OneAHN: One-step Active Learning for Heterogeneous Network

Group Members:
Hang Zhang (605523350), Xinyu Zhao(205638923), Haowei Jiang(505728615), Nuocheng Pan(905726493)

I. Problem & Goal

Heterogeneous graphs which are composed of various types of nodes and relations contain abundant information and are more common in real applications. However, it's difficult and expensive to achieve labeled heterogeneous graph data. Active learning is an efficient approach to handle the scenario in which we are given an unlabeled dataset and can only label a few data points under the budget. A well-designed active learning algorithm would enable us to select the most informative nodes by query function, and get their labels with the oracle.

Currently, we want to solve the node classification task of heterogeneous networks in an active learning approach. ActiveHNE[1] firstly explored this research with a framework similar to AGE[2] and ANRMAB[6]. In our research, we aim to improve the performance and reduce the expenses of HNE on node classification tasks by proposing a graph partition-based idea.[3] Except for the advantage that the graph partition-based method can find out the most representative samples in different graph regions, it can also help us query all nodes at once and save the does not need to retrain the network comparing to batch training method.

II. Data Plan

To explore this problem under limited computation resources, we decide to first examine our proposed approach and the baselines on small-size benchmark datasets. Some datasets we are considering are AIFB, MUTAG, and BGS. AIFB is a Semantic Web dataset and its graphs stand for the hierarchical structure of the AIFB research institute. Each vertex represents a staff, a research group, or a publication. Each edge represents the relationship of affiliation or belonging. The MUTAG dataset is a group of nitroaromatic compounds. Graphs in the MUTAG dataset stand for chemical compounds where each vertex is an atom and each edge is a bond. BGS is a dataset of rock types published by the British Geology Survey. Vertices are different types of rocks and edges are the relationships of lithogenesis. Information about the three datasets mentioned above is given in the following table. After downloading the dataset from the websites below. We may do some filtering work, such as sampling the data and removing relations that are used to create entity labels.

Dataset	Entities	Edges	Edge Type	Classes	Target Category	Source
AIFB	8,285	29,043	45	4	Personen	https://figshare.com/articles/AIFB_DataSet/745364
MUTAG	23,644	74,227	23	2	d	https://ls11-www.cs.tu-dortmund.de/staff/morris/graphkerneldatasets
BGS	94,806	672,884	103	2	NamedRockUnit	https://docs.dgl.ai/generated/dgl.data.BGSDataset.html

III. Solution Plan

GraphPart[3] is the state-of-the-art active learning solution to homogeneous networks in one-step setting, which proposes a graph-partition-based active learning framework following the setting that divides graphs into disjoint partitions and assumptions on the label and model smoothness. Inspired by GraphPart, we proposed a similar method OneAHN for heterogeneous networks which also follows its setting and assumptions.

First, we apply a modularity-based graph partition method to the graph to obtain a *K*-partition of the graph, specifically the Clasuset-Newman-Moore greedy modularity maximization[5] method. Then, given the partition result and a total budget *b*, we equally distribute budgets to each partition, apply K-Medoids algorithm to each partition based on the aggregated feature embeddings and select the clustering centers as the data points to be queried. Besides, to avoid problem that the clustering centers between two partitions are closed to each other, we also proposed OneAHNFar which penalize each node for the distance to the nearest selected node, which is similar to the idea of GraphPartFar, a variant of GraphPart.

For underlying GNN models, we plan to perform experiments on R-GCN[4] and DHNE part in ActiveHNE.

Our proposed method should be the first attempt to apply one-step batch-mode active learning on heterogeneous network settings. Compared with ActiveHNE, our proposed method only queries once thus minimizing the cost of training; compared with GraphPart, our proposed method can solve node classification problems on heterogeneous graphs and thus can be applied to a wider range of real world applications.

IV. Evaluation Plan

We will experiment on the 3 benchmark datasets mentioned above by comparing our method with three baseline active learning approaches on training two GNN models (R-GCN and DHNE). In the evaluation, we analyze the following baselines using a sequence of label budgets and record the **accuracy & total time** for node classification on the whole graph. For our proposed one-step active learning setting, the seed set is established as zero at start.

Baseline 1: R-GCN/DHNE; Nodes are randomly selected during each query in the case.

Baseline 2: AGE; An Active Graph Embedding framework which chooses nodes that scored the highest among all the three informativeness scores combined linearly.

Baseline 3: ActiveHNE; A novel Active Heterogeneous Network Embedding framework that designs three selection metrics and uses a batch selection method that assembles those metrics by a multi-armed bandit mechanism.

V. Schedule & Work Division

Week 5: Further literature review, make sure we survey the related field completely.

Week 6: Read baseline codes, Design OneHNE and training

Week 7: Implementation of OneHNE, training on baselines

Week 8: Continue coding, training on both baseline and OneHNE, starting report draft

Week 9: Finish up code and demo, make slides and practice for presentation

Week 10: Presentation, finishup report

For the work division, we plan to implement OneHNE together. Hang is in charge of supervising and coordinating the whole project. Xinyu is responsible for baseline 1. Haowei is responsible for baseline 2. Nuocheng is responsible for baseline 3.

VI. Reference

[1] Chen, X., Yu, G., Wang, J., Domeniconi, C., Li, Z., & Zhang, X. (2019). Activehne: Active heterogeneous network embedding. *akrXiv preprint arXiv:1905.05659*.

[2] Cai, H., Zheng, V. W., & Chang, K. C. C. (2017). Active learning for graph embedding. *arXiv preprint arXiv:1705.05085*.

[3] Ma, J., Ma, Z., Chai, J., & Mei, Q. (2022). Partition-Based Active Learning for Graph Neural Networks. *arXiv preprint arXiv:2201.09391*.

[4] Schlichtkrull, M., Kipf, T. N., Bloem, P., Berg, R. V. D., Titov, I., & Welling, M. (2018, June). Modeling relational data with graph convolutional networks. *European semantic web conference (pp. 593-607)*. Springer, Cham.

[5] Clauset, A., Newman, M. E., & Moore, C. (2004). Finding community structure in very large networks. *Physical review E*, *70*(6), 066111.

[6] Gao, Li & Yang, Hong & Zhou, Chuan & Wu, Jia & Pan, Shirui. (2018). Active Discriminative Network Representation Learning. 2142-2148. 10.24963/ijcai.2018/296.