Project Proposal

Team members: Eduardo Wang Zheng, 黎元, 曾丰

Introduction

The **node classification** task is one where the algorithm has to determine the labelling of samples (represented as nodes) by looking at the labels of their neighbours. **Link Prediction** in a network is how to predict the likelihood of a link between two nodes in a network that have not yet produced a link, such as known network nodes and network structures. These two methods have played an important role in network science.

In this project, we intend to solve the following two academic network problems with the multi-classification of nodes and link prediction between nodes：

1)Judge which of the 10 meetings each scholar has published in 2016-2019

2)Determine whether two scholars have collaborated in publishing articles in 2020

Related works

There have been multiple studies about node classification and link prediction.

For link prediction, since Sarukkai[1] first applied the Markov chain to link prediction and path analysis in the World Wide Web in 2000, research on link prediction has been going on for over 20 years. Hasan[2] et al. reviewed some representative social network link prediction methods, and mainly considered three types of models according to the classification of the network: two-class model, probability model and linear algebra model.

As for node classification, Bingbing Xu[3] propose label-consistency based graph neural network , leveraging node pairs unconnected but with the same labels to enlarge the receptive field of nodes in GNNs. But this method sometimes brings out-of-memory problem. Recently, Zou Haodong[4] propose a sequence correlation preserving method for attributed network embedding which transforms the network properties into three types of sequences and preserves the correlations among them.

Plan: Using DeepWalk for node and link prediction

1. Load Dataset

Algorithm:

1) Read in **author_paper_all_with_year.csv** and **labeled_papers_with_authors.csv**

2) Encapsulate attributes (**paper_id, author_id, label**) into a class **author** (set the label as "?" if the corresponding **author_id** does not have **label**)

3) Read in **paper_reference.csv**

4) Construct homogeneous network **G** (using attribute **paper_id**) and store the connected

authors' **author_id** into a list **relation**

2. Random Walk

Algorithm:

1) Defined a function **get_randomwalk** that will take a node and length of the path to be traversed as inputs. It will walk through the connected nodes from the specified input node in a random fashion. Finally, it will return the sequence of traversed nodes

2) Capture the **random walk sequences** for all nodes in the graph **G**

3. Train the skip-gram (word2vec) model with random walks

Algorithm:

1) Set some parameters and initialize the **word2vec** model

2) Build vocabulary from the **random walk sequences**

3) Train the **skip-gram (word2vec)** model

4. Make predition

4.1 link prediction

Algorithm:

1) Read in **author_pairs_to_pred_with_index.csv**

2) Store the **prediction** for every **author_pair (node pair)** and **index** into a .csv file

4.2 Node prediction

Algorithm:

1) Read in **authors_to_pred.csv**

2) Find the highest prediction value (means the corresponding node is the most similar to the current node) and the corresponding node for every **author_id (node)**

3) Store the **label** of the corresponding node and the **author_id** current node into a .csv file

Backup plan

Heterogeneous network modelling

1.Build network:

1) Read in **labeled_papers_with_authors.csv** to build graph of paper nodes and author nodes. An edge is put between an author and a paper if the paper belongs to the author. The conference are labeled in paper nodes.
2) Read **paper_reference.csv** to include directed "reference" edges.
3) Pass the conference label from paper nodes to author nodes by each paper nodes.

2.Random walk with normalized probability

Like random walk algorithm, but the probability of each step is normalized according to each neighbors of current nodes. The detailed calculation is left out here.

3.Train a metapath2vec model

Repeatedly obtain random walk sequence. This model is basically the skip-gram model of word2vec.

4.Make prediction

Taking **author_pairs_to_pred_with_index.csv** or **authors_to_pred.csv** as input, we will obtain the prediction result for both problems.

Schedule

Week 8~10 : Making proposal
Week 11~14 : Coding
Week 15~16 : Presentation

Reference

[1] RAMESH R S.Link prediction and path analysis using markov chains[J]. Computer Networks,2000,60(33):377-386
[2] HASAN M A,ZAKI M. A surwey of link prediction in social networks[J]. Social Network Data Analytics,2011,40(5):243-275
[3] Bingbing Xu,Junjie Huang,Liang Hou,Huawei Shen,Jinhua Gao,Xueqi Cheng. Label-Consistency based Graph Neural Networks for Semi-supervised Node Classification[P]. Research and Development in Information Retrieval,2020.
[4] Zou Haodong,Duan Zhen,Guo Xinru,Zhao Shu,Chen Jie,Zhang Yanping,Tang Jie. On embedding sequence correlations in attributed network for semi-supervised node classification[J]. Information Sciences,2021,562.