Fea2Fea: Exploring Structural Feature Correlations via Graph Neural Networks

Jiaqing Xie[1], Rex Ying[2]

[1]School of Engineering & Informatics, University of Edinburgh
[2]SNAP Lab, School of Computer Science, Stanford University

ECML PKDD 2021

Stanford Computer Science

Motivation

Let's think about **an interesting question:** In LesMiserables dataset, if we already know character **Valjean** has 35 ***Degrees***, can we predict ***PageRank*** or ***Clustering Coefficient*** of to measure Valjean's importance in the regional relationship? Besides, how to measure if a structural feature is easier to predict is mainly considered in this research.

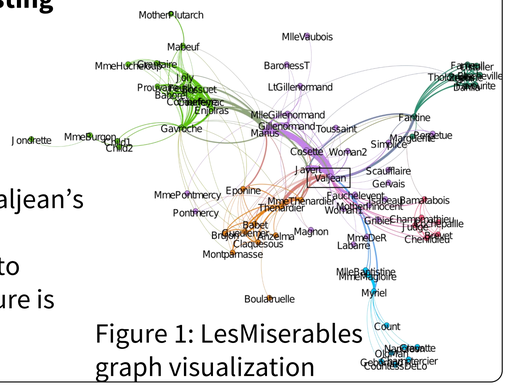


Figure 1: LesMiserables graph visualization

Introduction

Structural features are important features in a geometrical graph. Although there are some correlation analysis of features based on covariance, there is no relevant research on structural feature correlation analysis with graph neural networks. In our paper, we propose a framework based on GNN models for processing graph structural feature(s) to structural feature prediction, simply called **Fea2Fea.**

Main Findings:

- **Node degree** is correlated with other features such as **pagerank** and **clustering coefficient. Clustering coefficient** is correlated with **pagerank.** Features are correlated with themselves generally**.**
- The expressive power of the graph neural netowrk in achieving the correlations of graph features, such as the ease with which other structural features to predict **node degree** and general **difficulty** to predict **pagerank** and **average path length.**

Main Advantages:

- illustrate feature correlations with **low dimensionalit**y, where the number of input dimensions is controlled to a maximum of 5.
- require an **empirically effective** constructed models to perform comparisons which means that we introduce the importance of adding non-correlated features other than defeating the state of the arts graph embedding methods.
- filter additional redundant node features via graph neural network based models **instead of** covariance explained models.

Methods

Fea2Fea-single

Firstly we obtain a feature matrix . Specifically, graph features have been indexed in order, which is given by constant feature, degree, clustering coefficient, pagerank and average path length. We extract two features each time, one for input and the other for output. We will reach a correlation matrix **R** since each feature can be taken both as input and output.

Fea2Fea-multiple

The value R(i, j) in the matrix R indicates whether feature f_i and feature f_j can be easy or difficult to predict each other. Moreover, we want to add more features to f_i to see if it will predict f_j more accurately. Therefore, we apply a threshold mechanism to filter irredundant structural features (above the threshold value). According to the feature concatenation in the embedding layers, we provide three methods: simple, bilinear and **Neural Tensor Network**(NTN). We record the possible feature groups for each predicted objective and its corresponding accuracy on prediction.

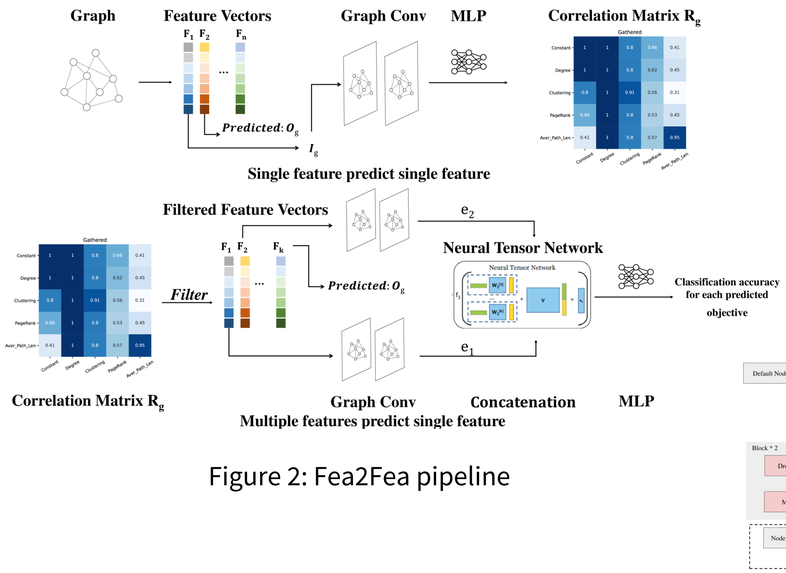
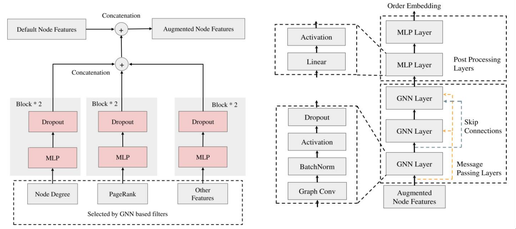


Figure 2: Fea2Fea pipeline

Model Architecture



Results

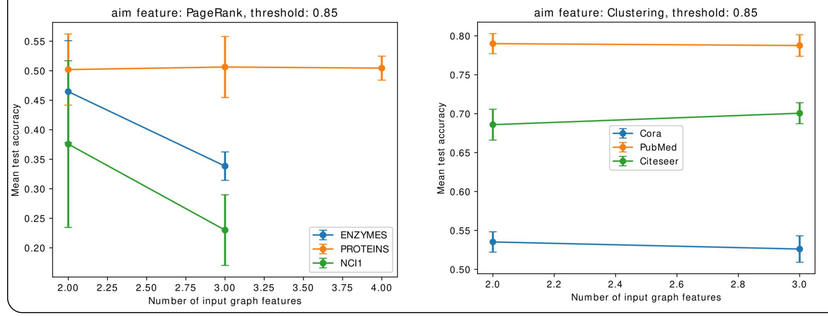
Table 1: Fea2Fea-single results (partial)

Task	Cora	Citeseer	PubMed	PROTEINS
Clu -> Deg	**1.000**	**1.000**	**1.000**	0.652
Deg-> PR	**0.792**	0.750	0.629	0.699
PR->AvgLen	0.421	**0.466**	0.459	0.172

Fea2Fea-single(GIN):

We figure out that **node degree** is the **easiest** to predict according to table 1. Predicting clustering coefficient and average path length is hard generally for both kinds of datasets.

Figure 3: : Fea2Fea-multiple results (partial)



Fea2Fea-multiple:

When predicting clustering coefficient from other features, possible combination number of features is only equal to 2 or 3. A possible 3-set concatenation for Cora is {**Cons, PR, Avglen**}, which is much better than a 1-set prediction for each feature.

Table 2: Real world applications(%)

Model	Enzymes	Proteins	NCI1
GCN	36.0	66.2	61.3
GAT	31.0	65.9	60.9
Fea2Fea-s2	**48.5**	**77.8**	**74.2**
Fea2Fea-b2	45.8	76.4	70.8
Fea2Fea-n2	42.8	74.9	68.5

With concatenating augmented structural features which are irredundant mutually in **tudatasets** with graph classification tasks, with original features, the accuracy is improved. NTN shows its powerfulness in some other tasks instead.

References

Xu, K., Hu, W., Leskovec, J., Jegelka, S.: How powerful are graph neural networks? Arxiv preprint arXiv:1810.00826 (2018)