arXiv:2304.01235v2 [cs.LG] 9 Feb 2024

图结构和标签依赖关系如何促进大型文档网络中的节点分类

Pirmin Lemberger
onepoint 萨布隆街29号，75016，巴黎（法国）
p.lemberger@groupeonepoint.com

Antoine Saillenfest
onepoint 萨布隆街29号，75016，巴黎（法国）
a.saillenfest@groupeonepoint.com

2024年2月12日

摘要

我们介绍一个名为 *WikiVitals* 的新数据集，其中包含一个大型图，由 4.8 万篇相互引用的维基百科文章构成，这些文章被划分为 32 个类别，并由 230 万条边连接。我们的目标是在半监督节点分类设置下，严格评估三种不同信息来源对标签预测的贡献，分别是文章内容、文章之间的连接，以及标签之间的相关性。我们使用图马尔可夫神经网络进行评估。该模型为这项任务提供了具有理论依据的方法；我们还通过明确区分模型选择与模型评估，详细评估了每种信息来源的贡献。一个有趣的发现是，与密集训练集相比，考虑标签依赖关系对稀疏训练集更为重要。

1 引言

图神经网络（GNN）已成为建模此类数据集的首选工具：这类数据集中的观测值并非独立同分布（i.i.d.），而是由依据关系图相互连接的实体组成（Yang 等，2016年；Defferrard 等，2016年；Kipf 和 Welling，2017年）。图神经网络既可用于图分类，例如分子分类（Dobson 和 Doig，2003年；Borgwardt 等，2005年），也可用于节点分类，例如引用网络中的文档分类（Sen 等，2008年）。最常见的任务是半监督节点分类：利用同一图中一个互不重叠的已标记节点子集，对给定子集中的未标记节点进行分类（Defferrard 等，2016年；Kipf 和 Welling，2017年）。本文聚焦于这一任务。

多年来，人们提出了许多架构，以应对图神经网络中出现的特定问题。有些方法用于缓解过平滑（即深层图神经网络倾向于为所有节点预测相同标签的现象）（Klicpera 等人，2018年）；有些方法处理同配性或异配性（即邻居节点很可能具有不同标签的情况）（Zhu 等人，2020年、2021年；Bo 等人，2021年）；还有一些方法则尝试通过适当的注意力机制从数据中学习连接权重（Veli˘ckovi´c 等人，2018年）。

尽管这些模型各不相同，但它们都有一个重要的局限：它们假设图中的标签可以针对每个节点独立预测。换言之，它们完全忽略了标签依赖关系。近来，图马尔可夫神经网络（GMNN）（Qu 等人，2019）作为一致的概率模型被提出，它通过结合图神经网络与条件随机场（CRF）的优势，以自然的方式纳入图中的标签相关性。

本研究旨在探讨如何在半监督环境下，利用三种不同的信息来源，对较大型文档网络中的文本进行分类。除了文档内容这一显然最相关的信息外，我们还计划严谨评估网络结构和标签依赖关系虽较弱但仍有意义的作用。为此，我们使用图马尔可夫神经网络，因为它们能生成融合这三种信息来源的文本表示。

简而言之，图马尔可夫神经网络由两个相互耦合的图神经网络组成。其中一个采用平均场近似，假设标签彼此不相关，并根据文本特征预测标签。另一个与其耦合的图神经网络则对标签依赖关系建模。这样便于区分这三种因素各自的作用。将图马尔可夫神经网络简化为一个简单的多层感知机，即可得到图无关基线；该基线仅考虑文档内容，并对每篇文档独立进行预测。只使用预测不相关标签的平均场图神经网络，则可以利用图结构而不考虑标签相关性。最后，完整的图马尔可夫神经网络还会纳入标签相关性。
Qu 等人（2019）在经典基准数据集 Cora、Pubmed 和 Citeseer（Sen 等人，2008）上评估了标签相关性对图马尔可夫神经网络（GMNN）预测准确率的贡献，发现其影响较小。如表 2 所示，与包含相当比例维基百科条目的图相比，这些数据集的链接数量仍然相当少。此外，评估采用了 Yang 等人（2016）为每个数据集定义的单一公开数据划分。在这些设置下，与现有未考虑标签依赖关系的基线模型相比，图马尔可夫神经网络模型表现出性能提升。
然而，正如近期多项研究（Shchur 等人，2018年；Errica 等人，2020年）所指出的，要公平评估图神经网络的性能，需要采用一种程序，对训练集-验证集-测试集划分进行系统性随机化，并明确区分模型选择与模型评估。

本文旨在对图马尔可夫神经网络进行严格的性能分析，研究对象是我们为此构建的一个规模较大、带有主题标签的文档图，名为*WikiVitals* 。为求完整，我们还在经典基准数据集 Cora、Citeseer 和 Pubmed 上进行了同样全面的分析。

总之，我们的贡献[1] 如下：

- 我们介绍了一个从英文维基百科中提取的互联文档新数据集，名为*WikiVitals* 。与经典基准数据集相比，这是一个规模较大的图，包含 48k 个节点，分为 32 个类别，并由 2.3M 条边连接。

- 我们将 Errica 等人（2020）提出的公平比较流程应用于图马尔可夫神经网络。迄今为止，只有图分类模型以这种方式接受过评估。

- 我们评估了*WikiVitals* 文章的内容、图结构以及其标签之间的相关性分别对分类准确率的贡献。我们区分稀疏训练集和密集训练集：在这两种情况下，带标签的节点分别占网络中节点的一小部分和较大部分。两种情况下，我们的结果都证实，考虑标签相关性可带来统计显著性提升；同时，这种影响在稀疏训练集中远强于密集训练集。

2 相关工作

2.1 评估图神经网络的性能

正如 Shchur 等人（2018年）的作者所指出的，图神经网络模型的评估几乎从未以严谨的方式进行。一方面，由于缺乏对评估流程的明确定义，许多实验无法复现。另一方面，他们认为，使用单一数据划分（通常是引入新基准数据集的论文中所定义的数据划分），不足以保证两个竞争性图神经网络架构的准确率之间存在显著差异。因此，作者建议统一超参数的选择，并在多个训练集-验证集-测试集划分上进行随机化。随后，他们在这些数据划分上寻找一组能够优化平均性能的超参数。令人惊讶的是，他们发现，对于半监督节点分类任务，最简单的架构（如图卷积网络）（Defferrard 等人，2016年；Kipf 和 Welling，2017年）往往比更复杂的模型（Monti 等人，2017年；Veli˘ckovi´c 等人，2018年）表现更好。

在我们的工作中，我们采用了一种更为严谨的准确率评估方法。该方法最初由 Errica 等人（2020）提出，作为图分类任务的最先进水平评估流程。对于给定模型，我们针对每个数据划分搜索最佳超参数，然后对这些经过优化的模型在各个数据划分上的准确率估计取平均。这样可以公平地比较两个模型：从某种意义上说，它保证了实践者随机选择一个数据划分，在训练集上训练模型、在验证集上优化超参数，并在测试集上估计准确率时，所得估计能够可靠地用于比较多层感知机、图卷积网络或图马尔可夫神经网络等模型。

[1]代码和数据可在此处获取：
https://github.com/ToineSayan/node-classification-and-label-dependencies/

2.2 在图神经网络中建模标签依赖

在图神经网络近期兴起之前，已有不少研究尝试通过各种启发式方法纳入标签依赖关系。标签传播便是一种早期尝试，其代价函数在惩罚错误标签预测的同时，要求节点标签平滑变化（Zhou 等人，2003年；Zhu，2005年）。

人们也提出了针对特定数据集的方法。就维基百科文章分类而言，Viard 等人（2020）的作者使用了一种简单的图神经网络，其权重会根据邻居节点标签的相似性进行经验性调整。尽管这些方法取得了一些经验上的成功（Huang 等人，2021），但由于缺乏坚实的概率基础，因此难以分析它们为何失败或成功。特别是，这些方法没有明确区分节点特征、图结构和标签相关性对预测准确率的贡献。因此，在我们的工作中，我们决定避免使用节点度、介数或同配性等拓扑节点特征（Newman，2003年，2005年；Blondel 等人，2008年），以便更清晰地区分这些因素。

图神经网络非常适合计算分布式节点表示，将节点特征所提供的信息（例如文档的内容）与其邻域中的图局部结构相融合。随后，每个这样的表示都会被单独用于预测相应节点的标签，而不考虑其他节点的标签。另一方面，条件随机场则适合描述标签相关性。然而，由于难以计算其配分函数，执行精确推断十分困难。图马尔可夫神经网络通过使用两个普通的图神经网络，并在使用期望最大化（EM）算法进行训练时将二者耦合起来，为这一难题提出了优雅的解决方案。Qu 等人（2019）以通常的方式对其性能进行了评估：使用小型引文网络的公开数据划分（Sen 等人，2008年），但未考虑在使用不同数据划分时该评估的稳健性，而这对于公平评估至关重要（Shchur等人，2018年；Errica等人，2020年）。我们的目标是为大型现实世界文档图提供详细且严谨的评估。

2.3 维基百科文章分类

维基百科文章提供了丰富的文本内容，可以据此计算表示，例如使用信息丰富的 n-gram。超链接定义了一个图，其节点是文章。包括 Squirrel 和 Chameleon（Rozemberczki 等人，2021年）[2] 在内的多个数据集均由维基百科创建而成，通常用于评估各种图神经网络架构。本文介绍的*WikiVitals* 数据集也是如此。

每个数据集中的标签语义各不相同。例如，Squirrel 和 Chameleon 的标签基于月度流量数据（通过文章的元数据获取），这些数据被离散化为 3 或 5 个类别（Bo等人，2021年）。在 Viard 等人（2020年）的研究中，标签是在维基百科之外定义的。然而，这些数据集都没有利用由维基百科编辑者达成共识后形成的主题分类，而我们正是使用维基百科的关键文章列表[3] 来定义 *WikiVitals* 数据集（第 4.1 节）。此外，该分类并不平衡，其中一些类别仅包含极少数代表，因此构成了一个困难的分类问题。

维基百科数据集（Squirrel、Chameleon 以及 *WikiVitals*）的一个共同特点是，它们的异配性比经典图数据集更强（Newman，2003年）[4]。这使它们特别适合作为节点分类任务的基准，因为 GCN 等基础模型在此类异配性情境中表现不佳（Bo 等人，2021年）。为解决这一问题，研究者提出了一些新模型，例如 H_2GCN 和 FAGCN，它们在这类情境中表现更好（Zhu等人，2020年；Bo等人，2021年）。我们对 GCN 和 FAGCN 都进行了实验（第 4.2 节）。

3 将公平比较方法应用于图马尔可夫神经网络

3.1 GMNN训练

为说明我们如何在下一小节中对图马尔可夫神经网络进行严格评估，我们先简要回顾这些模型的定义和训练方式。详细介绍请参见 Qu 等人（2019）。

我们考虑一个图 $G = (V,\ E,\ \mathbf{x}_V)$，其中 V表示节点集合，E表示边集合，而 $\mathbf{x}_V := \{\mathbf{x}_n\}_{n \in V}$表示与每个节点关联的特征集合 n。假设已知子集 $L \subset V$ 中节点的独热编码标签（对应 K 个类别）$\mathbf{y}_L := \{\mathbf{y}_n\}_{n \in L}$，以及所有节点的特征 $\mathbf{x}_V$。我们要预测 $U = V \setminus L$中其余未标记节点的标签 $\mathbf{y}_U$。图马尔可夫神经网络模型包含两项内容。首先，它定义了联合概率 $p_\phi(\mathbf{y}_L,\ \mathbf{y}_U|\mathbf{x}_V)$ 的模型，以刻画邻居节点之间的相关性。

[2] http://snap.stanford.edu/data/wikipedia-article-networks.html [3] https://en.wikipedia.org/wiki/Wikipedia:Vital_articles/Level/5 [4] 异配性这一概念也很常用，指的是大多数相邻 n' 节点 n 的标签 $\mathbf{y}_n$ 都与$\mathbf{y}_{n'}$ 不同$\mathbf{y}_n$。

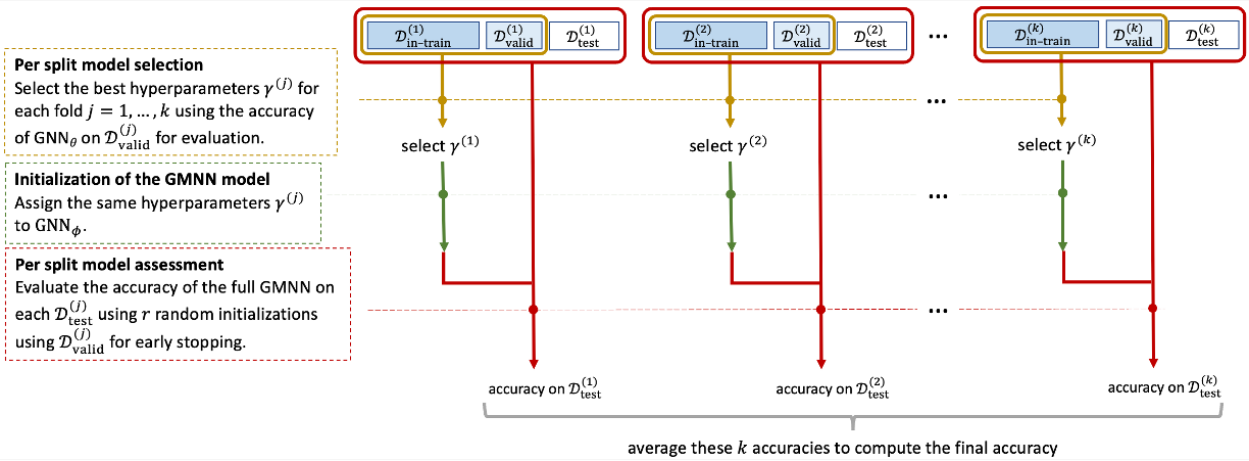


图 1：GNN 的公平评估流程及其针对图马尔可夫神经网络的调整，使用 k 训练/验证/测试划分$\mathcal{D}^{(i)}$ 其中–训练集 $\mathcal{D}^{(i)}$ 验证集$\mathcal{D}^{(i)}$ 测试集，这些划分由 k 分层折 $\mathcal{F}_i$生成。

其次，它描述了一种基于 EM算法的实用训练过程，用于求得参数 ϕ，使其最大化观测标签上的边际似然 $p_\phi(\mathbf{y}_L|\mathbf{x}_V)$ 的变分下界。

训练图马尔可夫神经网络需要定义两个常规图神经网络。第一个记为 GNN$_\phi$，其中 ϕ 是其参数集合，用于描述给定邻居节点的标签（记为 NB(n)）和节点特征 $\mathbf{x}_V$时，各个节点标签 $\mathbf{y}_n$ 的分布 $p_\phi(\mathbf{y}_n|\mathbf{y}^{\text{NB}(n)}, \mathbf{x}_V)$。其定义方式与通常相同：对从 GNN$_\phi$最后一层读出的 d 维节点嵌入 $\mathbf{h}_{\phi,\ n}$应用 Softmax，再与一个 $K \times d$ 可学习矩阵W_ϕ相乘

$$p_\phi(\mathbf{y}_n|\mathbf{y}_{\text{NB}(n)}, \mathbf{x}_V) = \text{Cat}(\mathbf{y}_n|\text{softmax}(W_\phi \mathbf{h}_{\phi,n})). \tag{1}$$

第二个图神经网络记为 GNN$_\theta$，它定义了一个平均场变分分布，用于在 EM算法中近似后验分布 $p_\phi(\mathbf{y}_U|\mathbf{y}_L,\ \mathbf{x}_V)$。它以节点为单位，以类似的方式定义

$$q_\theta(\mathbf{y}_n|\mathbf{x}_V) = \text{Cat}(\mathbf{y}_n|\text{softmax}(W_\theta \mathbf{h}_{\theta,n})). \tag{2}$$

直观而言，图神经网络$_\theta$ 是一种完全忽略标签相关性的模型。随后，图神经网络$_\phi$ 会考虑邻居节点标签之间的相关性，并据此调整这些预测；这些调整又会在 EM循环训练过程中修正图神经网络$_\theta$。训练过程使用以下两个目标函数。其中一个用于更新 θ，同时固定 ϕ:

$$O_\theta = \sum_{n \in U} \mathbb{E}_{p_\phi(\mathbf{y}_n|\hat{\mathbf{y}}_{\text{NB}(n)}, \mathbf{x}_V)}[\log q_\theta(\mathbf{y}_n|\mathbf{x}_V)] + \sum_{n \in L} \log q_\theta(\mathbf{y}_n|\mathbf{x}_V), \tag{3}$$

其中，$\hat{}\mathbf{y}_n$ 表示真实标签$\mathbf{y}_n$ （若 $n \in L$），若 $n \in U$，则从 $q_\theta(\mathbf{y}_n|\mathbf{x}_V)$中采样。使用相同的记号，另一个用于优化 ϕ、同时固定 θ 的目标函数为：

$$O_\phi = \sum_{n \in V} \log p_\phi(\hat{\mathbf{y}}_n|\hat{\mathbf{y}}_{\text{NB}(n)}, \mathbf{x}_V). \tag{4}$$

训练图马尔可夫神经网络的第一步是通过最大化 θ对应的式（3）中的最后一项来初始化 q_θ。这相当于训练一个不考虑标签相关性的普通图神经网络。因此，这个初始 q_θ 模型的准确率可作为与完整图马尔可夫神经网络模型比较的基线。第二步，固定 θ 并优化式（4）中的 ϕ，即 M步。最后，在固定 ϕ 的情况下，为 θ 优化式（3），即 E 步。重复 M 步和 E 步，直至收敛。经验表明，q_θ 始终能比 p_ϕ 得到更好的预测器（Qu等人，2019）。

3.2 图马尔可夫神经网络的公平比较

回顾一下我们的主要目标。首先，我们希望确定，在像*WikiVitals*这样的大型真实世界数据集上应用时，使用图马尔可夫神经网络模型考虑标签相关性，是否能以具有统计显著性的方式提高预测器的准确率。因此，我们详细说明如何将 Errica 等人（2020）的方法调整为图马尔可夫神经网络，以进行这项严格的评估。对于同一数据集，我们还希望比较像多层感知机这样的结构无关基线与

利用图结构但不考虑标签相关性的常规图神经网络（如图卷积网络或 FAGCN）。对于这些更简单的模型，我们直接采用其方法，不作修改。

一种粗略的评估方法是将带标签文章的数据集 $\mathcal{D} = ((\mathbf{x}_1, \mathbf{y}_1) \ldots, (\mathbf{x}_N, \mathbf{y}_N))$ 划分为三个互不重叠的集合：训练集 $\mathcal{D}_{\text{train}}$、用于选择最优超参数的验证集 $\mathcal{D}_{\text{valid}}$ γ^*，以及用于评估该最优模型准确率的测试集 $\mathcal{D}_{\text{test}}$。遗憾的是，研究表明，这种简单的方法极不稳定，改变数据划分就可能彻底打乱各种图神经网络架构的相对排名（Shchur等人，2018年；Errica等人，2020年）。

可靠评估的主要要求是明确区分模型选择与模型评估。为此，我们按常规方式划分数据集 $\mathcal{D}$，使用 k 互不重叠的分层[5] 折 $\mathcal{F}_1, \ldots, \mathcal{F}_k$。然后，定义 k 组不同的训练集和测试集，如下所示：

$$\mathcal{D}_{\text{train}}^{(i)} := \bigcup_{j \neq i} \mathcal{F}_j, \quad \mathcal{D}_{\text{test}}^{(i)} := \mathcal{F}_i, \quad i = 1, \ldots, k.$$

每个训练集又被划分为一个内部训练集和一个验证集：

$$\mathcal{D}_{\text{train}}^{(i)} := \mathcal{D}_{\text{in-train}}^{(i)} \cup \mathcal{D}_{\text{valid}}^{(i)}, \quad i = 1, \ldots, k.$$

图 1 展示了适用于图马尔可夫神经网络的公平评估流程。

模型选择是指选择一组最优超参数。与标准的 k折交叉验证流程不同，我们针对每个 $\mathcal{D}^{(i)}$ 训练集分别进行选择。与 Qu 等人（2019）一样，我们优化平均场网络图神经网络$_\theta$ 的超参数，然后将这些相同的最优值用作另一个编码标签相关性的图神经网络$_\phi$ 网络的超参数。更确切地说，图神经网络$_\theta$ 模型在 $\mathcal{D}^{(i)}$ in−训练集上进行训练，并使用$\mathcal{D}^{(i)}$ 验证集作为留出集，选择使损失 $\gamma^{(i)}$ 最小的超参数。

模型评估会使用数据划分 i的最优超参数 $\gamma^{(i)}$，分别在每个测试集 $\mathcal{D}^{(i)}$ 测试集上评估整个图马尔可夫神经网络（由图神经网络$_\theta$ 和图神经网络$_\phi$组成）的准确率。最终的准确率评估是这些结果在 k 个数据划分上的平均值。更确切地说，完整的图马尔可夫神经网络模型在每个 $\mathcal{D}^{(i)}$ in−训练集上使用 EM算法进行训练，使用 $\mathcal{D}^{(i)}$ 验证集进行提前停止，然后在 $\mathcal{D}^{(i)}$ 测试集上评估。

请注意，在完成图神经网络$_\theta$ 的初始训练后、进入 EM 优化之前对模型进行公平评估，相当于对普通图神经网络进行公平评估，因此不需要额外计算。

4 实验

4.1 创建*WikiVitals*

我们从截至 2022年4月的完整维基百科转储中提取了 48512 篇重要维基百科条目，并据此构建了 *WikiVitals* 异配文档网络。节点对应重要维基百科条目。节点特征是文章的稀疏二值词袋表示。这些表示中的 4000 个特征分别对应文章引言、标题或章节标题中某个信息性一元词或二元词的存在或缺失。边对应语料库中重要条目之间的超链接。

重要条目由维基百科编辑者选定，并按主题进行分类。我们提取了一个 3 层主题层级，并将该层级中的 32 个中间主题用作分配给图中每个节点的标签。每个节点都被分配了一个单独的标签。表 1 展示了主题层级的部分视图，重点介绍了我们在本文中使用的 32 个中间类别[6]。我们采用了英文维基百科的转储，因为它提供了英文维基百科的冻结视图，并且包含文章文本。

我们对重要条目的摘要、标题和页眉进行了特征提取。为预处理这些数据，我们使用 nltk.corpus 忽略停用词，并使用来自 nltk 的 Snowballstemmer 进行词干提取。接下来，我们从摘要和页眉中提取频率大于 1×10^{-3} 的一元词组和二元词组，并从标题中提取频率大于 1×10^{-4} 的一元词组和二元词组。最后，我们保留了在卡方意义上对标签最具预测性的前 4000 个特征。

[5]在本文中，节点的分层抽样是指抽样集合中的标签分布与其所抽取自的节点集合中的标签分布相同。[6]层级结构的顶层包含 11 个粗粒度主题，中间层包含 32 个主题，最细层包含 251 个主题。

类别名称	#文章
Arts	
01-艺术	3310
生物与健康科学	
02-动物	2396
03-生物学	886
04-健康	791
05-植物	608
日常生活	
06-日常生活	1191
07-体育、游戏与娱乐	1231
地理	
08-城市	2030
09-国家	1386
10-自然地理	1902
历史	
11-历史	2979
数学	
12-数学	1126
人物	
13-艺术家、音乐家和作曲家	2310
14-演艺人员、导演、制片人和编剧	2342
15-军人、革命者和活动人士	1012
16-其他	1186
17-哲学家、历史学家、政治学家和社会科学家	1335
18-政治家和领导人	2452
19-宗教人物	500
20-科学家 ，发明家 ，和数学家	1108
21-体育人物	1210
22-作家和记者	2120
哲学与宗教	
23-哲学与宗教	1408
物理科学	
24-天文学	886
25-基础与测量	360
26-化学	1207
27-地球科学	849
28-物理学	988
社会与社会科学	
29-文化	2075
30-政治与经济	1825
31-社会研究	355
技术	
32-技术	3148

表1：按主题分类的*WikiVitals*数据集所使用的32个节点标签。

4.2 数据集与设置

数据集： 除了 WikiVitals 数据集外，我们还对三个知名的同配引用网络数据集进行了全面评估：Cora、Citeseer 和 Pubmed。这些网络中的边表示两篇科学文章之间的引用关系，节点特征 $\mathbf{x}_n$ 是文章的词袋向量，标签 $\mathbf{y}_n$ 对应文章所属的领域。对于所有数据集，我们都将图视为无向图。这些数据集的统计信息如表 2 所示。

训练细节： 所有基线模型（多层感知机、图卷积网络和 FAGCN）均使用 PyTorch 重新实现，并包含两层（输入表示 $\rightarrow$ 隐藏层 $\rightarrow$ 输出层）。对于所有模型，所有层均进行 L^2 正则化，并在输入数据和所有层上应用丢弃。对于图卷积网络和 FAGCN，我们使用了邻接矩阵的所谓重归一化技巧（Kipf 和 Welling，2017年）。对于 FAGCN，传播次数（Bo 等人，2021年）设为 2，以将信息聚合限制在与节点距离最多为 2 的节点范围内。对于图马尔可夫神经网络，我们使用退火采样方法，因子设为 0.1（Qu 等人，2019），EM 循环次数设为 10。为了训练

数据集	同配性	节点数	边数	类别数	#特征
Cora	0.771	2,708	5,429	7	1,433
Citeseer	0.675	3,327	4,732	6	3,703
Pubmed	0.686	19,717	44,338	3	500
WikiVitals	0.204	48,512	2,297,782	32	4,000

表 2：文档图的统计信息

图神经网络ϕ 网络，我们同时使用邻居节点 $n' \in \text{NB}(n)$ 的标签 $\hat{\mathbf{y}}_{n'}$ 和式（4）中定义的节点特征 $\mathbf{x}_V$ 来预测$\mathbf{y}_n$。

所有模型均采用相同的训练流程。对于所有数据集，训练前先将节点特征二值化，然后进行归一化（单位 L^1-范数）。我们使用 Adam优化器（Kingma和Ba，2015年），采用默认参数且不进行学习率衰减；所有数据集使用相同的最大训练轮次、早停准则和耐心超参数（更多详情见附录A）。每轮结束时评估用于提前停止的验证损失。所有模型参数（FAGCN 的卷积核系数以及所有模型的权重矩阵）均同时初始化并优化（权重按照 Glorot 方法初始化，偏置初始化为零）。所有情况下，我们均采用全批次训练（每轮使用训练集中的所有节点）。

公平的评估设置：在评估阶段，我们使用不同的随机权重初始化执行 $r = 20$ 次训练，以减轻可能较差的配置所带来的影响。
对于每个数据集，我们遵循 Errica 等人（2020）提出的最佳实践，预先计算整个节点集的分层数据划分（$\mathcal{D}^{(j)}_{\text{在–训练}}\, \mathcal{D}^{(j)}_{\text{验证}}\, \mathcal{D}^{(j)}_{\text{测试}}$) $j = 1, \ldots, k$，对应比例分别为 81%、9% 和 10%。下文将集合 $\mathcal{D}^{(j)}_{\text{在–训练}}$ 称为密集训练集。
为了便于与现有研究进行比较，我们另外创建了两组数据划分，其训练集较为稀疏。需要指出的是，针对 Cora、Citeseer 和 Pubmed 的图神经网络以及图马尔可夫神经网络评估，传统上采用 Planetoid 划分（Yang 等人，2016年），或采用类似构造的稀疏划分：从整个数据集中随机选取每个类别 20 个节点（Shchur等人，2018年；Qu等人，2019；Bo等人，2021年）。为构造稀疏训练集的数据划分，我们从每个$\mathcal{D}^{(j)}_{\text{在–训练}}$ $j = 1, \ldots, k$中分别独立提取了两个子集：$\mathcal{D}^{(j)}_{\text{稀疏–平衡}}$ 和 $\mathcal{D}^{(j)}_{\text{稀疏–分层}}$。每个子集包含 $20 \times K$ 个节点（其中 K 为类别数）。

每个 $\mathcal{D}^{(j)}_{\text{稀疏–均衡训练集}}$的构建方法是：从 $\mathcal{D}^{(j)}_{\text{训练集–}}$的每个类别中选取 20 个节点。下文将这些集合称为稀疏均衡训练集，因为每个集合中的各类别所占比例相同。
每个 $\mathcal{D}^{(j)}_{\text{稀疏–分层训练集}}$的构建方法是：以分层方式从 $\mathcal{D}^{(j)}_{\text{训练集–}}$中选取节点。我们将这些集合称为稀疏分层训练集。
严格的模型选择阶段需要在超参数集合上进行大量网格搜索，这在计算上非常昂贵。因此，我们实现了自己的进化网格搜索算法，通过利用验证准确率引导进化过程来寻找合适的超参数。此类算法通过探索超参数集合中的一小部分来计算出合适的配置（Young 等，2015年）。

5 结果

针对我们的 *WikiVitals* 数据集以及经典的 Cora、Citeseer 和 Pubmed 数据集执行节点分类任务所得的定量结果，分别列于稀疏训练集对应的表3和密集训练集对应的表4中。报告的准确率是 GMNN 模型在 EM 阶段前后、基于 k 次数据划分计算的平均值。每项平均准确率均使用 r 种不同的初始化方式计算，以估计其方差；方差数值以括号形式列于表中。我们所说的准确率提升具有显著性，是指其显著性水平优于 $p < 0.05$ [7]。表3显示，在稀疏训练集中，分层抽样与均衡抽样的结果几乎相同。

[7]我们进行了 t-检验，备择假设为 EM 阶段后的准确率高于阶段前的准确率。

	WikiVitals		Cora		Citeseer		Pubmed	
	平衡	分层	平衡	分层	平衡	分层	平衡	分层
MLP	68.60 (0.92)		58.54 (3.98)	58.32(2.17)	59.84 (3.54)	58.35 (2.48)	71.23 (2.85)	70.39 (1.70)
图神经网络（			.7)	80.78 (2.58)	81.31(2.16)	69.05 (3.66)	70.94 (2.16)	80.20 (1.88)	80.50 (2.38)
+ 图马尔可夫神经网络	**74.80 (1.18)**		**81.67 (3.00)**	**81.91(2.21)**	**69.61 (3.96)**	**71.62 (2.20)**	**81.67 (1.32)**	**81.70 (2.45)**

表3：稀疏训练集：测试准确率为 k 个数据划分的平均值，以百分比表示。最佳结果已突出显示。对于 *WikiVitals*，基础图神经网络为 FAGCN；对于 Cora、Citeseer 和 Pubmed，基础图神经网络为图卷积网络。

	WikiVitals	Cora	Citeseer	Pubmed
MLP	86.55 (0.42)	78.49 (2.39)	75.02 (2.15)	88.68 (0.86)
GCN	72.74 (0.61)	88.84 (2.39)	77.24 (1.73)	89.20 (0.86)
+ 图马尔可夫神经网络	74.19 (0.42)	**89.26 (1.91)**	*77.43 (1.70)*	*89.18 (0.84)*
FAGCN	87.84（0.32）	88.87（1.99）	78.27（3.53）	90.23（0.90）
+ 图马尔可夫神经网络	**87.92（0.31）**	89.08 (1.76)	***78.32 (3.64)***	**90.34 (0.88)**

表4：密集训练集：在文中报告的 k 个数据划分上的平均测试准确率。最佳结果以粗体标出。斜体表示考虑标签相关性后没有显著提升的情况。

5.1 图结构的贡献

表3和表4中的结果证实，无论训练集是密集还是稀疏，考虑图结构都能显著提升所有数据集上的节点分类任务表现。

不过需要注意的是，由于 *WikiVitals* 是一个异配性数据集，因此表现最好的是 FAGCN 模型。实际上，在*WikiVitals* 上使用密集训练集时，简单的多层感知机比基础图卷积网络表现更好，这进一步证实，选择与图的同配性程度相适应的架构十分重要。

对于经典数据集 而言，GCN 和 FAGCN 模型都优于仅使用多层感知机的做法，而后者只考虑节点 特征，而忽略图结构。

5.2 标签依赖关系的贡献

首先可以观察到，无论训练集稀疏还是稠密，利用标签依赖关系对 *WikiVitals* 都很重要，尽管其影响在稀疏训练集上的效果远强于在稠密训练集上的效果。我们认为，这是因为标签间相关性所提供的信息在稀疏情况下更有用，可以弥补可用于训练的节点数量较少这一不足。另一方面，当训练集稠密时，用于准确预测的信息更可能来自大量节点的特征。

利用标签依赖关系也显著提升了 Cora 在稀疏和密集训练集上的表现。对于密集情形，Pubmed 只有在使用 FAGCN 时准确率才明显提高，而 Citeseer 的准确率没有显著提升。对于那些使用 GMNN 将标签相关性纳入考量后未取得显著提升的情形，表4中以斜体标示。

6 结论与展望

本文介绍了一个名为*WikiVitals* 的大型异配性文档间图数据集，并将 GNN 与 GMNN 的公平比较方法相结合，以评估半监督节点分类任务中三种不同信息来源的作用：编码文档内容的节点特征、底层图结构以及标签相关性。

考虑标签相关性是我们的主要关注点，因为这是最微妙的影响。我们确实证实，对于 *WikiVitals* ，无论整个图中可用于训练的节点是稀疏还是密集，这些相关性都能在分类准确率方面带来统计显著性提升。不过，稀疏训练集的准确率提升幅度要大得多。对于经典基准数据集 Cora、Citeseer 和 Pubmed，稀疏训练集上的提升具有统计显著性，这证实了已有研究结果；但在密集训练集上，这种提升则不那么明显。

未来研究的一个有趣方向是探讨：如果我们恰当地定义并考虑这些更丰富的标签结构中的相关性，层次分类和多标签分类是否能够得到改进。例如，这将使我们能够处理一些模糊情形：某些文档被分配了多个标签，这在实践中很常见。

更广泛影响声明

我们创建的 *WikiVitals* 可用于任何关于文档大型图的研究。我们预计不会产生任何负面的社会影响。

致谢与资金披露

Thi这项工作使用了 GENCI–IDRIS 的高性能计算资源（资助项目 2021-AD011013266） .

参考文献

Vincent D Blondel、Jean-Loup Guillaume、Renaud Lambiotte 和 Etienne Lefebvre。《在大型网络中快速发现社群》。统计力学杂志：理论与实验，2008(10):P10008，2008年。

Deyu Bo、Xiao Wang、Chuan Shi 和 Huawei Shen。《超越图卷积网络中的低频信息》。载于*AAAI*人工智能大会论文集，第35卷，第3950–3957页，2021年。

Karsten M Borgwardt、Cheng Soon Ong、StefanSchönauer、SVN Vishwanathan、Alex J Smola 和 Hans-Peter Kriegel。《通过图核预测蛋白质功能》。生物信息学，21(suppl_1):i47–i56，2005年。

Michaël Defferrard、Xavier Bresson 和 Pierre Vandergheynst。《基于快速局部谱滤波的图卷积神经网络》。神经信息处理系统进展，第29卷，2016年。

Paul D Dobson 和 Andrew J Doig。无需比对即可区分酶结构与非酶。分子生物学杂志，330(4):771–783，2003年。

Federico Errica、Marco Podda、Davide Bacciu 和 Alessio Micheli。图分类中图神经网络的公平比较。发表于*ICLR*，2020年。

Qian Huang、Horace He、Abhay Singh、Ser-Nam Lim 和 Austin R Benson。结合标签传播与简单模型，效果优于图神经网络。发表于*ICLR*，2021年。已故

erik P Kingma 和 Jimmy Ba。Adam：一种随机优化方法。发表于*ICLR*，2015年。

Thomas N Kipf 和 Max Welling。使用图卷积网络进行半监督分类。2017年。*arXiv abs/1609.02907*，2017年。

Johannes Klicpera、Aleksandar Bojchevski 和 Stephan Günnemann。先预测再传播：将神经网络与个性化 PageRank 结合，用于图分类。发表于*ICLR*，2018年。

Federico Monti、Davide Boscaini、Jonathan Masci、Emanuele Rodola、Jan Svoboda 和 Michael M Bronstein。使用混合模型卷积神经网络在图和流形上进行几何深度学习。载于电气与电子工程师协会计算机视觉与模式识别会议论文集，第 5115–5124 页，2017年。

Mark EJ Newman。网络中的混合模式。《物理评论*E*》，67(2)：026126，2003年。

Mark EJ Newman。基于随机游走的介数中心性度量。《社会网络》，27(1)：39–54，2005年。

Meng Qu、Yoshua Bengio 和 Jian Tang。图马尔可夫神经网络：图马尔可夫神经网络。载于*ICML*，第 5241–5250 页。PMLR 97，2019年。

Benedek Rozemberczki、Carl Allen 和 Rik Sarkar。多尺度属性节点嵌入。《复杂网络杂志》，9(2)，2021年。ISSN 2051-1329。doi：10.1093/comnet/cnab014。URL https://doi.org/10.1093/ comnet/cnab014。cnab014。

Prithviraj Sen、Galileo Namata、Mustafa Bilgic、Lise Getoor、Brian Galligher 和 Tina Eliassi-Rad。《网络数据中的集体分类》。《人工智能杂志》，29(3):93–93，2008。

Oleksandr Shchur、Maximilian Mumme、Aleksandar Bojchevski 和 Stephan Günnemann。《图神经网络评估中的陷阱》。载于关系表示学习研讨会论文集（*R2L 2018*），*NeurIPS 2018*，2018年。

Petar Veli˘ckovi´c、Guillem Cucurull、Arantxa Casanova、Adriana Romero、Pietro Liò 和 Yoshua Bengio。《图注意力网络》。载于*ICLR*，2018年。Tiphaine Viard、Thomas McLachlan、Hamidreza Ghader 和 Satoshi Sekine。《以细粒度层级对维基百科进行分类：图能带来什么》。*CoRR*，abs/2001.07558，2020年。URL https://arxiv.org/abs/2001. 07558。Zhilin Yang、William Cohen 和 Ruslan Salakhudinov。《借助图嵌入重新审视半监督学习》。载于*ICML*，第40–48页。PMLR 48，2016年。Steven R Young、Derek C Rose、Thomas P Karnowski、Seung-Hwan Lim 和 Robert M Patton。《通过进化算法优化深度学习超参数》。载于*MLHPC '15*：机器学习与高性能计算环境研讨会论文集，第1–5页，2015年。Dengyong Zhou、Olivier Bousquet、Thomas Lal、Jason Weston 和 Bernhard Schölkopf。《基于局部与全局一致性的学习》。*NeurIPS2003*，第16卷，2003年。Jiong Zhu、Yujun Yan、Lingxiao Zhao、Mark Heimann、Leman Akoglu 和 Danai Koutra。《超越图神经网络中的同配性：当前局限与有效设计》。*NeurIPS 2020*，第33卷，2020年。Jiong Zhu、Ryan A Rossi、Anup Rao、Tung Mai、Nedim Lipka、Nesreen K Ahmed 和 Danai Koutra。《具有异配性的图神经网络》。载于*AAAI*人工智能大会论文集，第35卷，第11168–11176页，2021年。Xiaojin Jerry Zhu。《半监督学习文献综述》。技术报告，威斯康星大学麦迪逊分校计算机科学系，2005年。

A 超参数

超参数和搜索空间：模型选择期间通过网格搜索在以下搜索空间 Γ 中进行：

- 隐藏维度： $[8, 16, 32, 64]$
- 输入丢弃： $[0.2, 0.4, 0.6, 0.8]$
- 丢弃： $[0.2, 0.4, 0.6, 0.8]$
- 学习率： [1e-1, 5e-2, 1e-2, 5e-3, 1e-3, 5e-4, 1e-4]
- L^2-正则化强度： [1e-1, 5e-2, 1e-2, 5e-3, 1e-3, 5e-4, 1e-4, 5e-5, 1e-5]

(仅适用于 FAGCN)： $[0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9]$

对于 *WikiVitals*，我们采用了缩减后的搜索空间，将隐藏维度设为 64，将 L^2-正则化强度设为 1e-5，学习率取值范围为 [1e-1, 5e-2], ，并且 ϵ 取值范围为 $[0.7, 0.8, 0.9]$。

图神经网络模型的训练流程：对于所有图神经网络模型的训练：

- 我们最多训练 1000 个训练轮次。
- 我们采用提前停止，耐心值设为 200。
- 不使用学习率衰减。
- 所有层都应用 L^2 正则化。
- 所有模型参数（FAGCN 的卷积核系数以及所有模型的权重矩阵）均同时进行优化。

- 训练停止后，我们将模型参数的状态重置为验证损失最低时对应的步骤。

对于多层感知机和图卷积网络，我们采用提前停止：如果验证损失连续 200 个训练轮次未下降，就停止优化。对于 FAGCN，如果验证损失和验证准确率连续 200 个训练轮次未下降，我们就停止优化。训练图马尔可夫神经网络时，我们训练模型 100 个训练轮次，并执行 10 个 EM 循环。

进化网格搜索：我们的进化算法会在各代中维护一个由 100 个超参数配置组成的随机初始化种群。每一代中，我们会保留验证准确率高于种群平均值的 2 至 50 个配置，用于下一代。通过双枢轴交叉生成新配置：评估表现较好的配置被选中的概率更高。变异步骤以 0.05 的概率为某个配置的超参数分配一个新值，以促进探索。每一代中，只有此前从未出现过的配置才会被加入种群，以补足种群规模。代数设为 10；超过这一数值后，最佳配置的评估表现不再显著提升。