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 inadequete 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 charicteristics 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(see Figure 1), 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. For node-wise classification tasks, our architecture has the form:

$$Z = \sigma([X\Theta_0, P_1 Y\Theta_1, ..., P_r Y\Theta_r])$$

$$Y^* = Classifier(Z)$$

where $\Theta_0, ..., \Theta_r$ are learnable parameters, F is the aggregating function and X_m is hidden embedding from pre-trained supervised or unsupervised model. In this work, we choose concatenate all the fearures from label propagation with model embedding from semi-superviced R-GAT for MLP classifier.

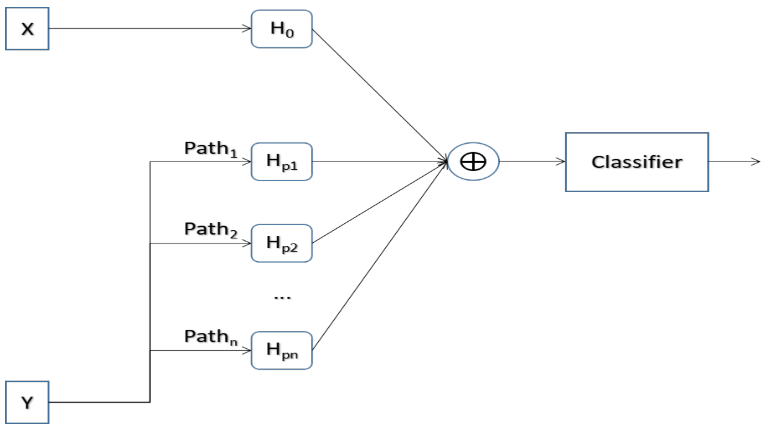
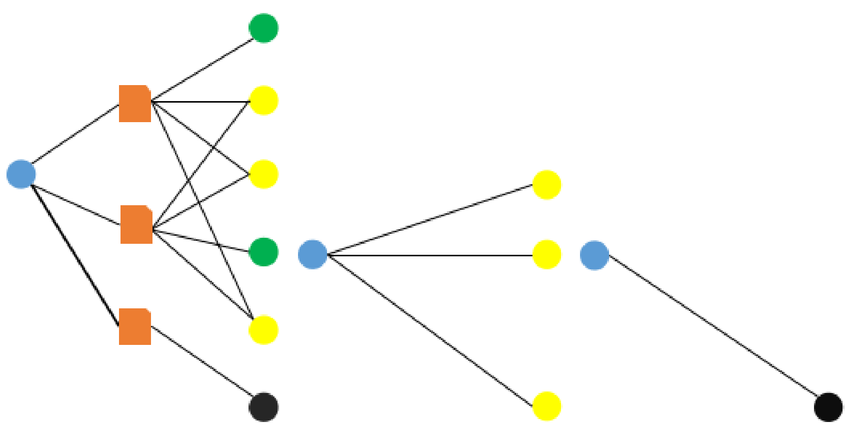


Figure 1: MPLP model.

In the first stage, we perform label propagation in different heterogeneous meta-paths including P-A-P (paper-write_by-author-write-paper), P-P (paper-cite-paper), P-C-P-C-P (paper-cite-paper-cite-paper) and so on. Furthermore, subgraphs are extracted within each meta-path to imporve SNR(Signal to Noise Ratio). To acquire more adequete homophily information without noise in each meta-path, several ways of label propagation are carryed out within pre-set subgraphs like >=2 subgraph, influencial author subgraph(see Figure 2) for denoising. In the second stage, we project label distributions from all meta-paths by an MLP separately and concatenate them with node original features as input for the final classifier.



(a) original graph (b) Type 1 (c) Type 2

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

3 Experimets

Three challenges emerge 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:

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

This function assign higher values to the classes which appear frequently in the year 2018 but rarely before 2018. We utilize the class weights in the calculation of loss function.

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 several strong baselines, as shown in Table 1. Our proposed method outperforms other methods in validation and test dataset.

Table 1: Performance on MAG240M-LSC

Model	Valid Accuracy	Test Accuracy	Parameters
MPLP	0.7669 ± 0.0003(ensemble 0.7696)	74.47	743449
R-GAT	70.02	69.42	12.2M
R-GraphSAGE	69.86	68.94	12.3M

4 Conclution

In this paper, we study subject prediction problem with scalable heterogenous graphs for 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. Experiements show that MPLP outperforms previous methods in mag240m datasets, including RGAT, UniMP etc.

This work shows that label infomation within different metapath-based topologies worths further researches on the graph. However, metapath-based label propogation is designed manually, which may limit the performance. Automatic metapath-based label propogation should be a promising area which we will explore continually. Meanwhile, different metapaths should show different importance. With certain attention methods which perform attention method on features of different metapaths, we could get the weights of metapaths, which may help us explain the influence of labels or node attributes of certain metapaths and further improve the interpretability of the network.

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.