Handling Sparse and Noisy Labels in Deep Graph Learning

by

Yayong Li

A thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

Under the supervision of Professor Ling Chen

Faculty of Engineering and Information Technology

University of Technology Sydney

November 2022

Declaration

I, Yayong Li declare that this thesis, is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the Faculty of Engineering and Information Technology at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis. This document has not been submitted for qualifications at any other academic institution. This research is supported by the Australian Government Research Training Program.

Signature: Production Note: Signature removed prior to publication.

Date: 11/17/2022

I would like to dedicate this thesis to my loving wife and parents ...

Acknowledgements

I would like to express my sincere gratitude to my supervisors, Prof. Ling Chen and A/Prof. Jie Yin, who have provided me with tremendous support and guidance during my PhD study.

As my principal supervisor, Ling has provided me with the opportunity to study in the Australian Artificial Intelligence Institute (AAII) at University of Technology Sydney (UTS) where I started a unique journey. At the beginning of my PhD, Ling led me to the academic path and gave me the promising research direction on graph representation learning. From scholarship application and research guidance to thesis submission and job recommendation, Ling has been always willing to offer any form of help to support my research and career development. She has also provided many valuable opportunities that remarkably broaden my horizon and gain my experiences. When I was depressed or disappointed because of experiment frustrations during my research, her encouragement cheered up my spirit and rebuilt my confidence.

I am also extremely grateful to my co-supervisor A/Prof. Jie Yin for guiding me throughout my PhD experience. During the four years, she has generously imparted the essential knowledge and expertise to me, which allows me to gradually build up solid academic skills from scratch and get me well prepared for the independent research in the next step of my career. Jie has a strong sense of responsibility. In our weekly meetings, she has been always patient to discuss every detail of my research with me, and her valuable advice can always help me tackle my problems more effectively. Despite having busy schedules, she can still devote her time to helping me resolve my research issues, review my presentation, revise and proofread our papers. The knowledge from her advice is only a small portion of what I learned from her. Her rigorous academic attitude, integrity, and irrepressible passion for research have deeply impacted and inspired me. Her behaviours have set a great example for my future career. Words cannot express my gratitude to my supervisors, and this thesis would not have been possible without them.

I am also grateful to my senior students Dr. Daokun Zhang and Dr. Wei Wu. When I was new to the school at UTS, their generous help and useful suggestions allowed me to avoid many mistakes and accommodate to the new academic life in Sydney faster. I also learned so much from my friend Xiaowei Zhou. His clear thought process and pleasant personality impressed me, and our every discussion inspired me with new perspectives and helped me get unstuck during research. Thanks should also go to my peer students, lab mates and other friends, including Lan Wu, Wei Huang, Shaoshen Wang, Yunqiu Xu, Leijie Zhang, Lu Huo, Tianyu Liu, Tao Zhang, Ying Tian, Wei Lin, Shulin Chen, Hongyang Zhang, Xinzhu Li, Yang Xu and Mingfei Tong. We shared a lot of experiences with laughter and happiness, which enriched my life in Sydney.

I would also like to acknowledge the financial support from the CSC-UTS scholarship, which significantly relieved my burden on living expenses and allowed me to pursue my degree dedicated.

Above all, I would like to express my special thanks to my wife. My wife has kept accompany with me for almost my whole PhD journey in Sydney, where we fell in love with each other and got married witnessed by my supervisors and all our friends. She has been the one with whom I have shared the most of my emotions and feelings, regardless, laughter or tears, happiness or sadness, excitement or depression, faith or frustration. There have been countless moments when I encounter setbacks in my experiments, her understanding and encouragement relieved my pressure and kept up my spirit to stick with it. She can always believe in me and fully support my decision, which gave me much courage and confidence to cope with varieties of difficulties and challenges. I am so fortunate to have her accompanying, which made my PhD journey much easier and happier.

The special thanks should also go to my parents, my sisters and my parents-in-law, for their constant and unconditional love. Although being thousands of miles away from each other in different countries, I can still feel their care and love. Our phone chat every week

was my most relaxing moment when I can put my research on hold for a while and just enjoyed the pleasing time. Their kind regards and words can always comfort me and sweep away all my depression and anxieties. They always gave me unwavering help and support with all their hearts, and no words can express my deepest gratitude for their endless love and self-sacrifice.

To them, I dedicate the dissertation.

Abstract

Nowadays, there are growing amounts of graph-structured data emerging from a broad variety of information industrial applications, such as social networks, financial networks, biomedical networks, traffic networks, and so on. The complex topological information among those graph nodes, along with their content-rich node attributes, pose a great challenge for data mining and analysis. Recently, Graph neural networks (GNNs) have been proposed as a novel learning paradigm to deal with graph-structured data, and have achieved a great success on a variety of graph-based tasks, especially on the node classification task. However, its success highly relies on the sufficient number of high-quality labels, which is often difficult to attain in the real world. On the one hand, acquiring node annotations is labour-intensive, time-consuming, and usually costs a lot of expenses for recruiting or paying annotators. This results in the label sparsity problem for GNNs learning. On the other hand, wrong labeling is almost inevitable while annotating nodes due to inter-observer variability, human annotator error, or errors in crowdsourced annotations[60]. Under this situation, GNNs are prone to overfitting to these corrupted labels, thereby leading to poor generalization abilities.

Considering these label-associated challenges, this thesis is developed to handle the label sparsity and label noise problem on graphs. Confronting the label sparsity problem on graphs, I first resort to Active Learning (AL) to improve the model performance. Within the limited labeling budget, AL can selectively construct the most informative label set for model training by querying labels for the most valuable nodes in the graph. Then I focus on the research of Pseudo-Labeling (PL) to relieve the label sparsity problem. It explores to fully exploit the unlabeled nodes to complement the severe lack of label information, and apply label augmentation techniques to enhance information propagation among graph nodes. Finally,

to cope with the label noise problem, I turn to the research of label-noise representation learning in GNNs, expecting to establish a robust GNN model that can effectively detect suspicious labels and minimize their influence on model training. Therefore, in this thesis, I would specialize in the three specific research topics and make efforts to effective solutions for them correspondingly.

In terms of Active learning, it aims to boost the labeling efficiency by selecting the most informative nodes for querying their labels, such that the selected nodes can maximize the model performance. Although AL has been widely studied for alleviating label sparsity issues with the conventional independent and identically distributed (IID) data, how to make it effective over attributed graphs remains an open research question. In Chapter 4, a SEmi-supervised Adversarial active Learning (SEAL) framework is proposed on attributed graphs, which fully leverages the representation power of GNNs and designs a novel AL query strategy in an adversarial way for node classification. Extensive experiments on real-world networks validate the effectiveness of the SEAL framework with superior performance improvements to state-of-the-art baselines on node classification tasks.

Pseudo-Labeling has been proposed to explicitly address the label scarcity problem. It aims to augment the training set with pseudo-labeled nodes so as to re-train a supervised model in a self-training cycle. However, the existing pseudo-labeling approaches often suffer from two major drawbacks. First, they tend to conservatively expand the label set by selecting only high-confidence unlabeled nodes without assessing their informativeness. Unfortunately, those high-confidence nodes often convey overlapping information with given labels, leading to minor improvements for model re-training. Second, these methods incorporate pseudo-labels into the same loss function with genuine labels, ignoring their distinct contributions to the classification task. In Chapter 5, a novel informative pseudo-labeling framework is proposed to facilitate learning GNNs with extremely few labels taking both informativeness and reliability of pseudo labels into consideration. Extensive experiments on six real-world graph datasets demonstrate that the proposed approach remarkably outperforms state-of-the-art pseudo-labeling and self-supervised baseline methods on graphs.

Label-noise representation learning has also been primarily studied on the tasks with IID data, such as the image classification task, but very little research effort has been on how to improve the robustness of GNNs in the presence of label noise. Furthermore, the graph topological information poses unique challenges when dealing with label noise - label sparsity and label dependency. To tackle these challenges, a unified framework is proposed to robustly train GNN models against label noise under the semi-supervised setting in Chapter 6. The key idea is to perform label aggregation to estimate node-level class probability distributions, and then use them to guide sample reweighting and label correction simultaneously, so as to reduce model sensitivity towards noisy labels. Experimental results on real-world datasets have been conducted to demonstrate the effectiveness of proposed algorithm with regard to different levels and types of label noise.

Table of contents

List of figures

List of tables