 关于图神经网络表示知识蒸馏的研究。*arXiv* 预印本 *arXiv:2111.04964* (2021年)。[20] Kingma, D. P., and Ba, J. Adam：一种随机优化方法。载于*ICLR* (2015年)。[21] Kipf, T. N., and Welling, M. 基于图卷积网络的半监督分类。载于*ICLR* (2017年)。[22] Li, P., Wang, Y., Zhao, H., Hong, P., and Liu, H. 关于二元公平性：探索并缓解图连接中的偏差。载于*ICLR* (2021年)。[23] M., N., M., F., S., N., L., K., and G., A. 机器学习中的偏差与公平性综述。*CSUR* (2021年)。[24] Navarin, N., Oneto, L., and Donini, M. 深度公平图神经网络的学习。载于*ECML* (2020年)。[25] Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. PyTorch中的自动微分。载于*NeurIPS* (2017年)。[26] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., 等。PyTorch：一种命令式风格的高性能深度学习库。载于*NeurIPS* (2019年)。[27] Saha, P., Mukherjee, D., Singh, P. K., Ahmadian, A., Ferrara, M., and Sarkar, R. GraphCovidNet：一种基于图神经网络的模型，用于从胸部CT扫描和X射线中检测新冠肺炎。*Scientific Reports 11*, 1 (2021年), 1–16。[28] Song, W., Dong, Y., Liu, N., and Li, J. Guide：图神经网络中由群体平等引导的个体公平。载于*KDD* (2022年)。[29] Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., and Su, Z. ArnetMiner：学术社交网络的提取与挖掘。载于*KDD* (2008年)。[30] Wang, D., Zhang, Z., Zhou, J., Cui, P., Fang, J., Jia, Q., Fang, Y., and Qi, Y. 用于信用风险预测的时间感知图神经网络。载于*SDM* (2021年)。[31] Wang, Y., Zhao, Y., Dong, Y., Chen, H., Li, J., and Derr, T. 通过缓解敏感属性泄漏提升图神经网络的公平性。载于*SIGKDD* (2022年)。[32] Wu, F., Souza, A., Zhang, T., Fifty, C., Yu, T., and Weinberger, K. 简化图卷积网络。载于*ICML* (2019年)。[33] Wu, X., He, R., Hu, Y., and Sun, Z. 通过大规模知识蒸馏学习演化嵌入。*International Journal of Computer Vision* (2020年)。[34] Yang, C., Liu, J., and Shi, C. 提取图神经网络的知识并超越它：一种有效的知识蒸馏框架。载于*WWW* (2021年)。[35] Yeh, I.-C., and Lien, C.-h. 比较数据挖掘技术对信用卡客户违约概率预测准确率的影响。*Expert systems with applications* (2009年)。[36] Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. 公平表示的学习。载于*ICML* (2013年)。

版权所有 © 2023年 SIAM。未经授权，禁止复制本文。

A 可复现性

本节将介绍所展示实验的可复现性，作为第5节的补充。具体而言，我们首先详细介绍实验设置，随后说明所提出的RELIANT框架、图神经网络、知识蒸馏骨干网络以及基线去偏框架的实现细节。最后，我们介绍实现中使用的若干关键软件包及其对应版本。代码将在论文录用后发布。

A.1 实验设置。 我们的实验实现主要分为三个部分：训练良好的教师图神经网络、知识蒸馏骨干网络，以及我们的 RELIANT 模块。通常，我们的实验在 Nvidia RTX A6000 上进行。我们使用 PyTorch [26]框架编写实验代码，并使用 Adam [20] 优化器来学习模型参数。所有实验均运行三次，并记录平均值和标准差；随机种子取自 {0,10,100,1000,10000}。

A.2 RELIANT 的实现。 我们使用 PyTorch 实现 RELIANT，并使用 Adam 作为可学习代理的优化器。对于表2和表4中的结果，CPF 的代理学习率设为 10^{-2}，权重衰减设为 10^{-2}；GraphAKD 的代理学习率设为 10^{-2}，权重衰减设为 5×10^{-4}。对于所有结果，我们都在集合 {1,10,100,1000,10000} 中搜索最优系数 λ。

A.3 图神经网络的实现。 对于教师图神经网络模型的训练，我们使用 CPF 框架中的代码，其实现设置的详细信息见表 3。

A.4 知识蒸馏框架的实现。 在我们的实验设置中，教师模型在训练阶段保持固定。因此，我们首先使用表 3 所示的参数训练一个教师图神经网络。获得训练充分的教师图神经网络后，我们将其作为监督信号来训练学生模型。学生模型训练的详细信息如下。

- 对于 CPF，我们遵循 [34], 中的实现，并使用 PLP 作为学生模型。学生模型的训练设置如下：最大训练轮次设为 1,000，早停设为 500，层数设为 5，特征丢弃率设为 0.8，边权重丢弃率设为 0.2，学习率设为 $\{10^{-2}, 10^{-3}, 10^{-4}\}$，权重衰减设为 10^{-2}。
- 对于 GraphAKD，我们遵循[16], 中的实现，并使用 SGC [32] 作为学生模型。关于详细的训练设置，我们将最大训练轮次设为 600，层数设为 3，学习率设为 $\{10^{-2}, 10^{-3}, 10^{-4}\}$，权重衰减设为 5×10^{-4}，丢弃率设为 0.5。

表 3：教师图神经网络模型的实验设置。

超参数	GCN	GraphSAGE
层	3	3
隐藏维度	64	128
训练轮次	1000	1000
早停	500	500
学习率	10^{-2}	10^{-2}
权重衰减	10^{-3}	5×10^{-4}
丢弃	0.8	0.5

A.5 基线去偏框架的实现。 对于原始知识蒸馏框架，我们遵循附录 A.4 中的设置，直接实现知识蒸馏框架。对于使用独热偏差代理的基线，我们将敏感属性的独热向量用作常量代理。在训练前将其添加到输入数据中，并在训练阶段保持固定。因此，常量代理没有损失项（如公式 3.4 所示）。独热基线的训练设置与原始方法相同，因为常量代理向量不包含任何额外参数。

A.6 实现所需的软件包。 我们列出了实现中使用的主要软件包及其对应版本，如下所示。

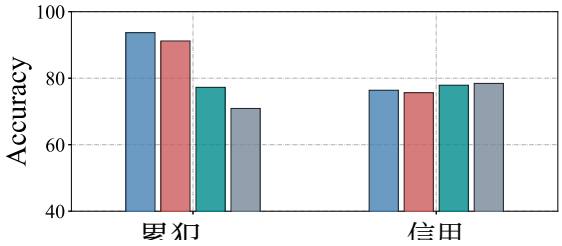
- Python == 3.7.10
- torch == 1.8.1
- torch-cluster == 1.5.9
- torch-geometric == 1.4.1
- torch-sparse == 0.6.9
- NumPy == 1.20.0
- NetworkX == 2.5.1
- scikit-learn == 0.24.1
- pandas==1.2.3
- SciPy==1.4.1

B 补充实验

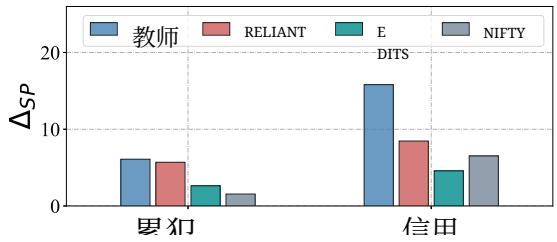
本节中，我们进行了实验，作为第5节的补充。具体而言，我们首先展示 RELIANT 在另一种广泛研究的公平性概念——机会均等下，在预测效用与公平性方面的表现。随后，我们通过实验比较 RELIANT 在平衡图神经网络的预测效用与公平性方面的表现，以及直接对图神经网络进行去偏的最先进方法。

B.1 RELIANT的有效性 我们在此呈现补充实验，以回答**研究问题1**，其中

版权所有 © 2023年 SIAM。未经授权，禁止复制本文。



(a) 节点分类准确率比较。



(b) Δ_{SP}比较。

图5：所提出的框架 RELIANT 与其他直接对图神经网络进行去偏的最先进方法，在平衡预测效用与偏差缓解方面的性能比较。所有数值结果均以百分比表示。

公平性概念具体采用机会均等（使用 Δ_{EO}进行测量）。此处我们采用与第5.2节相同的设置。我们比较教师图神经网络模型与采用三种不同框架变体训练的学生模型的表现，这些变体包括原始知识蒸馏框架、使用偏差的独热代理的知识蒸馏框架，以及我们提出的 RELIANT。我们在表4中给出了节点分类准确率和 Δ_{EO} 的定量结果。我们从表4中得出以下观察结果。

- 从预测效用的角度来看，使用所采用的三种知识蒸馏框架训练的所有学生模型，其性能均可与教师模型相媲美。这证实了这三种知识蒸馏框架都能够有效地实现知识蒸馏。

- 从偏差缓解的角度来看，使用原始知识蒸馏框架训练的学生模型在所有情况下都会继承并放大教师图神经网络的偏差。与使用原始知识蒸馏框架和独热代理训练的学生模型相比，RELIANT 在机会均等这一公平性概念下始终表现出更低的偏差。在某些情况下，使用 RELIANT 训练的学生模型甚至比教师模型的偏差更低，这进一步证实了 RELIANT 在训练偏差更低的学生模型方面的有效性。

•根据 RELIANT 在上述两个方面的表现，可以证明 RELIANT 能够有效地对学生模型进行去偏，同时与教师模型相比保持相当的预测效用。因此，我们认为 RELIANT 在去偏与保持预测效用之间实现了令人满意的平衡。

B.2 RELIANT 与图神经网络去偏方法的比较在本小节中，我们开展实验，将 RELIANT 与其他最先进的图神经网络去偏方法进行比较，以评估它们在平衡预测效用与偏差缓解方面的表现。

基线。 我们在此选取两种最先进的图神经网络去偏方法作为基线，即 EDITS [6]和 NIFTY [1]。EDITS 是一种新近提出的图神经网络去偏方法，可在预处理阶段学习偏差较低的网络数据。去偏后，网络数据将输入图神经网络模型进行评估。NIFTY 是另一种新近提出的基于图神经网络的去偏框架，可在处理阶段实现偏差缓解。在训练过程中，经过反事实扰动后，节点表示被学习为对敏感属性保持不变。

骨干网络。 在这里，我们选择应用最广泛的图卷积网络（GCN）作为所有方法的骨干图神经网络模型。我们选择 GraphAKD 作为所提出的 RELIANT 的知识蒸馏骨干网络。值得注意的是，我们在其他图神经网络骨干网络上也观察到了类似现象。

讨论。 我们在图 5 中展示了节点分类准确率的性能比较结果 Δ_{SP} 。我们得出以下观察结果。

•从预测效用的角度来看，RELIANT 的表现与教师 GCN 模型相当，而其他去偏方法则在不同程度上损害了预测效用。因此，在所有方法中，RELIANT 在维持预测效用方面表现令人满意。

- 从偏差缓解的角度来看，当所有模型的预测效用相近时（例如，在 Credit数据集上），RELIANT 能够取得与其他基线方法相当的去偏性能；当基线方法在偏差缓解方面优于 RELIANT 时，其预测效用的牺牲也大得多（例如，在再犯数据集上）。考虑到在有偏监督下对学生模型进行去偏，远比直接对图神经网络进行去偏困难，我们认为 RELIANT 在这两种情况下的表现都令人满意。

•根据 RELIANT 在上述两个方面的表现，我们认为 RELIANT 的性能可与其他最先进的图神经网络去偏方法相媲美，这进一步证明了它在平衡预测效用与偏差缓解方面表现出色。

版权所有 © 2023年 SIAM 未经授权，禁止转载本文

表4：基于节点分类准确率和 Δ_{EO}的实验结果。我们分别使用后缀“(T)”和“(S)”表示教师模型和学生模型。此处，Vanilla(S)表示使用原始知识蒸馏框架训练的学生模型；One-Hot(S)表示使用独热偏差代理训练的学生模型；RELIANT(S)表示使用我们提出的模型训练的学生模型。 ↑ 表示越大越好；而↓ 则表示越小越好。所有定量结果均以百分比表示。最佳结果以**粗体**显示。

			DBLP	DBLP-L	信用	累犯
CPF +GCN	准确率（↑）	GCN(T)	92.37 ± 0.06	**94.20 ± 0.09**	76.39 ± 0.48	**93.68 ± 0.21**
		Vanilla(S)	**93.24 ± 0.18**	94.15 ± 0.04	77.58 ± 0.20	89.36 ± 0.16
		One-Hot(S))	93.07 ± 0.35	94.15 ± 0.07	**77.65 ± 0.10**	89.38 ± 0.15
		RELIANT(S)	93.20 ± 0.12	94.15 ± 0.08	77.00 ± 1.57	89.30 ± 0.19
	Δ_{EO} (↓)	GCN(T)	2.31 ± 0.13	2.29 ± 0.34	12.63 ± 0.24	0.52 ± 0.06
		Vanilla(S)	2.56 ± 0.11	2.34 ± 0.29	11.56 ± 0.38	1.25 ± 0.19
		One-Hot(S)	2.21 ± 0.32	2.17 ± 0.26	10.69 ± 0.31	0.96 ± 0.28
		RELIANT(S)	**0.42 ± 0.21**	**1.08 ± 0.10**	**6.02 ± 4.78**	**0.35 ± 0.12**
CPF +SAGE	准确率（↑）	SAGE(T)	92.57 ± 0.28	94.10 ± 0.25	77.88 ± 0.06	**89.71 ± 0.14**
		Vanilla(S)	93.22 ± 0.03	**94.37 ± 0.08**	**78.30 ± 0.23**	89.15 ± 0.27
		One-Hot(S)	93.13 ± 0.11	94.36 ± 0.06	78.01 ± 0.23	88.98 ± 0.55
		RELIANT(S)	**93.24 ± 0.09**	94.32 ± 0.06	78.11 ± 0.40	89.01 ± 0.26
	Δ_{EO} (↓)	SAGE(T)	2.51 ± 0.33	2.67 ± 0.19	11.05 ± 0.71	0.86 ± 0.03
		Vanilla(S)	2.83 ± 0.34	2.00 ± 0.18	11.07 ± 4.61	1.17 ± 0.11
		One-Hot(S)	2.16 ± 0.27	2.05 ± 0.21	12.73 ± 2.29	1.23 ± 0.08
		RELIANT(S)	**0.63 ± 0.42**	**0.86 ± 0.18**	**6.72 ± 4.49**	**0.51 ± 0.25**
AKD +GCN	准确率 (↑)	GCN(T)	**92.37 ± 0.06**	**94.20 ± 0.09**	76.39 ± 0.48	**93.68 ± 0.21**
		Vanilla(S)	92.12 ± 0.09	94.06 ± 0.06	**78.12 ± 0.65**	92.29 ± 0.06
		One-Hot(S)	91.68 ± 0.28	93.98 ± 0.13	77.87 ± 0.48	92.28 ± 0.13
		RELIANT(S)	91.69 ± 0.19	94.09 ± 0.12	77.88 ± 0.82	92.46 ± 0.09
	Δ_{EO} (↓)	GCN(T)	2.31 ± 0.13	2.29 ± 0.34	12.63 ± 0.24	**0.52 ± 0.06**
		Vanilla(S)	2.76 ± 0.33	1.88 ± 0.08	8.26 ± 3.41	0.82 ± 0.17
		One-Hot(S)	2.69 ± 0.28	1.87 ± 0.17	8.43 ± 5.08	0.97 ± 0.45
		RELIANT(S)	**1.79 ± 0.31**	**1.43 ± 0.09**	**4.96 ± 3.77**	0.66 ± 0.21
AKD +SAGE	准确率 (↑)	SAGE(T)	**92.57 ± 0.28**	94.10 ± 0.25	77.88 ± 0.06	89.71 ± 0.14
		Vanilla(S)	92.23 ± 0.07	94.45 ± 0.03	78.10 ± 0.24	90.56 ± 0.14
		One-Hot(S)	92.31 ± 0.06	**94.52 ± 0.11**	78.24 ± 0.45	**90.85 ± 0.20**
		RELIANT(S)	92.15 ± 0.16	94.42 ± 0.05	**79.08 ± 0.29**	90.00 ± 0.64
	Δ_{EO} (↓)	SAGE(T)	2.51 ± 0.33	2.67 ± 0.19	11.05 ± 0.71	**0.86 ± 0.03**
		Vanilla(S)	2.06 ± 0.06	2.23 ± 0.23	10.56 ± 0.43	1.61 ± 0.39
		One-Hot(S)	2.21 ± 0.39	2.11 ± 0.21	8.38 ± 0.73	1.10 ± 0.37
		RELIANT(S)	**1.60 ± 0.45**	**1.89 ± 0.21**	**2.33 ± 0.80**	0.91 ± 0.22

版权所有 © 2023年 SIAM。未经授权，禁止转载本文。