METAPATH-BASED LABEL PROPAGATION FOR LARGE-SCALE HETEROGENEOUS GRAPH

Qiuying Peng
Topology Lab
Data Intelligence Department
OPPO Research
pengqiuying@oppo.com

Wencai Cao
Topology Lab
Data Intelligence Department
OPPO Research
caowencai@oppo.com

Zheng Pan
Topology Lab
Data Intelligence Department
OPPO Research
panzheng@oppo.com

June 11, 2021

ABSTRACT

MAG240M-LSC is the first large-scale heterogeneous academic graph extracted from the Microsoft Academic Graph (MAG) dedicated to the task of semi-supervised node classification. The complexity and efficiency of current best baseline model are unsatisfactory. Meanwhile, methods involving label propagation have shown great potential in performance gain. Our proposed model, MPLP (**M**eta**P**ath-based **L**abel **P**ropagation), combines efficient scalable metapath-based random walk and label propogation to yield excellent performance in the node classification task.

Keywords Label Propagation · Metapath · Graph Neural Network · Node Classification

1 Introduction

In recent years, machine learning on graphs is prevailing, as graph-structured data is widely used in real-world areas such as text classification, recommender systems, knowledge graphs and many others. Graph Convolutional Networks (GCNs) [1] and subsequent variants which generalize classical convolutional architectures (CNNs) to graph-structure data, has emerged as frequent winners to the major graph benchmarks. However, most of these models were developed and evaluated on relatively small datasets due to the need of placing the graph into memory during full-batch training. Although many graph sampling methods have been proposed, most of them still suffer from inadequate computation efficiency and efficacy. Researchers have developed various techniques in simplifying GNNs to improve their scalability via pre-computing graph structures and utilizing neighbor-averaging features [2, 3] as well as combining GCNs with Label Propagation (LP). MAG240M-LSC[4] is a heterogeneous academic graph extracted from the Microsoft Academic Graph (MAG) which aims to predict the subject areas of papers whose features are represented by their RoBerta[5] embedding of titles and short descriptions. However, such representations usually live in a concentrated subspace and suffer from low separability.

Inspired by [2, 3, 6, 7] and to better introduce label information, we propose a novel model MPLP (Metapath-based Label Propagation) which combines label propogation and scalable metapath-based random walk techniques. Specifically, MPLP extracts label propogation features from different types of metapath-based topologies, and integrates them into subsequent classifier such as MLP, GCN, GAT, etc. Given the label imbalance and time-evolving characteristics of MAG240M-LSC dataset, we also design a label weighting scheme for training and propose a dynamic finetuning method to address these problems. Currently, our proposed MPLP model ranks among top-3 in the MAG240M-LSC node level prediction challenge.

2 Methodology

In this work we propose MPLP, a metapath-based label propagation for large-scale heterogeneous graph. The key idea of MPLP is to propagate labels by specified metapaths with random walk. In the first stage, we perform label

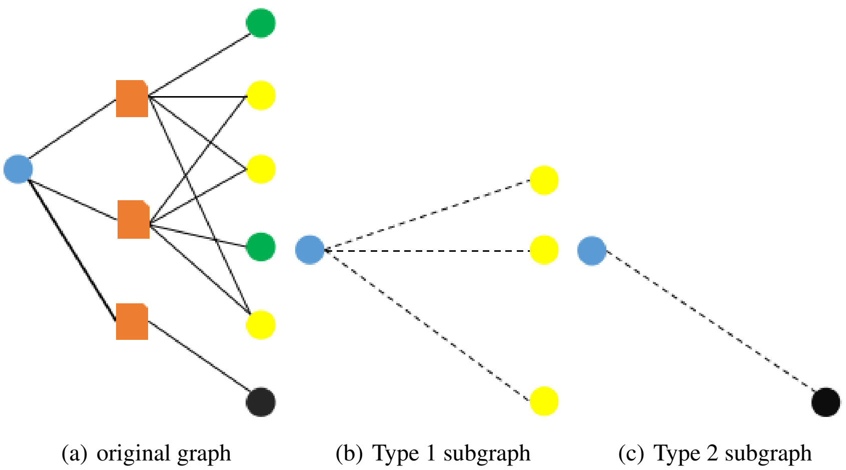

Figure 1: Different strategies of MPLP on paper-write_by-author-write-paper (P-wb-A-w-P) meta-path. (a) original graph. (b)Type 1 subgraph contains two-hops paper nodes sharing more than 2 authors (three yellow nodes). (c) Type 2 subgraph contains nodes sharing the author who writes the fewest paper.

propagation in different heterogeneous meta-paths including P-wb-A-w-P (paper-write_by-author-write-paper), P-c-P (paper-cite-paper), P-cb-P-c-P (paper-cite_by-paper-cite-paper) and so on. Furthermore, to acquire more adequate homophily information without noise in each meta-path, several ways of label propagation are carried out within pre-set subgraphs like type 1, type 2 subgraph (see Figure 1). In the second stage, we concatenate information propagated from all meta-paths and feed them into the final classifier.

For node-wise classification tasks, our architecture has the form (see Figure 2):

$$\boldsymbol{Z} = \sigma([\boldsymbol{H}_{emb}, \boldsymbol{X\Theta}_0, H_{p1}\boldsymbol{X\Theta}_1, ..., H_{pk}\boldsymbol{X\Theta}_k, H_{p1}\boldsymbol{Y\Theta}_1, ..., H_{pr}\boldsymbol{Y\Theta}_r])$$

$$\boldsymbol{Y}^* = Classifier(\boldsymbol{Z})$$

where $\boldsymbol{X}$ and $\boldsymbol{Y}$ denote features and labels. H_{pk} denotes metapath k for $\boldsymbol{X}$ and $\boldsymbol{\Theta}_k$ is the corresponding learnable parameter, and the same works in H_{pr}, $\boldsymbol{\Theta}_r$ for $\boldsymbol{Y}$. $\boldsymbol{H}_{emb}$ is graph embedding from supervised or unsupervised model. In this work, we get $\boldsymbol{H}_{emb}$ from pre-trained R-GAT and set $\boldsymbol{\Theta}_0$ to identity matrix $\boldsymbol{I}$, and then concatenate all the fearures for MLP classifier.

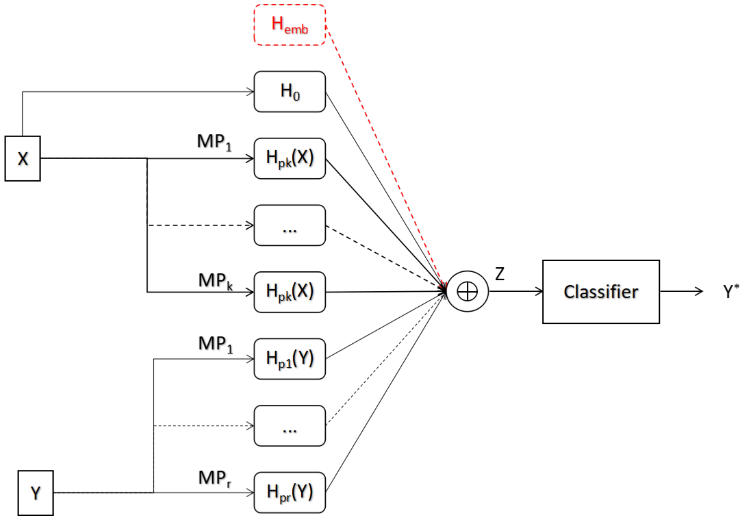

Figure 2: MPLP model.

3 Experiments

It is observed that three challenges exixts when dealing with MAG240M-LSC dataset:

1. the graph involves more than 240 million nodes and 3 billion edges
2. the distributions of papers' subjects or labels vary greatly across years
3. the number of samples among different subjects is extremely imbalanced

Concerning the first challenge, our MPLP model has a natural advantage in scalability and efficiency in that the inputs for training can be calculated in advance, which is similar to SIGN and NARS. To tackle the problem of unidentical distributions of labels across years, we finetune our trained model on the latest two years (2018 and 2019). With regard to the imbalanced number of samples, we manually design a weight for each class as the following function:

$$weight = N_{class} \times normalise(log_{10}(\frac{cnt_{2018}}{cnt_{<=2018}} + \boldsymbol{\alpha}))$$

where cnt is a vector of dimension N_{class} with each element representing the total number of specific subject papers published in certain years, $\boldsymbol{\alpha}$ is a hyperparameter which is chosen to be 5 in our model, and $N_{class} = 153$ for the current dataset.

Table 1 demonstrates our intermediate experiments w.r.t. different input features. *label* denotes label propagation information from $\boldsymbol{Y}$, *feat* denotes feature propagation information from $\boldsymbol{X}$, and *R-GAT* or *Line-2nd* embedding is generated through pre-training. In consideration of both model complexity and accuracy on validation set, we choose MPLP (*label + R-GAT embeddings).*

Table 1: Performance on Different Models

Model	Valid Acc($\%$)	Valid Acc($\%$) fine-tuned	Valid Acc($\%$) fine-tuned+class weight	Parameters
MPLP (*label)*	74.60	75.31	-	614,169
MPLP (*label + feat)*	74.67	75.40	-	908,953
MPLP (*label + R-GAT embeddings)*	75.24	75.82	75.94	743,449
MPLP (*label + feat + R-GAT embeddings)*	75.24	75.96	-	1,018,553
MPLP (*label + feat + R-GAT embeddings + line-2nd embeddings)*	75.41	75.99	-	1,061,817

One special circumstance in this contest is that we only have one chance to submit our predictions for partial testing. Thus, overfitting is a potential risk that may greatly influence the performance of our model. To alleviate overfitting, we implement 5-fold cross validation for training, repeat the process for several times with different random seeds and ensemble all the models' outputs through averaging for final prediction. As an evaluation of this method, we view the data of 2018 as validation set and the data of 2019 as test set. After repeating 5-fold cross validtion with 4 random seeds, the validation accuracy on 2018 is 0.7770 ± 0.0003 *and the ensembled accuracy is 0.7794, while the test accuracy on 2019 is 0.7605.*

We compare our proposed method with official strong baselines, as shown in Table 2. Our proposed method outperforms other methods in validation and test dataset.

Table 2: Performance on MAG240M-LSC

Model	Valid Acc($\%$)	Test Acc($\%$)	Parameters
MPLP	**76.69** $\pm$ 0.03(ensemble 76.96)	**74.47**	743,449
R-GAT	70.02	69.42	12.2M
R-GraphSAGE	69.86	68.94	12.3M

4 Conclusion

In this paper, we study subject prediction problem with scalable heterogenous graphs by comprehensively exploring label information within various metapath topologies in academic scenario. Inspired by NARS, UniMP and label reuse *methods, we propose a novel MPLP model which combines label propogation and scalable metapath-based random walk techniques. MPLP could extract label propogation features within different scale of metapth-based topologies beforehand, which could be utilized by various following methods (e.g., MLP, GCN, GAT, etc.). Furthermore, we propose a time-based finetune method to tackle time-evolving problems.*

This work shows that label information within different metapath-based topologies worths further study. However, label propagation from manually-designed metapath may limit the performance. Therefore, automatic metapath-based label propagation should be a promising area in which we will explore further. Meanwhile, different metapaths could embrace different level of importance and so that attention mechanism may help to improve model accuracy and interpretability.

References

[1] *Thomas N. Kipf and Max Welling. Semi-Supervised Classification with Graph Convolutional Networks.* ICLR*, 2017.*

[2] *Felix Wu, Tianyi Zhang, Amauri Holanda de Souza Jr., Christopher Fifty, Tao Yu, and Kilian Q. Weinberger. Simplifying Graph Convolutional Networks.* ICML*, 2019.*

[3] *Fabrizio Frasca, Emanuele Rossi, Davide Eynard, Ben Chamberlain, Michael Bronstein, and Federico Monti. SIGN: Scalable Inception Graph Neural Networks.* CoRR*, 2020.*

[4] *Weihua Hu, Matthias Fey, Hongyu Ren, Maho Nakata, Yuxiao Dong, and Jure Leskovec. OGB-LSC: A Large-Scale Challenge for Machine Learning on Graphs.* arXiv:2103.09430 [cs]*, 2021.*

[5] *Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach.* ICLR*, 2019.*

[6] *Lingfan Yu, Jiajun Shen, Jinyang Li, and Adam Lerer. Scalable Graph Neural Networks for Heterogeneous Graphs.* ICLR*, 2020.*

[7] *Yunsheng Shi, Zhengjie Huang, Shikun Feng, Hui Zhong, Wenjin Wang, and Yu Sun. Masked Label Prediction: Unified Message Passing Model for Semi-Supervised Classification.* ICLR*, 2021.*