AI506 Term Project
Ensemble Learning for Hyperedge Prediction using Graph Structure based on features in Hypergraph

20205007 Junhyeok Kang
20193154 ChangHwan Kim
20203221 Hyangsuk Min

Contents

1. Motivation

2. Method

3. Performance Evaluation

4. Conclusion

1. Motivation

Problem: Hyperedge prediction in large hypergraph

Challenge I. Scalability issue

- It is difficult to directly apply the methodology visiting the entire large graph such as Girvan-Newman and DeepWalk.
  >>> *We extract **graph structure based features** of hyperedge*

Challenge II. Low level representation

- It is limited to predict hyperedge using only graph structure based features of hyperedge because high level features cannot be considered.
  >>> *Abstract hyperedge through **n-order expansion** and extract graph structure based features for each abstracted hyperedge*

2. Method

Our method has the following steps:

*Step 1. Abstract hypergraph through **n-order expansion** to extract high level features*

*Step 2. Extract **graph structure based features** of n-projected graph*

*Step 3. Boost performance through **ensemble learning***

2. Method

n-order expansion

: We adopt n-order expansion which is a method of incrementally representing high-order interactions as n-projected graphs in a given hypergraph[1]. We extract high-order latent features from n-projected graphs of hypergraph.

graph structure based features

: We used 8 different features for hyperedge prediction and extracted each feature from n-projected graphs.

>> *Features* *Harmonic Mean (HM), Arithmetic Mean (AM), Common neighbors (CN), Jaccard Index (JI), Adamic Adar Index (AA), Resource Allocation (RA), Hub Promoted Index (HPI), Number of Nodes in a hyperedge(NN)*

ensemble learning

: We use following classifiers and conducted ensemble learning for hyperedge prediction

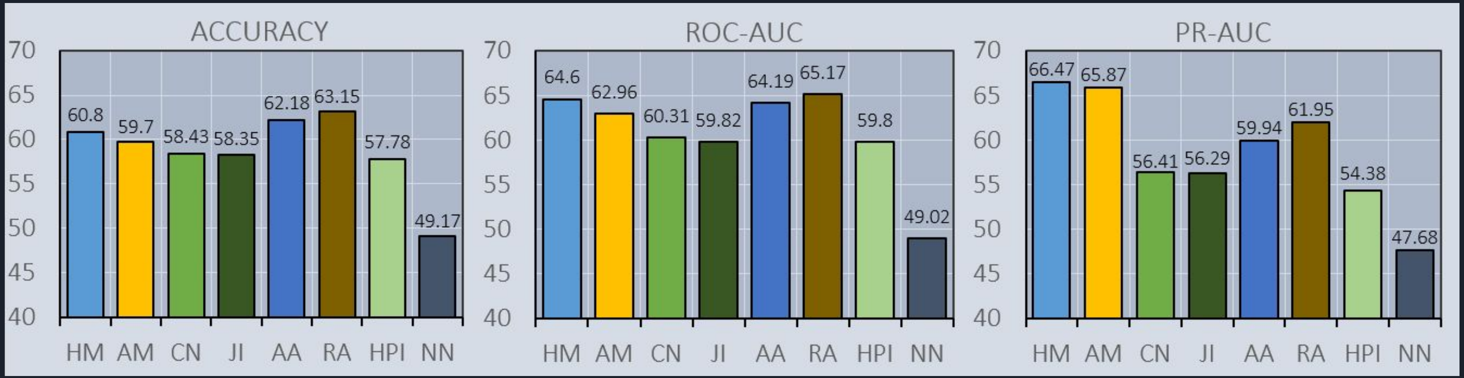
>> Classifiers *Random Forest(RF), LightGBM(LGBM)*

[1] Yoon, Se-eun, et al. "How Much and When Do We Need Higher-order Informationin Hypergraphs? A Case Study on Hyperedge Prediction." Proceedings of The Web Conference 2020. 2020.

3. Performance evaluation

We construct a graph using 58646 nodes and 137958 hyperedges. For hyperedge prediction task, we use 34479 candidate hyperedges 27583 for train, and 6896 for test.

Effectiveness of graph structure based features
: To show the performance of each feature, we ran 8 different hyperedge prediction models with each feature. The results are displayed in following chart.



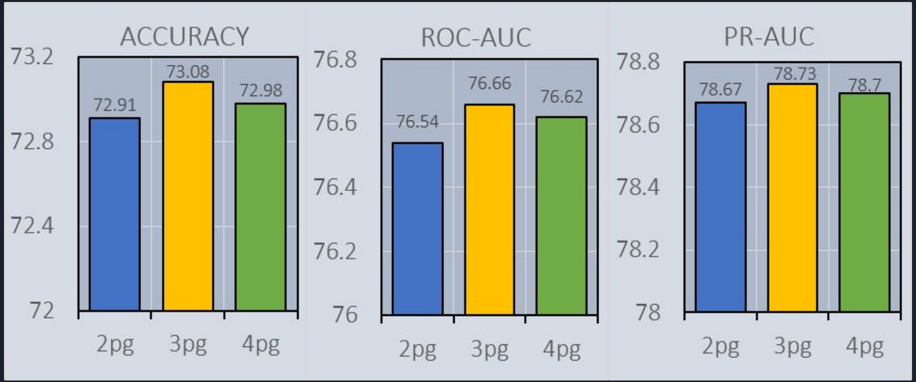
>> ***Features*** *Harmonic Mean (HM), Arithmetic Mean (AM), Common neighbors (CN), Jaccard Index (JI), Adamic Adar Index (AA), Resource Allocation (RA), Hub Promoted Index (HPI), Number of Nodes in a hyperedge(NN),*

3. Performance evaluation

We construct a graph using 58646 nodes and 137958 hyperedges. For hyperedge prediction task, we use 34479 candidate hyperedges 27583 for train, and 6896 for test.

Effectiveness of n-order expansion

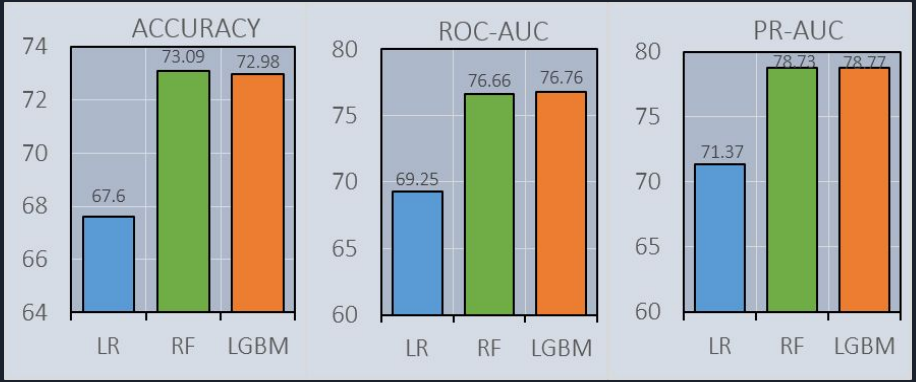
: To show the performance of n-order expansion, we trained models by gradually concatenating features from n-projected Graph. (i.e 3pg has total 16 features extracted from 2-projected Graph and 3-projected Graph) We extended the graph to 4-order expansion to represent the high-order interaction in hyperedge.



3. Performance evaluation

We construct a graph using 58646 nodes and 137958 hyperedges. For hyperedge prediction task, we use 34479 candidate hyperedges 27583 for train, and 6896 for test.

Effectiveness of ensemble learning
: To show the performance of the ensemble learning, we ran 3 models(Logistic Regression, Random Forest, LightGBM). We use 16 features extracted from 2-projected Graph and 3-projected Graph. We compare ensemble learning method(Random Forest and LightGBM) with normal linear model, Logistic Regression.



3. Performance evaluation

Effectiveness of graph structure based features. *Every features have great performance than Random Guessing except NN. But NN complements other features and shows good results.*

Effectiveness of n-order expansion. *N-order expansion gives better accuracy, ROC-AUC, PR-AUC than using 2-projected graph. Especially, model with graph extended to 3-order shows the best performance.*

Effectiveness of ensemble learning. *Ensemble learning method achieves high performance than simple linear method. Especially, Random Forest gives the best performance.*

FINAL MODEL. ***Random Forest trained with 16 features extracted from 2-pg Graph and 3-pg Graph.***

4. Conclusion

- To overcome two challenges(scalability issue, low level representation), we abstracted the graph through n-order expansion and conducted *Ensemble Learning for Hyperedge Prediction using Graph Structure based on features in Hypergraph*.

- In experiment using co-authorship network data, our method showed 73.09% accuracy, 76.66 ROC-AUC and 78.73 PR-AUC score for co-author relationship predictions.

- We showed the effectiveness of our features, n-order expansion and ensemble learning

THANK YOU!

AI506 Term Project

20205007 Junhyeok Kang
20193154 ChangHwan Kim
20203221 Hyangsuk Min