CS3319-02 Project (2023 Spring)

将图算法和图神经网络应用于现实世界问题的大作业项目。

在这个项目中，你需要从2项作业中选择1项作业（同时进行这两项作业也可以）。

Assignment 1. GNN over Recommendation Senario

Introduction

GNN在异质图和同构图的节点分类和链路预测任务中表现非常好。在很多推荐场景中，利用用户之间的信任关系和产品的本质属性，结合用户的购买记录，使用GNN生成的表征也能达到很好的效果。同时，在学术平台中，合作者推荐、论文推荐、期刊和会议的审稿人推荐是主要任务，GNN在这些任务上的表现依然良好。

对于上述问题，都有相对较好的基准指标和模型。但是，在论文推荐的场景中，如果是通过合作者和论文所在的领域和合作社群进行科学研究的学者，引文推荐仍然缺乏更好的数据集和模型。本选题提供的数据集用在一个学术推荐系统中，其中"用户"为学术论文的作者，"产品"为学者的论著。当用户和产品都有关联网络时，我们可以提取用户群体和产品类别的群体特征，基于产品之间的关系和用户之间的联系，为用户提供更好、更多样的推荐，这可能有助于解决推荐系统冷启动造成的用户行为较少的问题。

由此，你需要解决一个学术网络中的推荐问题。我们从地理科学领域的顶级期刊中收集了6,611位作者和相应的79,937篇论文，以及他们出版物的引文信息。你需要利用收集到的信息形成一个学术网络，这里有一个可行的方法：

建立一个*异质*网络，其中包含两类节点，一类节点代表作者，另一类代表论文。在这个网络中，作者节点和论文节点之间的每条边都表示作者阅读过该论文（连接作者和被作者所写的论文所引用的论文），两个作者节点之间的每条边表示合著关系，两个论文节点之间的每条有向边表示引用关系。

你可以使用其他方式来构建学术网络。请注意，我们所提供的是迄今为止的作者和论文信息，它揭示了不同作者的工作之间的关联性。因此，假设你正在设计一个学术阅读推荐系统，你需要挑选出与作者以前研究相关的论文。这个问题可以被模拟成一个链路预测问题，你的任务是根据所提供的信息预测测试集中的每个作者-论文对。如果该论文被推荐给作者，则标记为1，否则标记为0。

Data

具体的数据说明请参见 `Data Description.md`。

Required Files

- 算法的设计报告，以会议论文的形式呈现。
- 最终用于性能评估的代码（与Kaggle最终决定用于测试的代码一致）

References

1. 【arXiv 2020】 Graph neural networks in recommender systems: a survey. paper
2. 【arXiv 2021】 Graph learning based recommender systems: A review. paper
3. 【KDD 2018】 Graph Convolutional Matrix Completion (GC-MC). paper

4. 【KDD 2018】 Graph convolutional neural networks for web-scale recommender systems (PinSage). paper
5. 【RecSys 2018】 Spectral collaborative fltering (SpectralCF). paper
6. 【SIGIR 2019】 Neural graph collaborative filtering (NGCF). paper
7. 【SIGIR 2020】 Lightgcn: Simplifying and powering graph convolution network for recommendation (LightGCN). paper
8. 【SIGIR 2019】 A neural influence diffusion model for social recommendation (DiffNet). paper
9. 【WWW 2019】 Graph neural networks for social recommendation (GraphRec). paper
10. 【WWW 2019】 Dual graph attention networks for deep latent representation of multifaceted social effects in recommender systems (DANSER). paper
11. 【RecSys 2019】 Deep social collaborative filtering (DSCF). paper

Attention

该项目将在**Kaggle**平台上进行。你需要在Kaggle上提交结果以参与性能评估和排名，同时你需要在Canvas上提交其他材料。关于项目细节、数据格式、评估方法等，请查看Kaggle竞赛页面。

注意：大作业为小组形式，在组队完成后，不要忘记在Canvas上登记并加入小组。

为了公平起见，我们制定了一些规则------违反规则将导致额外的扣分：

1. 请不要复制别人的代码。我们会在提交后进行抄袭检查。
2. 请不要在其他地方下载数据集来训练你的模型。我们已经对数据集进行了随机打乱，并将使用该数据集重现你的实验结果。如果你报告的结果和我们重现的结果之间有很大的差距，将被视为违规。
3. 请不要使用预先训练好的模型。

Other references

图机器学习的最新成果和性能最先进的模型在这可以找到：

1. OGB leaderboard
2. 机器学习/数据挖掘领域顶级国际会议： 要找到与图机器学习相关的论文，你可以在对应的会议网站中搜索 "graph" 这个词。

- ICML 2021: https://proceedings.mlr.press/v139/
- NeurIPS 2022: https://papers.nips.cc/paper/2022
- ICLR 2022: https://openreview.net/group?id=ICLR.cc/2022/Conference
- KDD 2022: https://kdd.org/kdd2022/toc.html

Assignment 2. Explaining Graph & GNNs

Introduction

通过课程我们已经简单的了解了图和图神经网络（GNN）。近年来，GNN在多种基于图的分析任务中取得了亮眼的成绩，但是GNN模型本身仍是一个黑箱模型，缺乏可靠的理论解释。因此，为了保证这个强大工具的正确性和可靠性，理解图，理解GNN如何工作，就具有非常重要的意义。

目前解释GNN的主流方式有：

- 对于模型本身进行解释
- 通过提出图上拓扑相关的指标衡量各节点/边的重要性，对指标进行分析从而解释模型
- 通过某种GNN explainer解释（GNNExplainer，PGMExplainer等）

在学术网络分析中，有许多问题值得我们关注。从结构上看，网络模型分为常规网络、随机网络、小世界网络和无标度网络。我们一般认为，学者网络是一个无标度网络。面对学者之间的合作关系，我们通常用聚类系数来衡量。通过计算一个点的邻居节点之间的互联程度，我们可以知道学者之间的相互了解程度。我们可以用这个指标来了解学者之间合作的密切程度。与此相关的是网络的稀疏度，不同领域的合作网络具有不同的稀疏度。

接下来，我们可以通过不同的模式进一步分析网络的拓扑结构。有些合作模式是放射状的，也就是一个强势的学者带领一些弱势的学者，有些合作模式是链状的，代表学者。他们互不相识，但却能得到他们需要的东西。还有一种合作模式是两个径向连接的，依靠两个有影响力的学者合作，把两个大团队结合起来。这些与两个最初不知名的学者相遇并合作的学者被认定为关键性学者，这些学者在科学研究中发挥着至关重要的作用。

我们有时可以认为，学者合作网络是一个小世界网络。这样的网络往往有较小的平均距离和较明显的聚类系数。这些指标都是用来衡量一个社区网络的，但是其中的某个（或某些）节点对这个网络的连通性有多大影响呢？目前，只有中心性指标可以衡量，一般有如下几个指标：

- 间性中心度（betweenness centrality，在网络中所有节点对的最短路径中，经过一个节点的最短路径越短，这个节点的影响力越大）
- 接近性中心度（closeness centrality，一个节点与网络中其他节点的平均距离，越是小，这个节点的接近性就越大）
- 特征向量中心线（eigenvector centerline，一个节点的重要性取决于其邻居的数量（即节点的程度）和每个邻居的重要性）
- 程度中心性（degree centrality，一个节点的邻居越多，它就越重要）
- 半本地中心性（semi-local centrality，计算二阶邻居的数量）

然而，这些指标忽略了学者本身的属性。因此，用以上指标识别关键学者是远远不够的,需要更全面的指标对学术网络中的学者及其合作关系进行分析评估。

Task

请从以下几个角度自行设计指标，选择一种或几种GNN算法以及GNN explainer，并选取合适的例子来解释你的工作。

- 评价图（自行设计图上指标）
    - 重要的点，根据重要性排序，并且根据节点的特征和拓扑属性，以及其对网络连通性的影响来分析节点的重要性来源
    - 重要的边，根据重要性排序，并且根据边的特征和所连接节点的拓扑属性，以及其对网络连通性的影响来分析边的重要性来源

可以参考的节点分类分析方式：

1. 网络的核心，对应学术大牛等学者
2. 网络桥梁节点，对应跨领域学者，连接若干个学术社群
3. 网络边缘节点

可以参考的边分类分析方式：

1. 社区内边，往往是同一研究领域、同一地域内学者的联系
2. 社区间桥梁，一般为跨领域、跨国的学术合作关系

- 从以下几种GNN算法中选择一种及以上进行实现，任务场景限定为节点预测
    - 2层 GCN
    - 2层 GraphSage
    - 2层 GAT
    - 2层 lightGCN
    - 2层 GIN

  注：以上GNN算法均可借助 dgl 库进行模块化构建。

- 评价GNN（根据选择的GNN算法，从GNN本身结构和设计思路解释，或设计实现GNN explainer）
    - 对某个预测值，哪些点更重要，为什么？
    - 对某个预测值，哪些边更重要，为什么？

Data

本选题提供一个完整的学术网络数据集，给出的数据包含：

1. 一个学者的机构、国家和学科，从中我们可以知道一个学者的基本属性。
2. 当两位学者来自同一机构，且属于同一学科时，其合作情况很可能是群体内的合作。
3. 当两位学者来自不同的机构，但仍属于同一学科时，其合作情况可能是学科合作。
4. 如果两位学者来自不同的国家和学科，那么这两位学者进行的是一些跨学科的国际合作，这种合作很可能是需要多个国家共同探讨的特定课题。

具体的数据说明请参见 `Data Description.md`。

此外，同学们也可自选公开数据集进行相应的分析，这里给出一些示例：

- cora, citeseer, pubmed
- Stanford Large Network Dataset Collection

Required Files

- 指标或算法的定义，原理书面报告
- 指标或算法的代码，统一要求输入为图结构，输出为图结构+结果值。要求使用 dgl graph 作为图的输入输出格式
- 对数据的分析结果报告

Visualization Support

本项目会提供一个可视化工具（正在开发，基于d3）帮助可视化和分析图。你只需要提供图的数据即可。工具使用 dgl 作为图数据的工具库。

Bonus

可视化是一种通过图形和动态效果传递信息的技术。随着理论研究的不断深入，所用到的知识就更加抽象和复杂，也就更加需要这种技术来简单的展示知识。随着chatgpt的出现，简单的可视化门槛进一步降低，但是复杂

可视化的领域依然属于人类的想象力和创造力。在你设计完对图或是GNN的评价指标后，你是否对于如何展示你的设计有一定的想法？请给出你的想法，以一个可视化设计方案的方式展现你算法的流程和结果分析。

References

网络分析相关

1. 【PNAS】Consolidation in a crisis: Patterns of international collaboration in early COVID-19 research. paper
2. 【PNAS】Collaboration and Knowledge Networks: A Framework on Analyzing Evolution of University-industry Collaborative Innovation. paper
3. 【Nature Communication】Dynamics of social network emergence explain network evolution. paper
4. 【Financial Science】The Evolving Network of Legal Scholars. paper

网络指标相关

5. 【PNAS】How humans learn and represent networks. paper
6. 【Scientific Report】Ollivier-Ricci Curvature-Based Method to Community Detection in Complex Networks. paper

GNN解释性相关

7. 【ICLR 2022】Understanding over-squashing and bottlenecks on graphs via curvature. paper
8. 【CVPR 2019】Explainability methods for graph convolutional neural networks. paper
9. 【NIPS 2019】GNNExplainer: Generating explanations for graph neural networks. paper
10. 【NIPS 2020】Parameterized explainer for graph neural network. paper
11. 【ICLR 2020】Interpreting graph neural networks for NLP with differentiable edge masking. paper
12. 【Openreview】Hard masking for explaining graph neural networks. paper
13. 【Openreview】Causal screening to interpret graph neural networks. paper
14. 【PMLR】On explainability of graph neural networks via subgraph explorations. paper
15. 【NIPS 2020】PGM-Explainer: Probabilistic graphical model explanations for graph neural networks. paper

Attention

该项目不以特定图数据集上的节点预测准确率为评分依据，以节点预测为任务训练GNN只是为了保证GNN模型的有效性，以便于分析，因此不要纠结于GNN算法的性能。

注意：大作业为小组形式，在组队完成后，不要忘记在Canvas上登记并加入小组。

Acknowledgements

大作业中提供的数据集均来自 *Acemap* 小组。