

F R A U D U S E R D E T E C T I O N

聚焦数创·连接未来

第七届信也科技杯算法大赛

7th Finvolution Data Science Competition



GNN in 欺诈用户风险识别

2022 第七届信也科技杯算法大赛

SYSU-GEAR

2022/09/19

目录

- 团队简介
- 数据分析
- 解决方案
- 问题定义
- 思路分享
- 未来工作

团队简介

SYSU-GEAR(Graph IEARning)

中山大学图学习研究团队

做一颗扭动科技进步的齿轮

指导老师：陈亮 副教授

队伍成员

李金膛、于宙鑫、孙王斌、金信洲、王其超

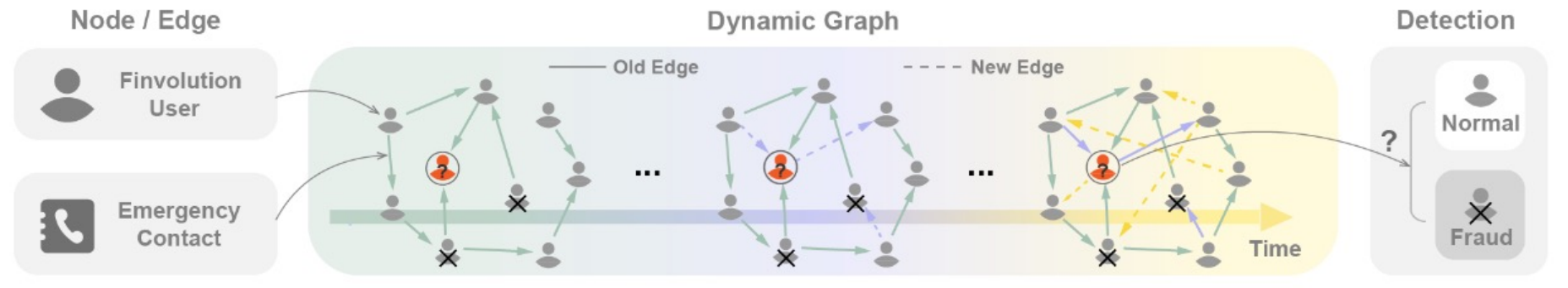
问题定义

动态社交网络视角下的智能风控

数据概况

| 任务类型 | 节点 | 节点类别 | 有向边 | 边类别 | 时间戳 |
|----------------|---------|------|---------|-----|-----|
| 欺诈检测 (节点分类) | 4059035 | 4 | 4962032 | 11 | 578 |

动态有向稀疏社交关系图



数据分析

数据决定性能上限

数据分析

- 正常用户远远多于欺诈用户
 - $22.07 \div 0.26 \approx 85$
 - 样本不均衡
- 背景节点多于前景节点
 - $(54.73 + 22.94) \div (22.07 + 0.26) \approx 3.5$
 - 前景只是冰山一角
存在大量无标签数据

| 已知节点类别 | 占比 |
|--------|--------|
| 0 | 22.07% |
| 1 | 0.26% |
| 2 | 54.73% |
| 3 | 22.94% |

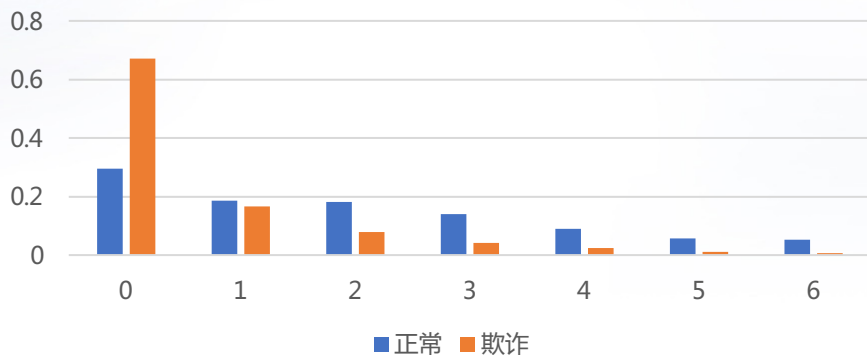
数据分析

- 节点特征
 - 节点特征维度较少
 - 匿名特征需要挖掘的工程量大
 - 缺失值占比: 56.37%
 - 欺诈节点缺失值占比(67%)
 - 正常节点缺失值占比(39%)
- 丰富的拓扑结构
 - 平均节点度数为1.22
 - 边上有丰富的属性信息

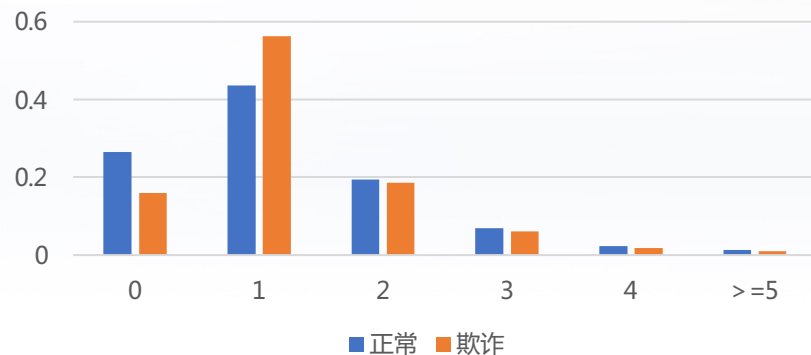
| 数据 | 规模 |
|------|---------|
| 节点 | 4059035 |
| 节点特征 | 17维匿名特征 |
| 边 | 4962032 |
| 边属性 | 类别，时间戳 |

数据分析

节点出度分布

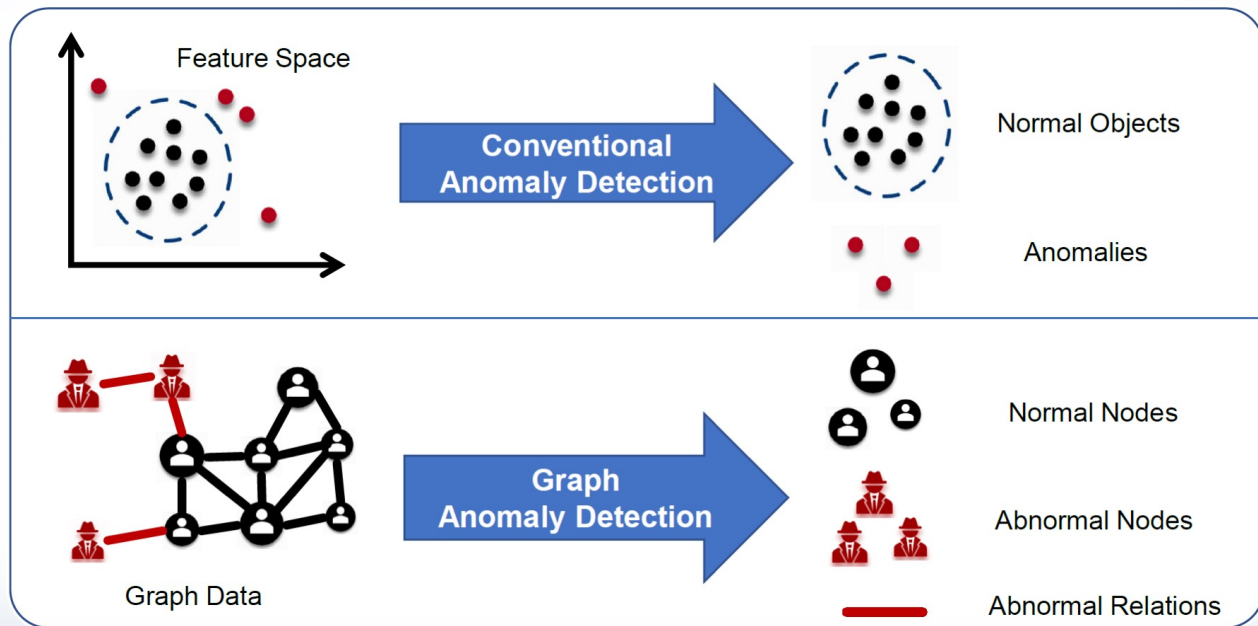


节点入度分布



- 两类节点出入度分布差异较大
- 出度表示信息完善度，入度表示被信任关系
- 节点入度不可自控，出度可以自控

方法选型



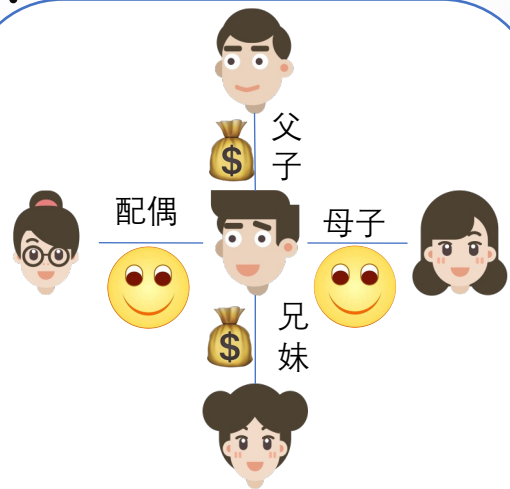
- 利用图算法进行欺诈检测：
半监督，契合拓扑结构，归纳学习，无需繁琐特征工程

思路分享

寻找勇攀高峰的路

思路分享

- 风控场景下传统GNN模型哪里做的不够好？
 - GNN基于同质性假设，对连边一视同仁
 - 风控场景的边存在多种类型
- 引文网络同质性强
- 引用与被引相似性强
- 传递的是信息
- 紧急联系人社交网络传递的是感情和金钱
- 因此需要考虑边的属性



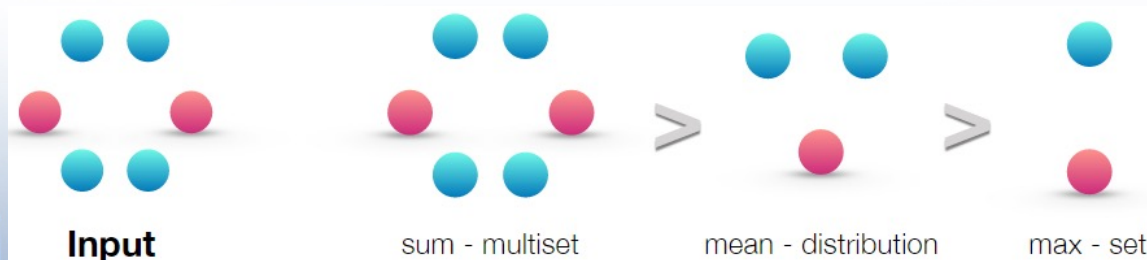
同样的消息，不同的含义

思路分享

- GCN的正则化形式

$$\tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}}$$

- 加大了低度节点的影响：mean(1,0)=0.5 sum(1,0)=1
- Mean的聚合方式不符合场景逻辑：
每个朋友给自己一元钱，两个朋友和两百个朋友肯定不一样



解决方案

不断逼近上限

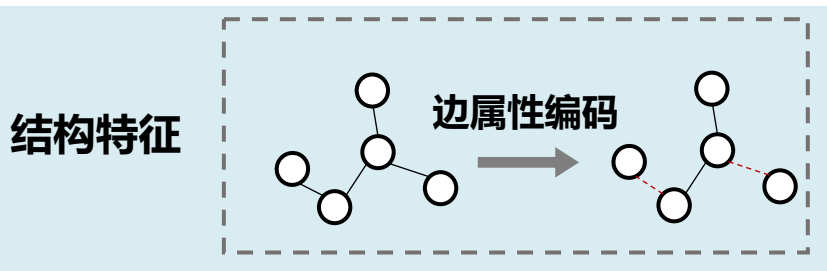
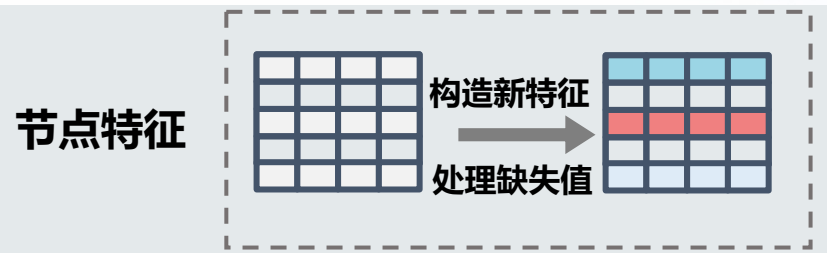
解决方案

• 节点特征工程

- 缺失值处理：缺失Flag，替换
- 增加节点度数
- 邻域时间戳信息
- 前景/背景节点信息
- 与邻域的相似度信息

• 边特征工程

- 加入反向边为图增加稠密度
- 构造边的方向性信息
- 边属性编码(embedding)

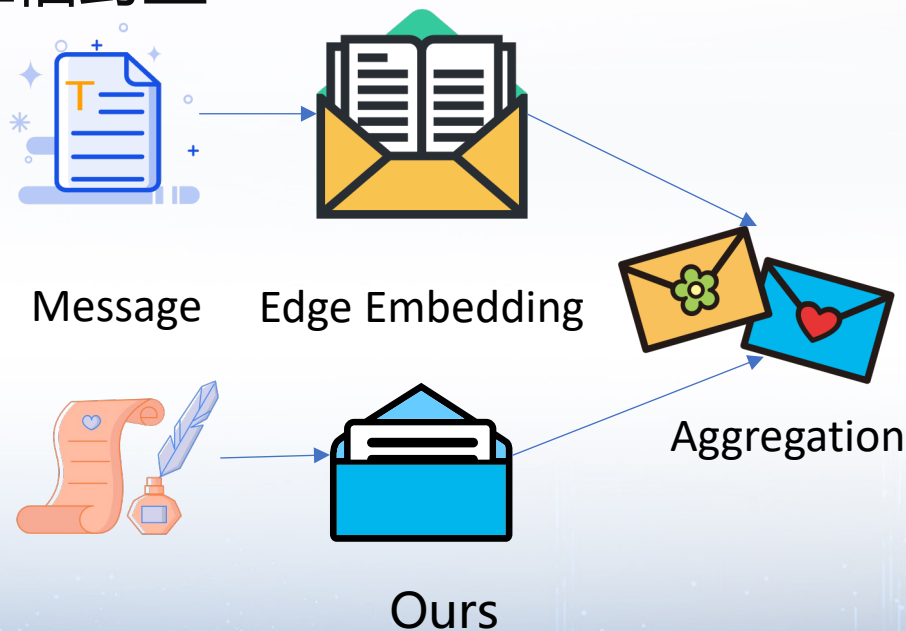


解决方案

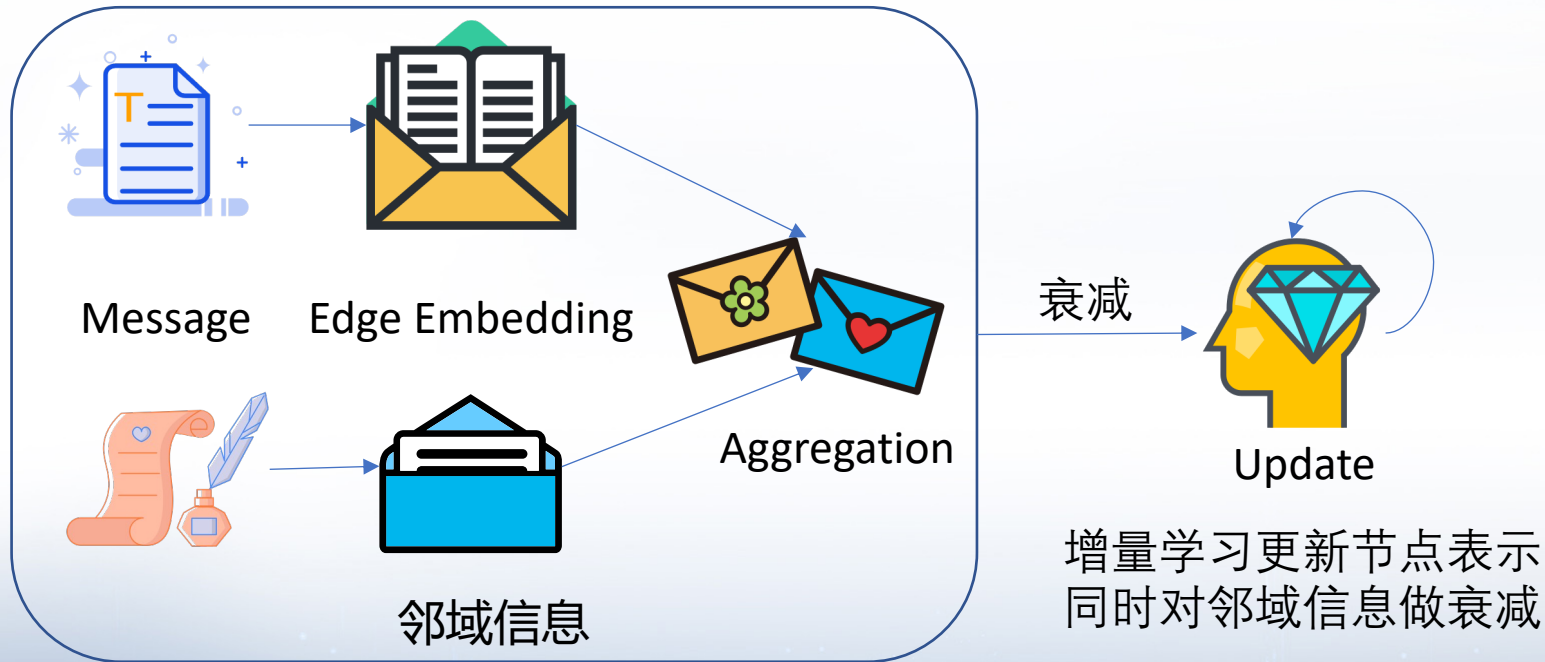
改进卷积算子——把消息装在信封里



传统GCN

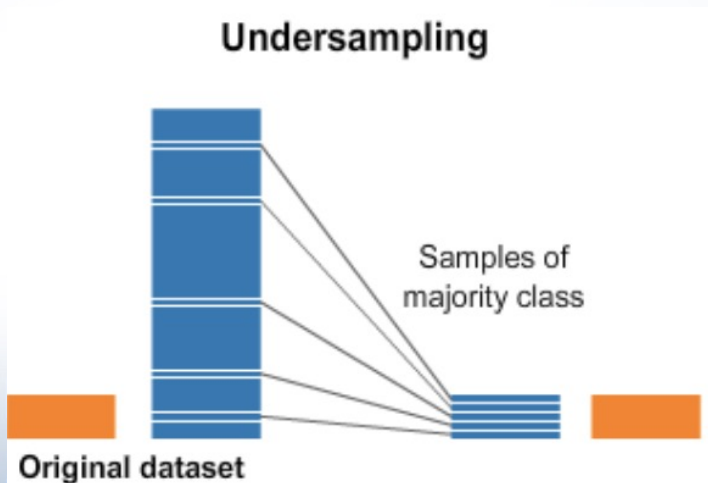


解决方案

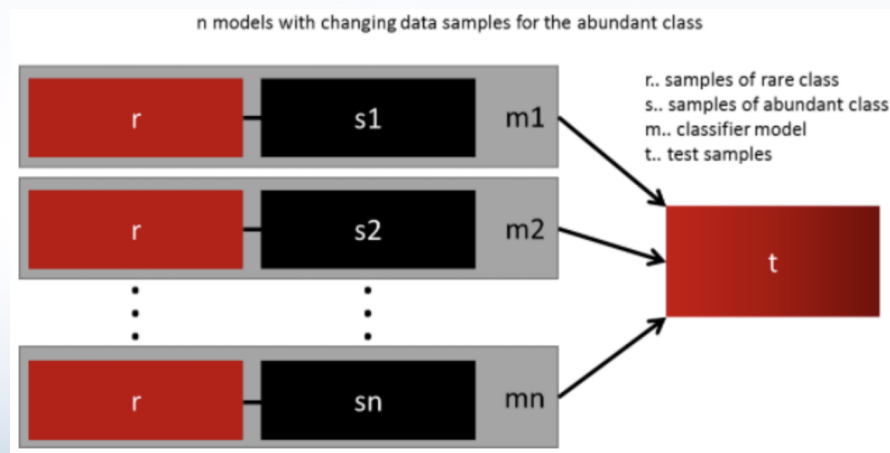


解决方案

- 样本极度不均衡
 - 降采样使得样本数量大致相同



- 采样后模型表现不稳定
 - 集成提高算法稳定性



研究成果

- 提交给DGraphFin(dgraph.xinye.com)且开源分享
 - Test AUC 0.8460
 - 相较于Leaderboard第一名(0.7761)提升9%
 - <https://github.com/storyandwine/GEARSage-DGraphFin>
- 复赛排名Top 10
- 首次提交第二名
 - 归纳学习方法可以更好的适应数据变化

消融研究

| 消去 | Test AUC |
|--------|----------|
| 增量学习表征 | 0.7806 |
| 手工构造特征 | 0.7912 |
| 边的方向 | 0.8399 |
| 边的时间 | 0.8404 |
| 边的种类 | 0.8426 |
| 邻域信息衰减 | 0.8447 |
| None | 0.8460 |

未来工作

路漫漫其修远兮 吾将上下而求索

未来工作

- 动态图模型
- 图净化方法——扰动边
- 社区搜索
- 个性化聚合函数
- 频域类模型研究
及其在归纳学习上的优化



THANK YOU

