Author = Song Jian

#### 阅读论文： A FAIR COMPARISON OF GRAPH NEURAL NETWORKS FOR GRAPH CLASSIFICATION

|  |  |  |
| --- | --- | --- |
| **@article{Errica2019,**  **author = {Errica, Federico and Podda, Marco and Bacciu, Davide and Micheli, Alessio},**  **pages = {1--14},**  **title = {{A Fair Comparison of Graph Neural Networks for Graph Classification}},**  **year = {2019}**  **}** | **[1] F. Errica, M. Podda, D. Bacciu, and A. Micheli, “A Fair Comparison of Graph Neural Networks for Graph Classification,” pp. 1–14, 2019.**  **源代码：**[**https://github.com/diningphil/gnn-comparison**](https://github.com/diningphil/gnn-comparison)  文章链接：https://openreview.net/pdf?id=HygDF6NFPB | **比萨大学** |

#### 文档由以下几个部分组成：

* 文章简述
* 作者贡献
* 受测模型介绍
* 实验部分
* 可以学到什么

#### 文章简述

摘要

实验的可重复性和可复制性是机器学习中的关键主题。最近，图形表示学习领域吸引了大量的研究者，从而产生了大量的作品。因此，已经开发了几种图形神经网络模型来有效地处理图形分类。但是，实验程序通常缺乏严格性，并且几乎不可重现。因此，作者提供了一些常见实验的概述，应该与常规技术进行公平比较。作者在一个受控且统一的框架中进行了超过47000个实验，以重新评估跨越9个通用基准的5种流行模型。此外，通过将GNN与结构不可知的基线进行比较，作者提供了令人信服的证据，即在某些数据集上，图结构信息并未被利用。

简介

该领域常见的可再现性问题：超参数选择，模型选择，对于模型评估中用数据拆分的正确使用。实际上，对模型的评估经历了两个不同的阶段，即验证集上的模型选择和测试集上的模型评估。显然，未能将这些阶段很好地分开可能会导致对模型真实性能的过度乐观和有偏见的估计，这使得其他研究人员如果不遵循相同的模棱两可的评估程序，就很难呈现

出具有竞争力的结果。

#### 作者贡献

1. 使用标准化，且可重现的的实验环境，对几种模型进行了公平比较。更详细地讲，在严格的模型选择和评估框架内进行了大量实验，其中所有模型都使用相同的功能和相同的数据分割进行了比较。
2. 研究当前的GNN模型是否以及在何种程度上可以有效利用图结构。
3. 研究节点度作为社交数据集中特征的影响。 的确，我们表明提供度可以提高表现，并且对达到良好结果所需的GNN层数也有影响。
4. 公开发布代码和数据集拆分以重现结果，以便允许其他研究人员以最少的额外努力进行严格的评估。

#### 受测模型

**作者将模型的评价分为了两个阶段**：

（i）model selection

（ii）model assessment。

前者作为超参数选择阶段，后者为模型评估阶段。

**根据以下标准选择合适的GNN模型进行测试：**

1. 用10折CV获得的表现；
2. 同行评审；
3. 强烈的架构差异；
4. 受欢迎程度

**因此选出了因此选择了DGCNN，DiffPool，ECC，GIN**

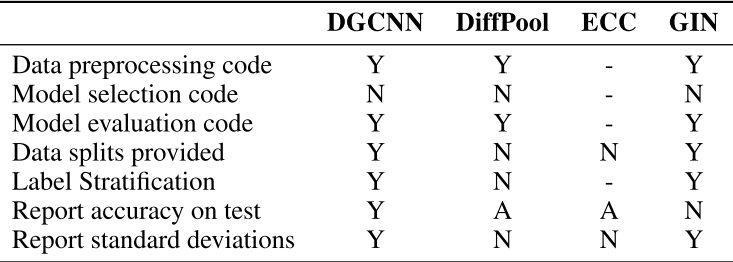


Table 1: Criteria for reproducibility considered in this work and their compliance among considered models.

* **DGCNN**作者使用10折CV评估模型。虽然所有数据集的结构都是固定的，但仅使用一个随机CV折叠调整学习率和epoches，然后在所有其他折上重用。尽管这种做法仍然可以接受，但可能会导致性能欠佳。而且没有提供用于重现模型选择的代码。此外，作者运行了10次，他们报告了10个最终分数的平均值。因此，提供评估的方差被减小了。但是，其他的对比模型却没有采取相同的方式。最后，CV数据分割被正确地分层并且公开，因此至少可以再现评估实验。
* **DiffPool**从本文和提供的代码中都未标明，是否在测试集而不是验证集上获得了报告的结果。尽管作者声明使用了10折CV，但未报告DiffPool及其对比模型的标准偏差。此外，作者确认对验证集应用了早期停止功能以防止过拟合。不幸的是，既没有模型选择代码也没有验证拆分。此外，根据代码，数据被随机分割（不分层）并且没有设置随机种子，因此每次执行代码时，分割都是不同的。
* **ECC**本文报告说ECC是用10折CV评估的，但结果不包括标准差。与DGCNN相似，超参数是预先固定的，因此尚不清楚是否以及如何执行模型选择。重要的是，在代码存储库中没有对数据预处理，数据分层，数据拆分和模型选择的引用。
* **GIN**作者正确列出了所有已调整的超参数。但是，正如论文和公众讨论中明确指出的那样，他们报告了10折CV的验证准确性。换句话说，报告的结果涉及模型选择而不是模型评估。没有提供用于模型选择的代码。
* **GraphSAGE**原始论文没有在图分类数据集上测试此模型，但是GraphSAGE在其他论文中经常被用作基线。因此，图分类的GraphSAGE结果应附有代码以重现实验。尽管如此，展现GraphSAGE结果的两个作品（DiffPool和GIN）还是没有这样做。

**评估和可重现性的标准是：**

1. 提供了数据预处理，模型选择和评估的代码；
2. 提供了数据拆分；
3. 通过分层抽样技术对数据进行拆分，以保留所有分区中的类比例；
4. 使用标准偏差正确报告了10倍CV的结果，它们涉及模型评估（测试集）而不是模型选择（验证集）。

#### 实验部分

* 使用9个数据集（5个社会类，4个化学类）
* 特征：在GNN文献中通常使用结构特征来扩充对节点的描述，比如DiffPool将度和聚类系数添加到每个节点特征向量，而GIN添加一个one-hot表示节点度。对于化学领域，节点使用原子类型的one-hot encoding。对于节点没有特征的社交图(如Reddit)，采用无信息特征(将特征设置为相同或者度的)。

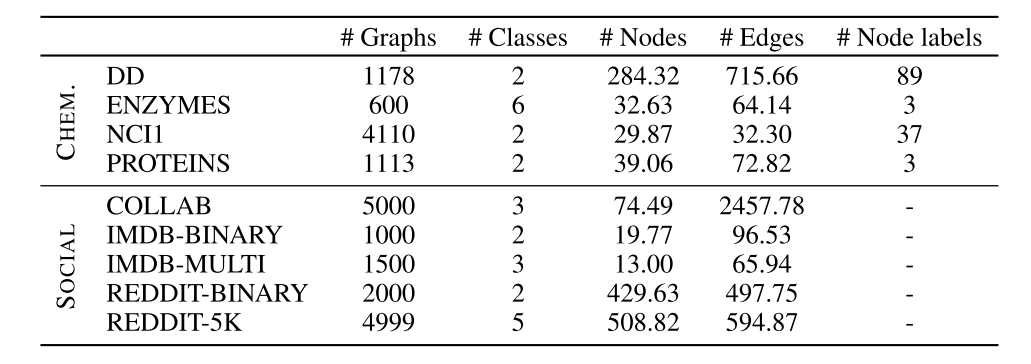


Table 2: Dataset Statistics.

* 基线：采用两种不同的基准，一种用于化学基准，另一种用于社会数据集。这两个基线都没有利用图拓扑。使用这些基线作为参考对于将来的工作至关重要，它们可以提供有关特定数据集上的GNN有效性的反馈。事实上，如果GNN的性能接近结构不可知的基线，则可以得出两个可能的结论：该任务不需要有效地解决拓扑信息，或者GNN没有充分利用图结构。前者可以通过特定领域的人员专业知识进行验证，而第二个则更难评估，因为多种因素会发挥作用，例如训练数据的数量，体系结构所施加的结构归纳偏差以及用于分析的超参数。但是，相对于这些基线的显着改进是图形拓扑已被利用的有力指标。因此，与结构无关的基准对于了解是否以及如何改进模型至关重要。
* 实验方法：作者提供了一个基于交叉验证的评测流程，例如采用10折CV用于模型评估，按照90%/10%的比例划分训练集/验证集用于模型选择。

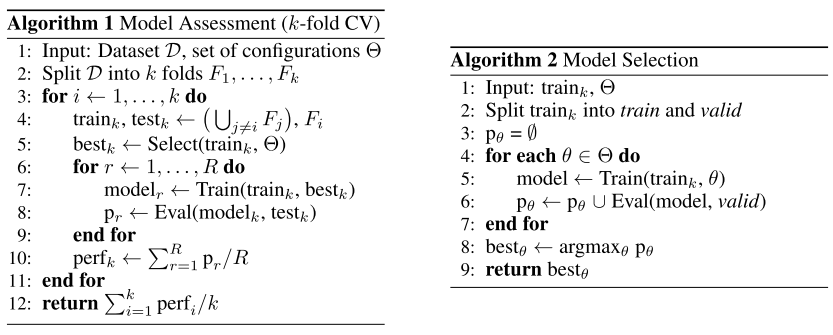
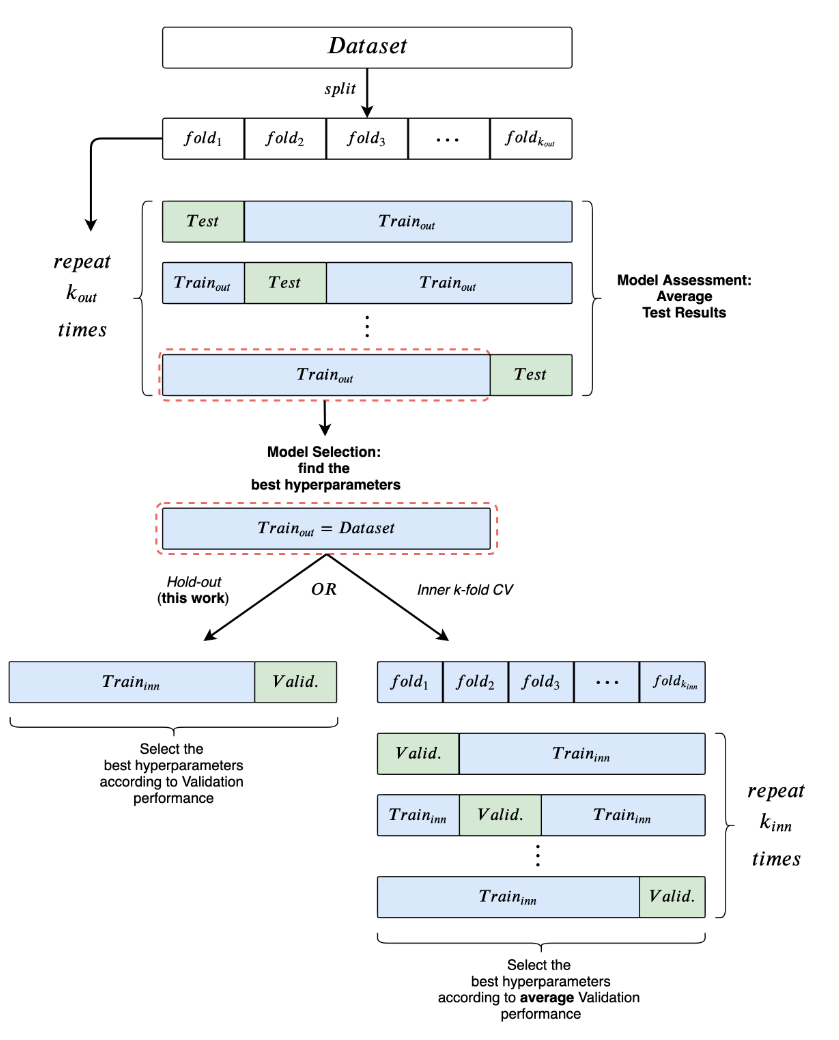


Table 2: Pseudo-code for model assessmen



#### 结果与讨论

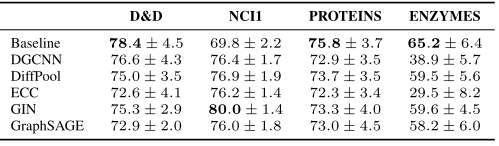


Table 3: Results on chemical datasets with mean accuracy and standard deviation are reported. Best performances are highlighted in bold.

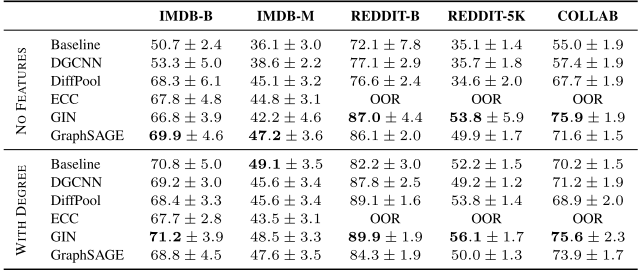


Table 4: Results on social datasets with mean accuracy and standard deviation are reported. Best performances are highlighted in bold.

表3和表4显示了我们的实验结果。总体而言，GIN在社交数据集上似乎很有效。重要的是，我们发现在D＆D，蛋白质和酶方面，没有一个GNN能够在基线水平上得到改善。相反，在NCI1上，明显胜过基线，该结果表明，我们分析的GNN可以实际利用该数据集中图形的拓扑信息。

###### 基准的重要性

我们的结果还表明，与结构无关的基线是了解GNN的有效性并提取有用见解的重要工具。例如，由于没有一个GNN超过D＆D，蛋白质和ENZYMES的基线，因此我们认为，我们分析的最新GNN模型还不能完全利用此类数据集的结构。实际上，在化学中，已知结构特征与分子特性相关（van Rossum，1963）。由于所有这些原因，我们建议将这些数据集的较小性能提升放到正确的角度来看，至少要等到基线明显好于预期为止。当前，这些数据集上的平均波动很小，可能是由其他因素（例如随机初始化）引起的，而不是结构的成功利用。总之，我们热烈推荐GNN的从业者在未来的工作中包括基线比较，以更好地描述他们的贡献程度。

###### 节点度的影响

根据我们的结果，使用节点度作为输入特征几乎总是有益于提高社交数据集的性能，有时甚至是大量提高。例如，度信息足以使我们的基线提高约15％的性能，因此在许多数据集中具有竞争力。特别是，基线在IMDB-BINARY上达到了最佳性能。相比之下，增加节点度对大多数GNN而言就不那么重要了，因为它们可以从结构中自动推断出此类信息。

DGCNN是一个值得注意的例外，它明确需要节点度才能在所有数据集上表现良好。此外，我们观察到，所有模型的排名在添加度之后都发生了急剧变化。

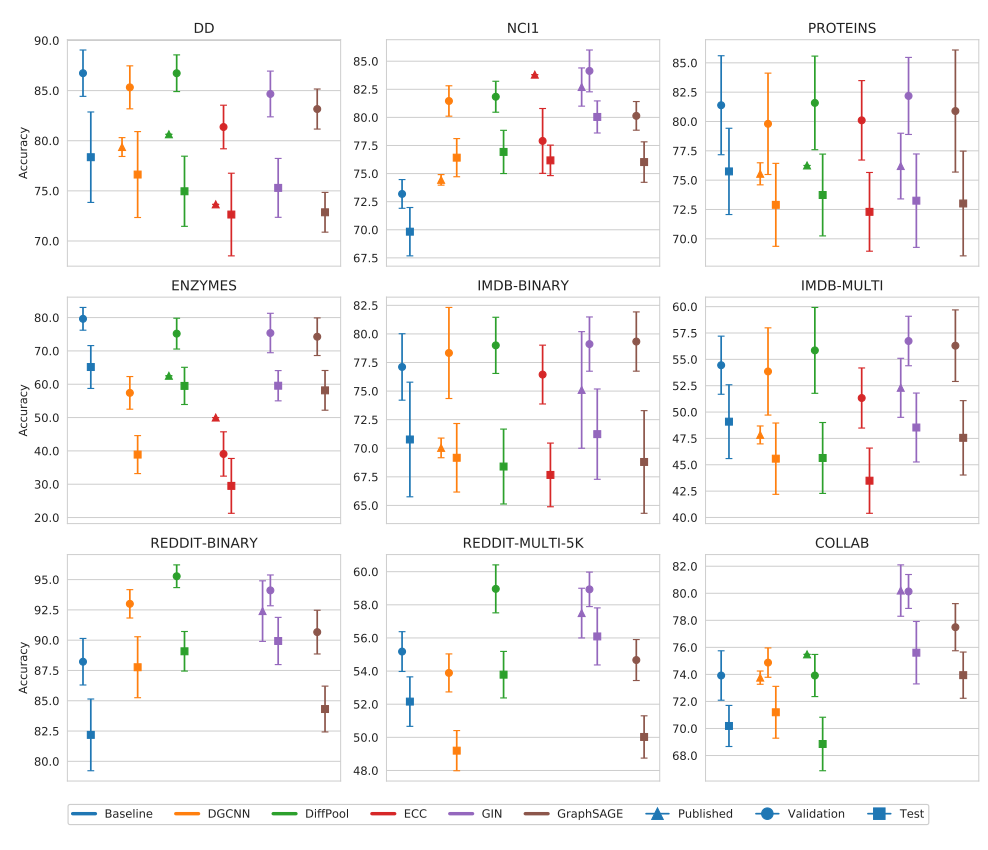


Figure 1: Chemical and social (with degree) benchmark results are shown together with published results (when available). For each of them, we report validation and test accuracies of the evaluated models, together with published results if available.