Beijing Forest Studio

北京理工大学信息系统及安全对抗实验中心

图神经网络的可解释性

慕星星　硕士研究生

2021年07月11日

内容提要

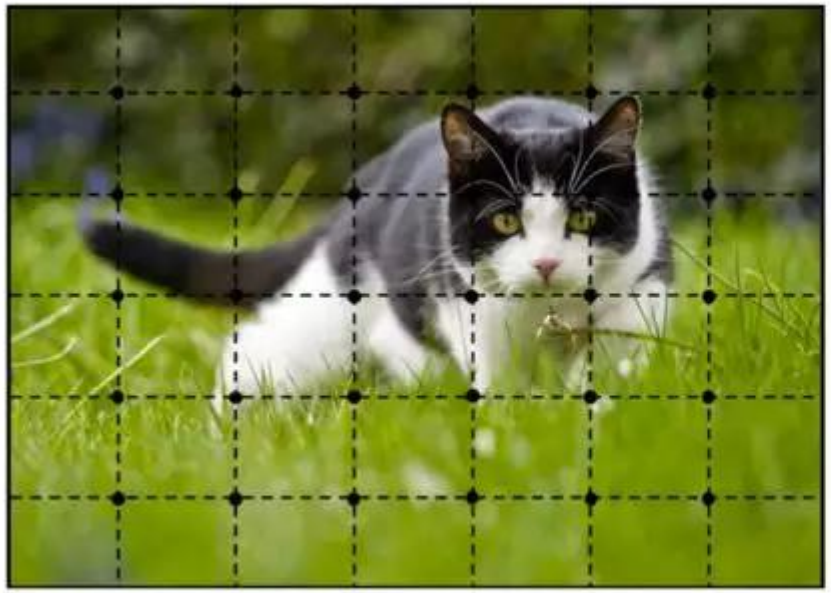
- 背景简介
- 基本概念
- 算法原理
- 应用总结
- 参考文献

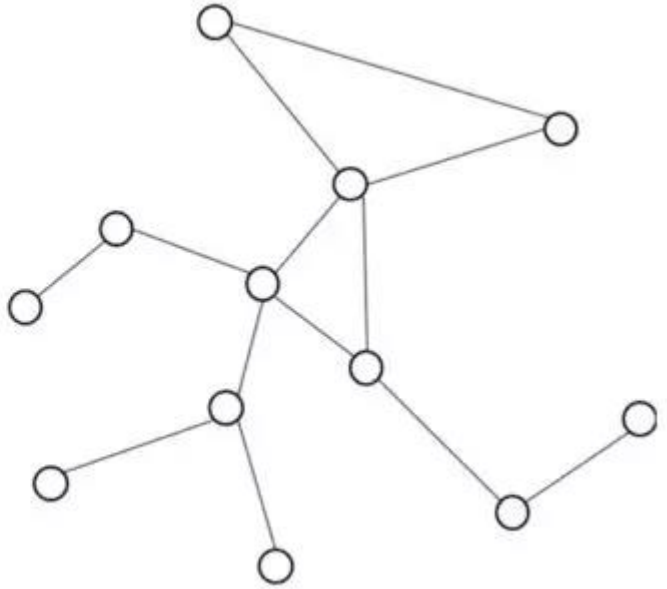
预期收获

- 预期收获
  - 1. 了解图神经网络可解释性的基本概念
  - 2. 了解图神经网络可解释性方法的分类
  - 3. 了解常用的图神经网络可解释性方法的基本原理

背景简介

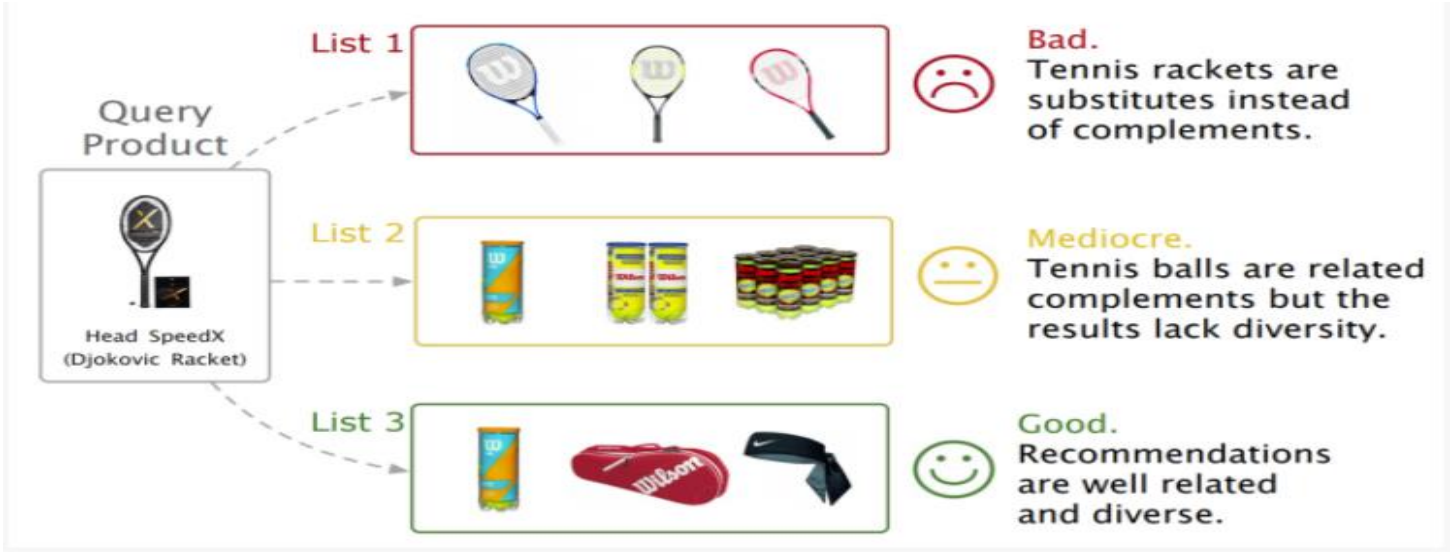
- 图神经网络(Graph Neural Networks, GNN)
  - 图是一种数据结构，它对一组对象（节点）及其关系（边）进行建模。近年来，由于图结构的强大表现力，用机器学习方法分析图的研究越来越受到重视。
  - 图神经网络（GNN）是一类基于深度学习的处理图域信息的方法。GNN充分利用了图结构信息，可以很方便的应用于节点级、边级和图级预测任务。

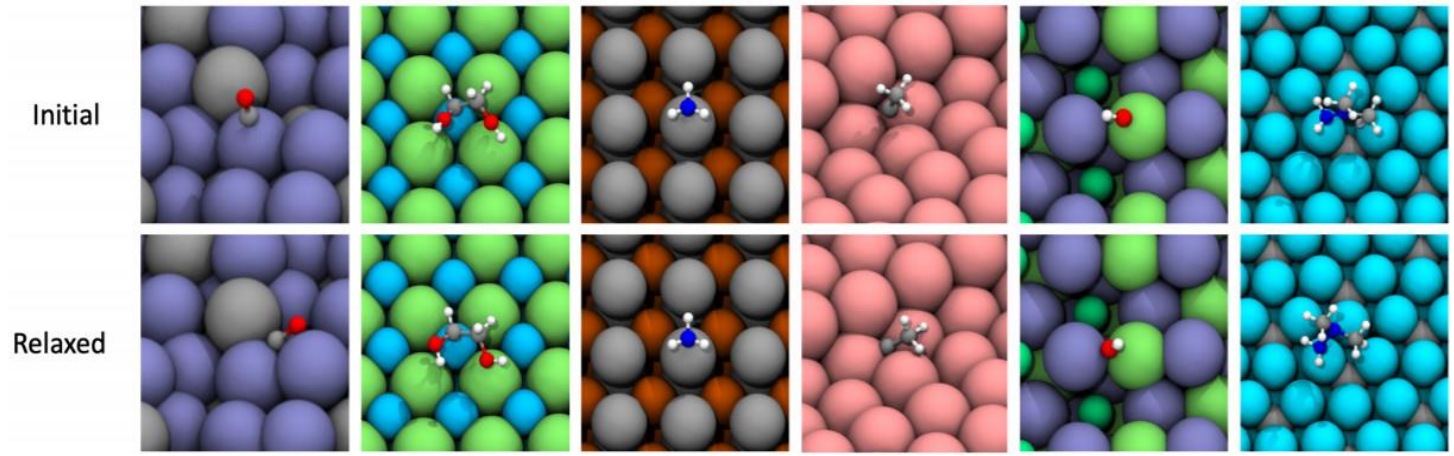




背景简介

- 图神经网络的应用
  - 推荐系统
  - 知识图谱
  - 计算机视觉
  - 粒子物理学
  - 药物研发





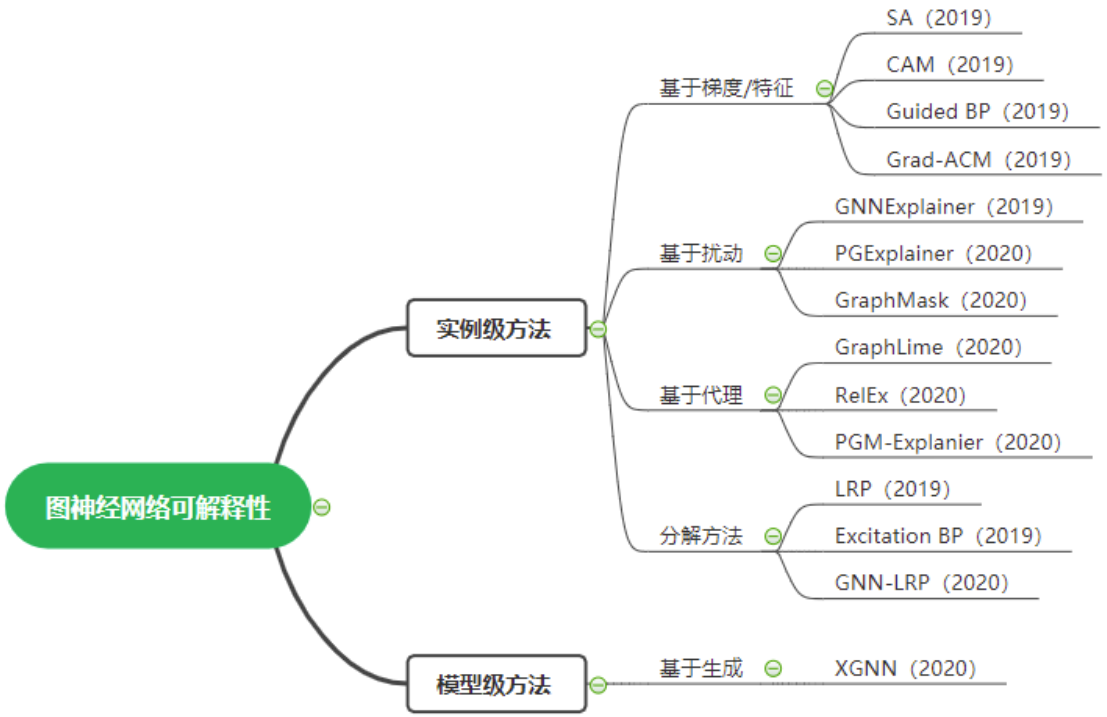
基本概念

- 可解释性名词说明
  - Interpretability：模型本身能够对其预测提供人类可理解的解释；
  - Explainability：模型是黑盒子，本身不能提供解释，其预测可以被一些事后解释技术所解释。

- 图神经网络可解释性的意义
  - 提高人对GNN模型的信任程度；
  - 在很多注重公平性，隐私性和安全性的应用领域中，提高模型的透明度；
  - 提高开发者对网络特征的理解、认知，并且减少模型出现系统性错误的风险。

基本概念

- 图神经网络可解释性方法
  - 实例级方法：实例级方法旨在探究影响模型预测的重要特征，实现对深度模型的解释
  - 模型级方法：模型级方法则提供了高层次的见解和对深度图模型工作原理的一般理解



基本概念

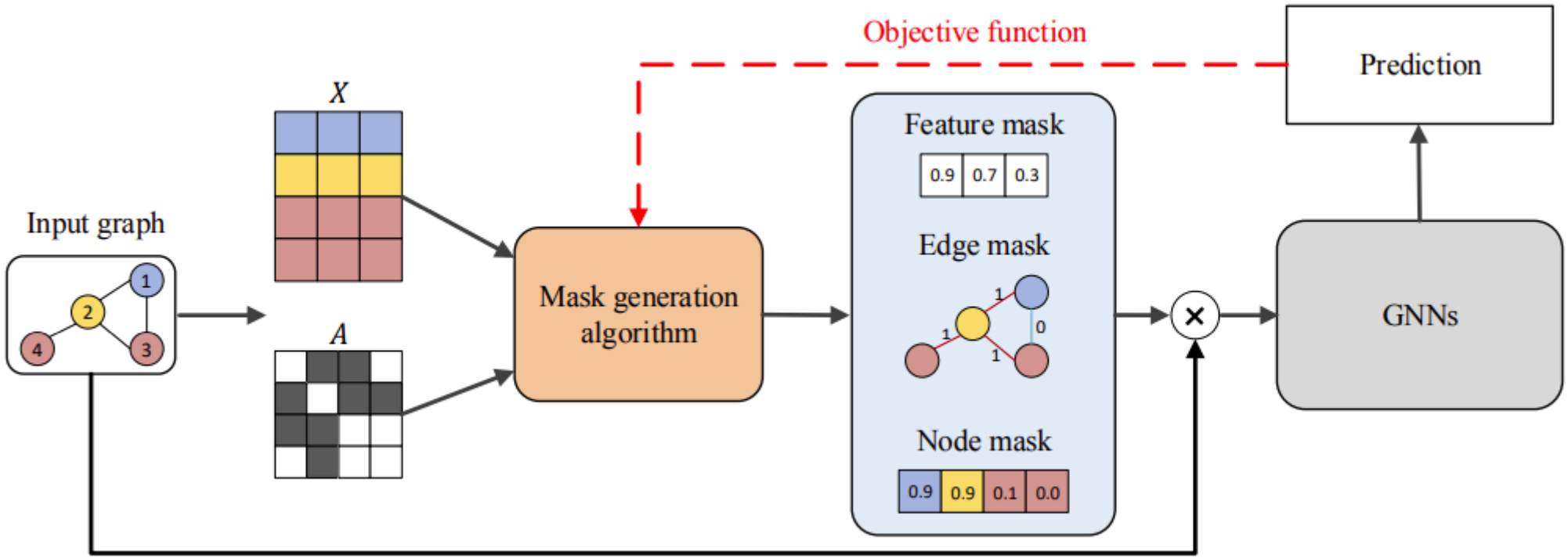
- 基于梯度/特征的方法
- 基本思想：是将梯度或隐藏特征的映射值作为输入特征重要性的近似值。

- 具体实现：基于梯度的方法通过反向传播计算目标预测相对于输入特征的梯度；基于特征的方法通过将隐藏特征映射到输入空间来度量输入特征重要性得分。

- 主要区别：梯度方向传播的过程以及将不同的隐藏特征结合起来的方法

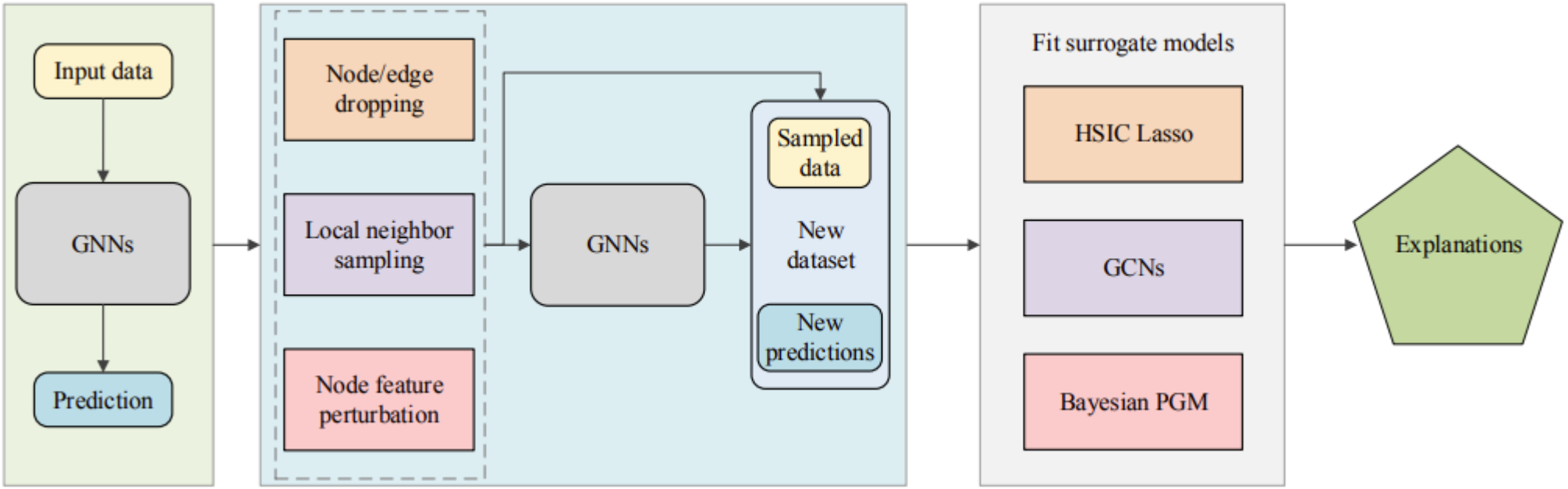
基本概念

- 基于扰动的方法
- 基本思想：研究不同输入扰动下的输出变化。
- 主要区别：掩码生成算法，掩码类型和目标函数



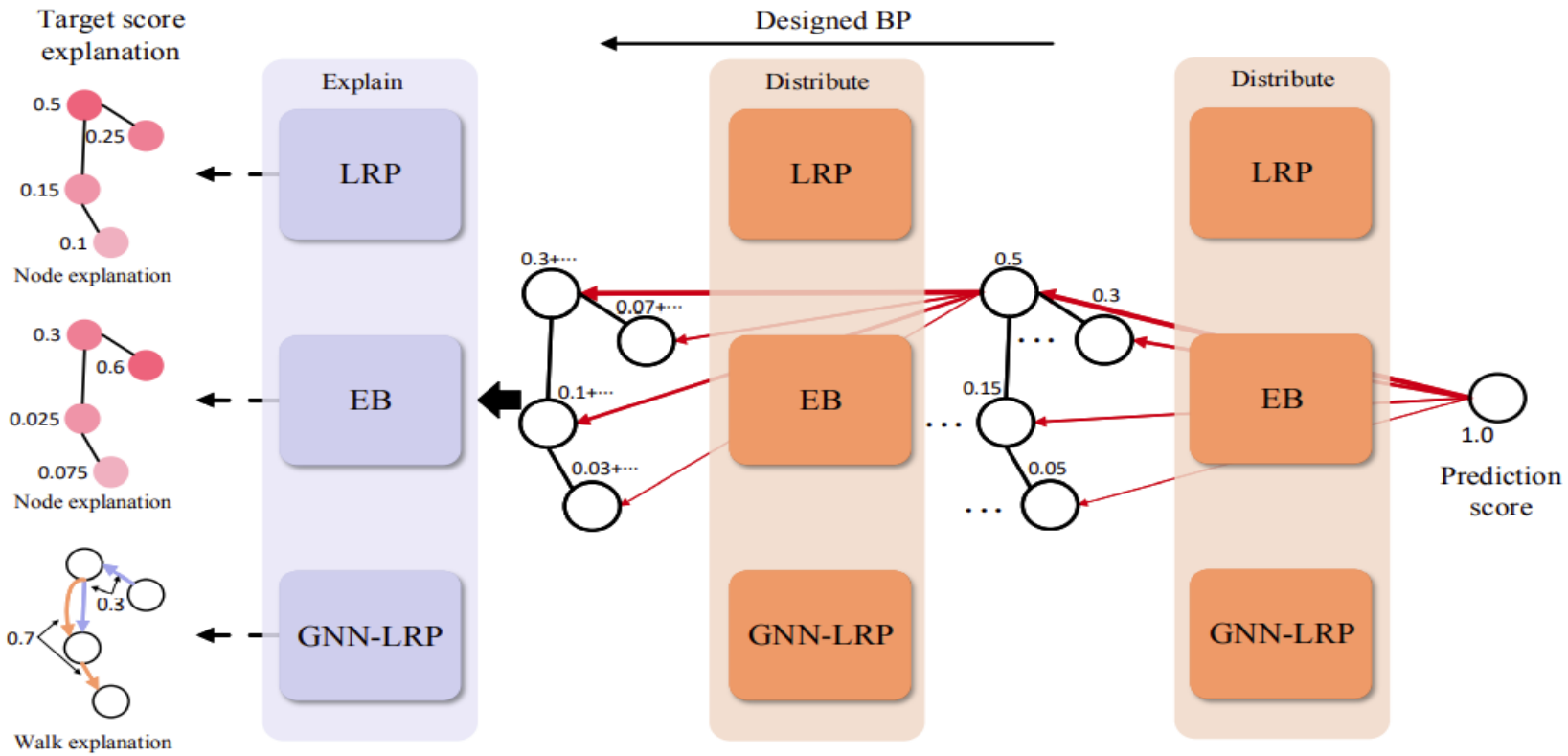
基本概念

- 基于代理的方法
- 基本思想：化繁为简，采用一个简单且可解释的代理模型来近似复杂的深层模型，实现输入实例的邻近区域预测。
- 主要区别：局部数据集获取方法和代理模型



基本概念

- 分解方法（Decomposition Methods）
- 基本思想：将原始模型预测分解为若干项来衡量输入特征的重要性。
- 主要区别：分数分解规则和解释的目标



算法原理

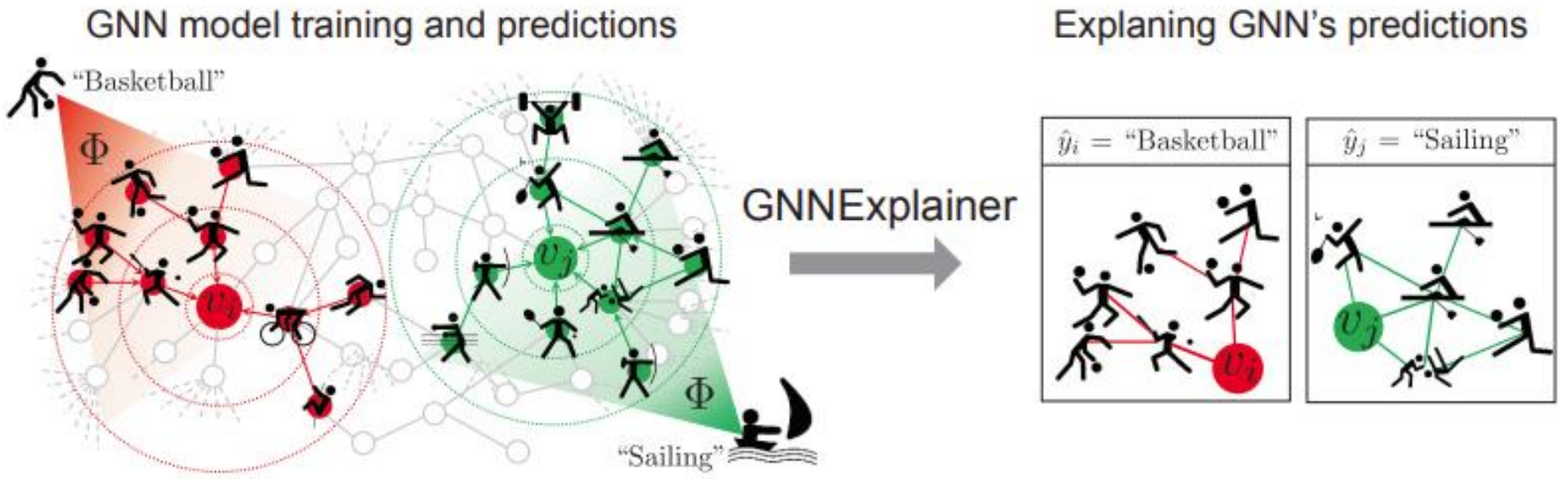
GNNExplainer

T	通用的、模型无关的图神经网络的解释器
I	训练好的GNN模型和其预测结果
P	学习边和节点特征的软掩码，进行掩码优化
O	输入图的结构子图与节点特征子集

P	GNN模型的预测无法解释
C	模型无关的方法
D	需要考虑图的结构特征
L	NIPS 2019

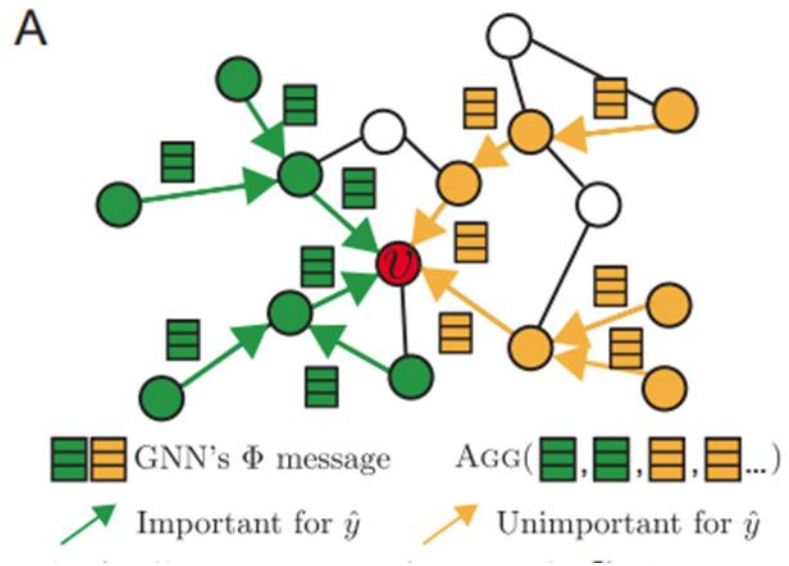
GNNExplainer

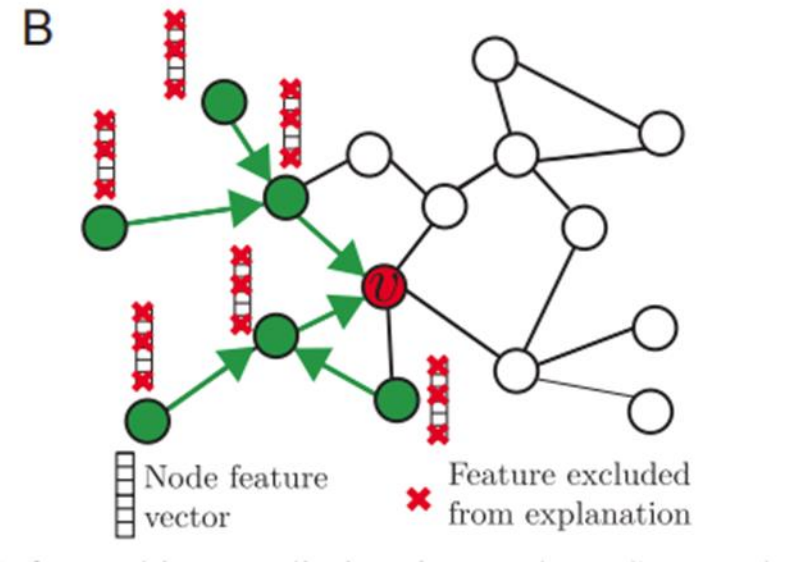
- GNNExplainer可以通过输出输入图的结构子图与节点特征子集的方式，表示最能影响预测结果的子图结构与特征向量子集。



GNNExplainer

- 基本原理：通过减少图中的冗余信息，获取图中影响模型决策的关键结构或特征来解释模型的预测结果。





GNNExplainer

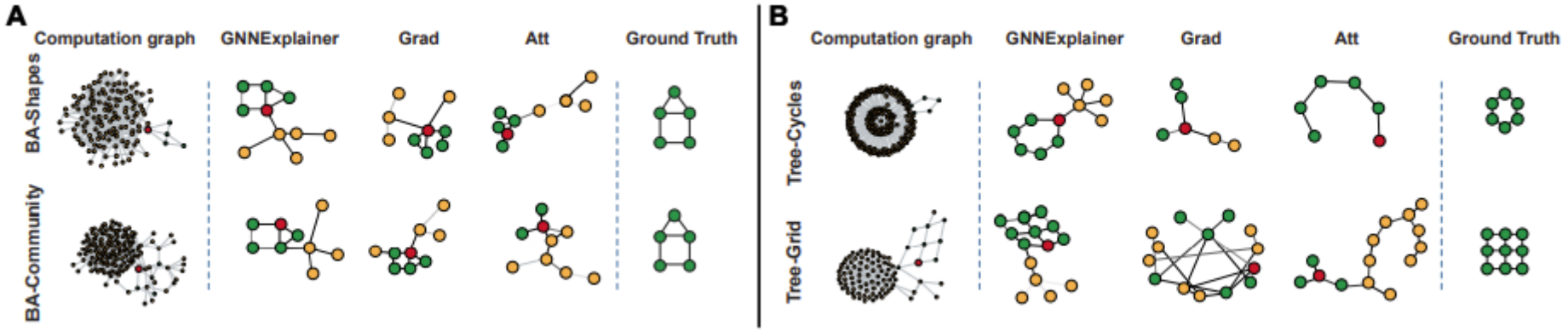
- 基本过程：
  - 学习边和节点特征的软掩码
  - 通过元素点乘将掩码与原始图结合得到最小图
  - 最大化原始图的预测和最小图的预测之间的互信息来优化掩码。 Y代表模型预测结果，G_S和X_S分别为最小图的结构和特征，MI为互信息

$$\max_{G_S} MI\left(Y,\left(G_S,X_S\right)\right)=H\left(Y\right)-H\left(Y\mid G=G_S,X=X_S\right)$$

GNNExplainer

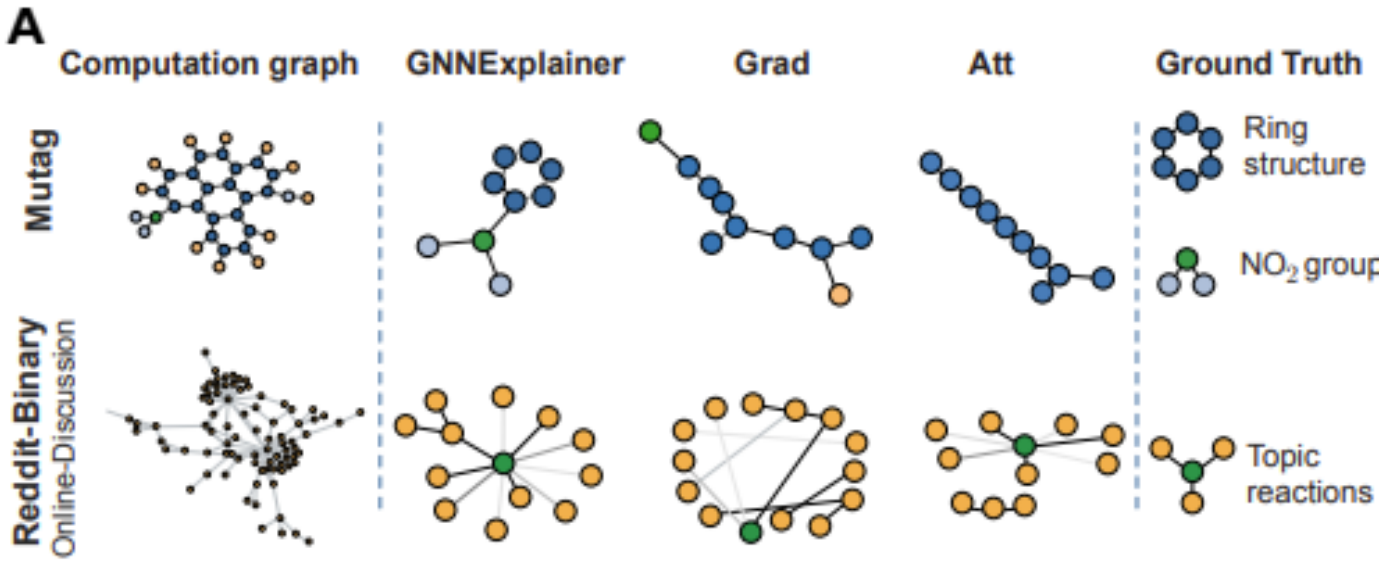
- 合成数据集实验结果

	BA-Shapes	BA-Community	Tree-Cycles	Tree-Grid
Att	0.815	0.739	0.824	0.612
Grad	0.882	0.750	0.905	0.667
GNNexplanier	**0.925**	**0.836**	**0.948**	**0.875**



GNNExplainer

- 真实数据集实验结果
  - Mutag数据集（分子数据集，根据对某类细菌的诱变效应进行分类）
  - Reddit-Binary数据集（Reddit在线讨论主题数据集）



GNNExplainer

- 优点：
  - 适用于任何在图上的机器学习任务（节点分类，图分类和边预测）
  - 模型无关，适用于任何GNN模型
  - 计算复杂度低

- 缺点：
  - 得到的掩码是软掩码，可能会给输入图引入新的噪声
  - 掩码是针对每个输入图单独优化的，解释缺乏全局视角

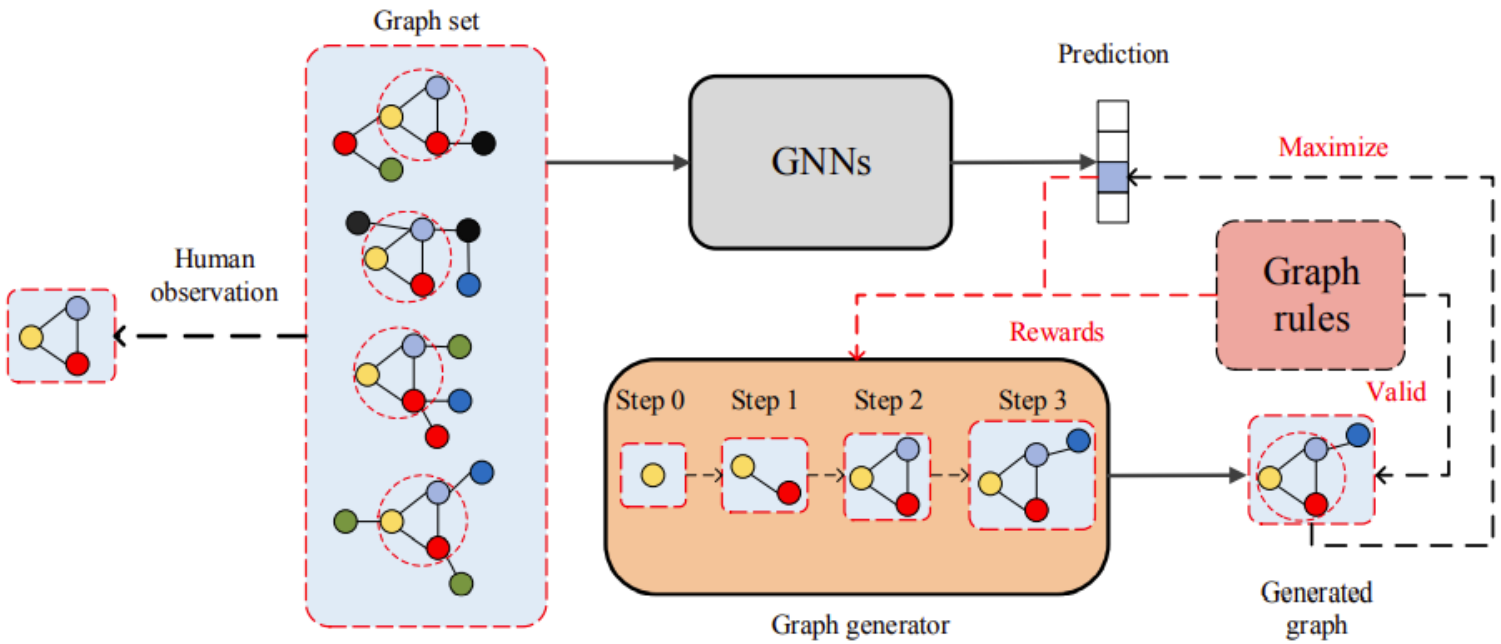
算法原理

T	在模型层面对GNN模型进行解释
I	训练好的GNN模型
P	基于强化学习的方法训练一个图生成器
O	图生成器，能最大化GNN模型的预测概率的生成图

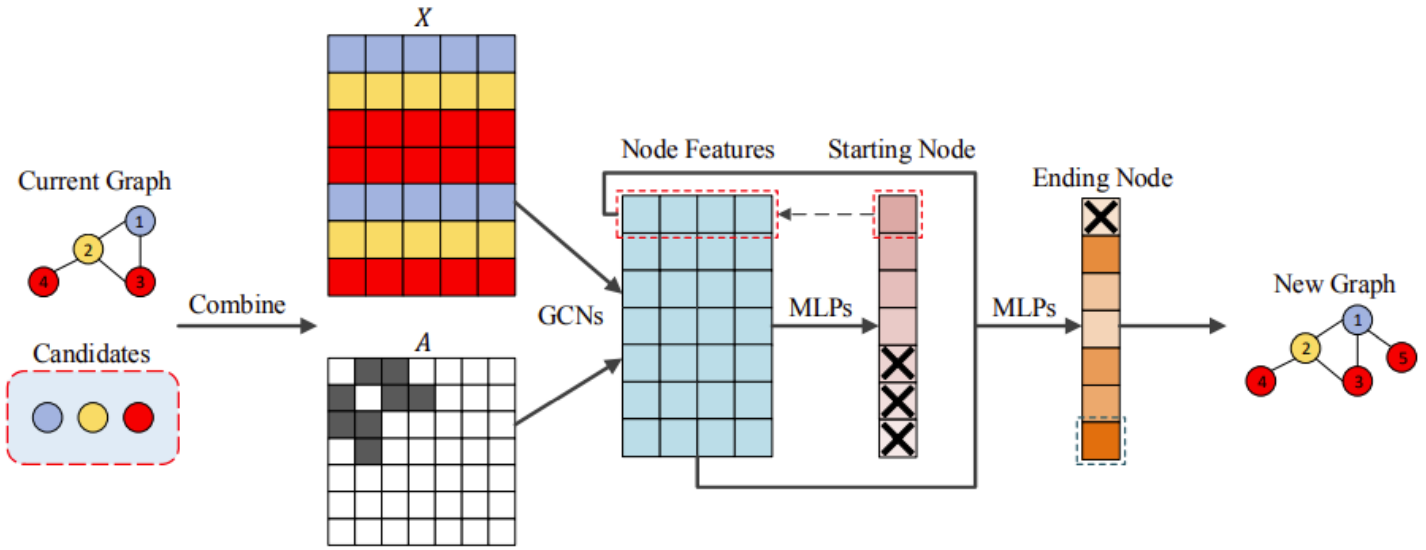
P	GNN被视为黑盒模型，缺乏人类可理解的解释
C	在模型层面
D	图结构离散，无法直接优化
L	KDD 2020

XGNN

- XGNN通过训练一个图生成器来解释GNNs，使用强化学习训练一个分步的图生成器，对于每一步，图生成器预测如何在当前图形中添加边。

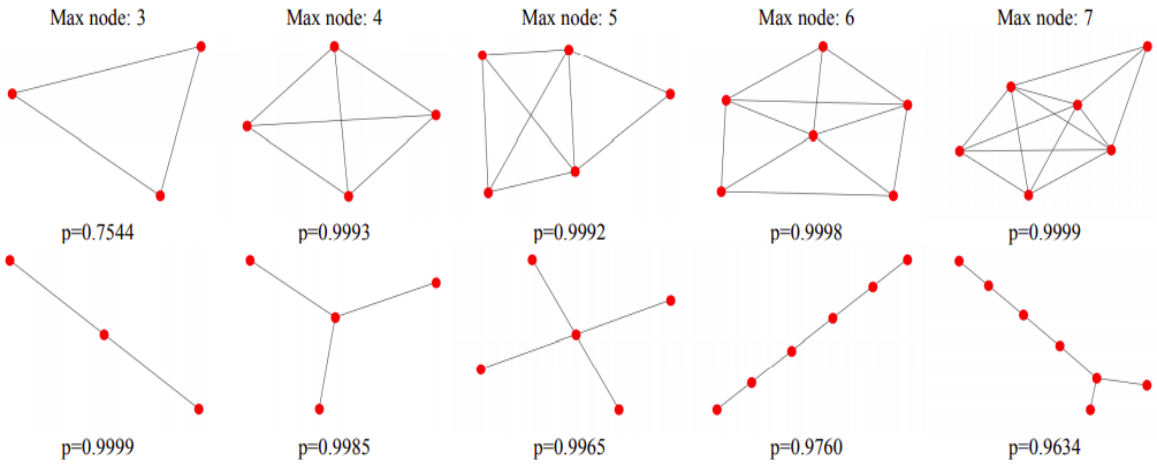


- 图生成策略
  - 将当前图和候选集组合，得到特征矩阵和邻接矩阵
  - 使用GCN（图卷积网络）学习节点表示
  - 使用MLP预测起始节点（限制：必须在当前图上选择）
  - 把起始节点加入，用另一个MLP预测终止节点（限制：不能与起始节点相同）
  - 连接起始节点和终止节点

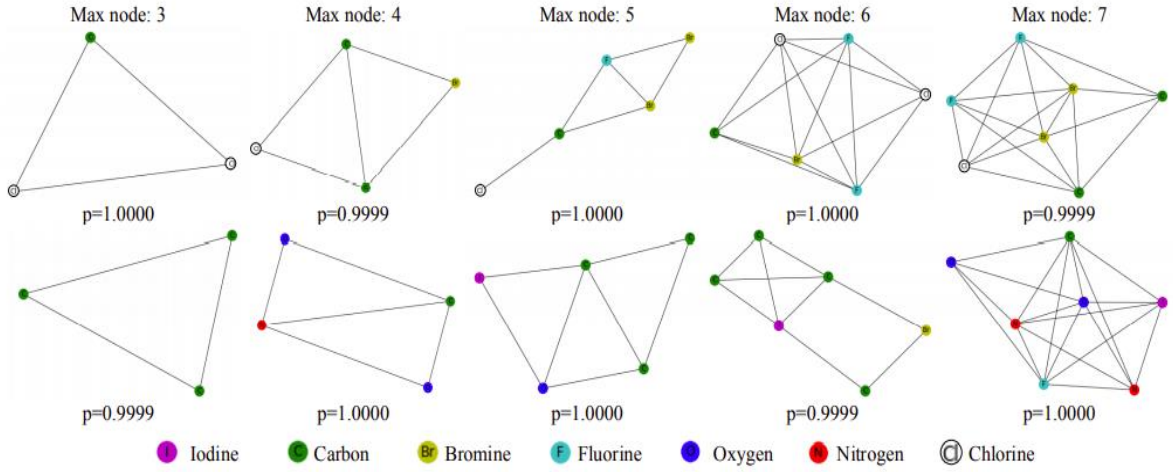


- 实验结果

Dataset	Classes	# of Edges	# of Nodes	Accuracy
Is_Acyclic	2	30.04	28.46	**0.978**
MUTAG	2	19.79	17.93	**0.963**



Experimental results for Is_Acyclic



Experimental results for MUTAG

XGNN

- 优点：
  - 模型级解释的通用框架，可以应用任何图生成算法
  - 提供了对GNN模型的全局解释

- 缺点：
  - 仅验证了在图分类任务中的有效性，不确定是否适用于其他图神经网络任务（节点分类任务和边预测任务等）

总结

总结

- 应用
  - 物理/化学（粒子物理学，低成本催化剂分子）
  - 药物研发（分子特性预测，蛋白质工程和药物再利用）

- 未来发展
  - 模型级的解释
  - 提升GNN模型本身的可解释性（引入注意力机制、解耦表示学习等技术）
  - 确定的评估方法和同一的度量标准

参考文献

- [1] Ying R , Bourgeois D , You J , et al. GNNExplainer: Generating Explanations for Graph Neural Networks[J]. Advances in neural information processing systems, 2019, 32:9240-9251.
- [2] Yuan H , J Tang, Hu X , et al. XGNN: Towards Model-Level Explanations of Graph Neural Networks[C]// KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. ACM, 2020.
- [3] Yuan H , Yu H , Gui S , et al. Explainability in Graph Neural Networks: A Taxonomic Survey[J]. 2020.
- [4] Wu Z , Pan S , Chen F , et al. A Comprehensive Survey on Graph Neural Networks[J]. IEEE Transactions on Neural Networks and Learning Systems, 2019.

道德经

大成若缺，其用不弊。
大盈若冲，其用不穷。
大直若屈。大巧若拙。
大辩若讷。静胜躁，寒
胜热。清静为天下正。

谢谢！