2021 MAXP 基于DGL的图机器学习大赛解决方案

队伍：Graph@ICT

2021.12

队伍介绍-Graph@ICT

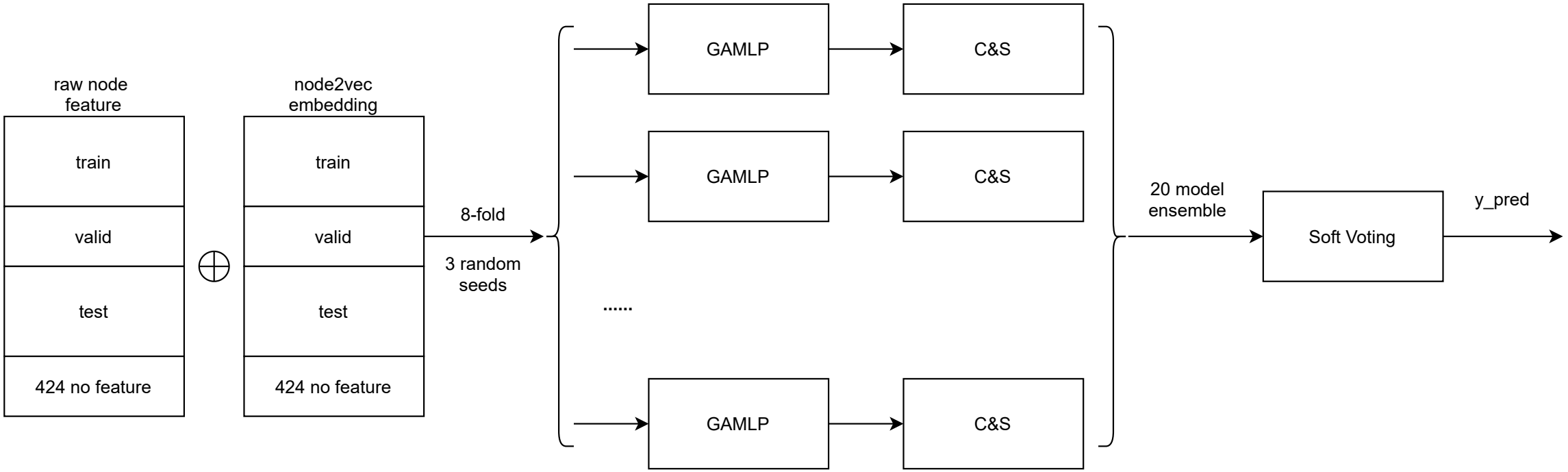
- 队长：迟慧璇[1] chihuixuan21@mails.ucas.ac.cn
- 队员：王玉莹[1]、黄琳焱[2]、孙志豪[1]

- [1]中科院计算所
- [2]厦门大学信息学院

整体框架-Scalable GNN

https://github.com/ytchx1999/MAXP_DGL_Graph

- 最终Test（b榜）：rank8（铜奖）
- Valid（a榜）：rank14



解决方案介绍-数据

- 原始数据处理
  - jupyter文件：process-*.ipynb、 gen_test_submitcsv.ipynb
  - 生成DGL的图格式

- Pre-processing
  - 生成node2vec embedding，并和原始特征进行拼接
  - 使用邻居聚合为424个没有feature的节点构造feature
  - GAMLP进行preprocess（类似Scalable GNN通常的做法）

解决方案介绍-模型

- Model：GAMLP (Graph Attention Multi-Layer Perceptron)
  - https://arxiv.org/abs/2108.10097
  - 8-fold cross validation with 3 random seeds

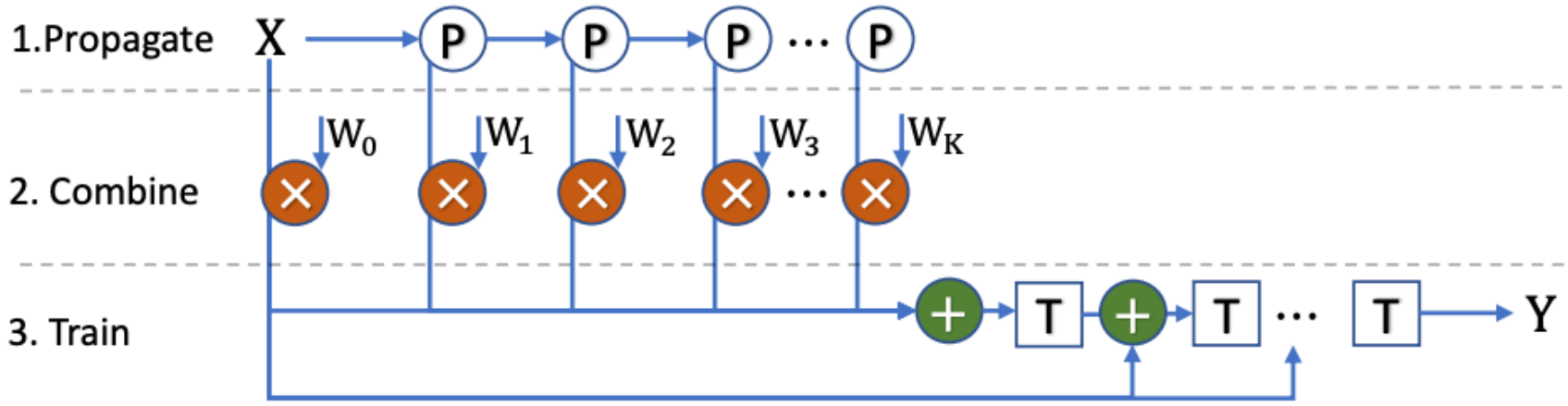


Figure 2: Overview of the proposed GAMLP, including (1) feature propagation, (2) feature combination with RF attention, and (3) MLP training. The feature propagation can be pre-processed.

解决方案介绍-后处理

- Correct and Smooth (C&S)
  - https://arxiv.org/abs/2010.13993
  - 每一折都进行C&S

- Ensemble
  - 为防止过拟合，只进行了简单的Soft-Voting（Average）

相关尝试及结果

- Valid (a榜)
  - SAGN+SE：比GAT效果好，速度较快
  - GAT：效果较差，速度较慢
  - GAMLP：效果最好，速度较快：最好的valid提交，rank14
- 结果

Model	Score
GAMLP (leaky-relu, 9 hops, 8-fold) + node2vec + C&S (DAD, AD) + Model Merge (+GAMLP_8fold_seed_{0-2})	55.53829808307
GAMLP + node2vec + C&S + Model Merge (+SAGN-SE, +SAGE, +GAT)	55.0070680604702
SAGN + node2vec + SE + Model Merge (+GAT, +SAGE) + C&S	54.5420166932282
GAT + node2vec + FLAG + C&S + Model Merge (+SAGE, +GCN)	54.2394856973069

相关尝试及结果

- Test (b榜)
  - GAMLP+node2vec (20 ensemble)
  - GAMLP (20 ensemble)：最好的test提交，rank8
  - GAMLP (8 ensemble)
- 结果

Model	Score
GAMLP (leaky-relu, 9 hops, 8-fold) + **node2vec** + C&S (DAD, AD) + Model Merge (+GAMLP_8fold_seed_{0-2})	49.7822086481499
GAMLP (leaky-relu, 9 hops, 8-fold) + C&S (DAD, AD) + Model Merge (+GAMLP_8fold_seed_{0-2})	49.7923833548815
GAMLP (leaky-relu, 9 hops, 8-fold) + C&S (DAD, AD) + Model Merge (+GAMLP_8fold_seed_{0})	49.7767704428278

谢谢！