Semi-Supervised Graph Federated Learning with a Lack of Labeled Data

Fan Yang, Ha Young Kim, Won-Yong Shin
Yonsei University

vanyeung@yonsei.ac.kr, hayoung.kim@yonsei.ac.kr, wy.shin@yonsei.ac.kr

Abstract

Existing graph federated learning (GFL) methods are limited mostly to supervised training settings, which may be often impractical for many real-world graph-related tasks since labeling a reasonable number of labels across different clients is both time-consuming and costly. When there are only a few labeled data in GFL scenarios, the performance is expected to be significantly degraded. To solve this problem, we propose a semi-supervised graph federated learning method using energy-based loss, which is robust to insufficiently labeled settings. Experiments on several datasets demonstrate the superiority of our method over competing graph federated learning method.

Ⅰ. Introduction

Graph federated learning (GFL), such as FedSage [1], focuses mainly on cross-entropy loss to train graph neural networks (GNNs) at each client, overlooking the insufficiency of labeled data. The insufficiency of labels inherently leads to a decrease in model performance. To tackle this challenge, we propose a semi-supervised GFL method that employs a multilayer perceptron (MLP) model at each client. Unlike the prior study in [1], the MLP model at each client is trained using both cross-entropy loss and energy-based loss. The energy-based loss [2] serves as an unsupervised learning function to leverage the unlabeled data, enabling the model to be robust in case where there is a lack of labeled data.

Ⅱ. Methodology

We consider a whole graph G divided into multiple subgraphs, each of which is assigned to each client. We aim to train the local MLP model more precisely at each client for node classification. Our method performs the following steps iteratively until convergence:

- **Step 1**: Each client initializes its MLP parameters;
- **Step 2**: Each client i trains and updates its parameters on the dedicated subgraph with two loss terms;
- **Step3**: Each client sends its updated parameters to the server;
- **Step4**: The server aggregates these parameters by simple averaging and broadcasts the aggregated parameters to all clients identically.

Ⅲ. Experimental Results

We assess the performance of node classification according to different proportions of labeled nodes as training data on the Cora and Citeseer datasets. We set the number of clients as 5. Each dataset is split into training/validation/test sets, where {1,10,20}% of the given dataset is used for training. Table 1 summarizes the performance of our method in comparison with the state-of-the-art GFL method, FedSage [1]. It is observed that our method substantially outperforms FedSage especially when portion of training is low. We also compare our method with energy-based loss with the one with standard contrastive loss used for unsupervised learning in [3], while keeping the architecture and training procedure unchanged. It is shown that the proposed method using energy-based loss is consistently superior to its counterpart (the case of contrastive loss).

Table 1. Classification accuracy.

Dataset	Method	Portion of training		
		1%	10%	20%
Cora	Proposed (w/ energy-based loss)	**0.58**	**0.71**	**0.79**
	Proposed (w/ contrastive loss)	0.45	0.69	0.78
	FedSage	0.33	0.51	0.55
Citeseer	Proposed (w/ energy-based loss)	**0.53**	**0.71**	**0.73**
	Proposed (w/ contrastive loss)	0.43	**0.71**	0.72
	FedSage	0.34	0.62	0.70

ACKNOWLEDGMENT

This research was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2021R1A2C3004345, No. RS-2023-00220762) and by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2021-0-00347, 6G Post-MAC (POsitioning- & Spectrum-aware intelligenT MAC for Computing & Communication Convergence)).

REFERENCES

[1] Zhang *et al.*, "Subgraph federated learning with missing neighbor generation," in *Proc. NeurIPS* (2021).

[2] Shin *et al.*, " Edgeless-GNN: Unsupervised Representation Learning for Edgeless Nodes," *IEEE Transactions on Emerging Topics in Computing (2024).*

[3] Hu *et al.*, "Graph-MLP: Node classification without message passing in graph," arXiv preprint *arXiv:2106.04051* (2021).