Efficient Subsampling for GNN Downstream Tasks

Hirad Daneshvar HIRAD.DANESHVAR@TORONTOMU.CA
Toronto Metropolitan University, Toronto, ON, Canada

Reza Samavi SAMAVI@TORONTOMU.CA
Toronto Metropolitan University, Toronto, ON, Canada
Vector Institute, Toronto, ON, Canada

Editors: Hung-yi Lee and Tongliang Liu

Appendix A. Clinical Dataset

We utilized two linked datasets: a medical records dataset and a mental health outpatient questionnaire dataset. Table 1 shows available information in the medical record's dataset. The dataset includes several categorical features, which are **triage level**, describing the type and severity of a patient's initial symptoms, **visit disposition code**, which indicates how a patient was discharged from ambulatory care after registration, **service utilization,** which refers to the specific health professional services accessed during the patient's visit, and **most responsible diagnosis code** that identifies the primary, most clinically significant problem determined by the healthcare provider during service utilization.

The patient graph consists of three feature categories: visit (including age, triage, and disposition), service utilization (based on service code), and diagnosis (using diagnosis code). Every visit includes at least one diagnosis, and each diagnosis is associated with at least one service. Connections extracted for each visit are: *visit-visit* (chronologically ordered), *visit-diagnosis*, and *diagnosis-service*. Figure 1 shows a sample EHR graph.

Table 1: Medical dataset information.

Attribute	Type	Insight	Example
Patient age at the time of visit	Numeric	Min: 4 Max: 17 Mean: 12.5 Std: 3.9	12
ED admission date	Date	Min: 2006-10-16 Max: 2021-07-31	-
Triage level	Categorical	6 Categories	Emergent
Visit disposition code	Categorical	16 Categories	Intra-facility transfer to the ED
Utilized service code	Categorical	35 Categories	Orthopedic Surgery
Diagnosis code	Categorical	707 Categories	Allergic Purpura

Table 2 shows the subset of the questions used in the mental health questionnaires dataset. The questions were selected through an ablation study designed to identify the most influential questions for the downstream task. The question categories have been used to create nodes and connections between them. We utilized the timeline of events in the questionnaire to augment the EHR patient graph with the questionnaire responses, as shown in Figure 2.

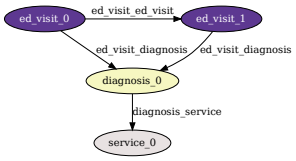

Figure 1: A sample patient graph using EHR data.

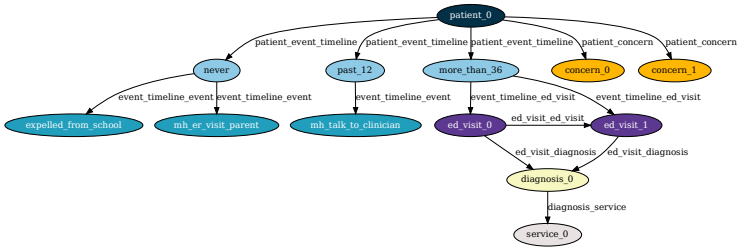

Figure 2: A sample augmented patient graph using EHR and questionnaire data.

Table 2: The subset of selected questionnaires, along with their category.

Question Category	Question
Patient Specific	Are you taking any medication or pills prescribed for mental health concerns?
	What sex were you assigned at birth, on your original birth certificate?
	How do you describe yourself (Male, Female, Transgender, Do not identify as female, male or transgender, Don't know)?
	Are you currently living in a friend's house or apartment?
	Do you think of yourself as (Straight, Gay or Lesbian, Bisexual, Transgender, transsexual or gender non-conforming, Don't know)?
	Were you born in Canada?
Events	During the past 12 months, did you visit an emergency room about concerns regarding your mental health?
	During the past 12 months, have you been suspended or expelled from school?
	During the past 12 months, did you see or talk to a doctor, psychologist, psychiatrist, or counsellor about concerns regarding your mental health?
Concerns	I am concerned about my physical appearance
	I am concerned about physically hurting myself

Appendix B. Additional Experiments with Enzymes Dataset

To assess the generalizability of the proposed approach, we conducted experiments using the Enzymes dataset (Morris et al., 2020). As this evaluation involves a single dataset, where the records are not linked to any other datasets, we compared the performance of a GNN classifier trained on a randomly split dataset with that of the same classifier trained on a subset selected by the proposed framework.

B.1. Experimental Setup

Dataset: We employed the Enzymes dataset, collected by TU Dortmund University Morris et al. (2020), which consists of graphs representing macromolecules. As each graph is

Table 3: Results of performance improvement using the proposed framework on the Enzymes dataset.

Method	Accuracy	ROC AUC	Precision	Recall	F1 Score
Random Splitting	0.47 ± 0.04	**0.8 ± 0.03**	0.51 ± 0.05	0.47 ± 0.04	0.46 ± 0.04
Proposed Framework	**0.53**	0.79	**0.55**	**0.53**	**0.53**

associated with a label, the task is formulated as graph-level classification. The dataset comprises 600 graphs, each containing on average 32.6 nodes and 124.3 edges. Each node is characterized by 18 features, and the dataset includes six distinct classes.

Networks: For the classification task, we trained a three-layer GNN following the architecture outlined in Dwivedi et al. (2023). The model begins with a linear transformation and dropout to generate initial node embeddings. These embeddings, together with the adjacency matrix of each graph, are processed through a three-layer GraphSAGE operator (Hamilton et al., 2017) with ReLU activations. To obtain a graph-level representation, we applied max pooling across node embeddings. The resulting representation is then fed into a two-layer multilayer perceptron (MLP) for classification. The configuration of the GraphSAGE operator was aligned with the recommendations in Dwivedi et al. (2023).

For the self-distillation approach used to estimate the model's prediction uncertainty, we employed a similar three-layer GNN. In this case, after each graph layer, the pooled representation is passed through a two-layer MLP to produce intermediate classifications. The final classifier, i.e., the deepest layer, is used as the reference model for predictions.

Training and Hardware Specifications: All models have been implemented using Python version 3.9 and PyTorch version 1.13.1. The models have been trained on a GPU (NVIDIA GeForce RTX 3050) with CUDA version 12.5. To optimize the networks, we used the Adam Optimizer. We only used 90% of the data during training of the network using both approaches. The remaining 10% of the data is used to assess network performance.

B.2. Results

For comparison, we first trained the GNN using 5-fold cross-validation, with 80% of the data allocated for training; this setting is referred to as random splitting. Next, we applied the self-distillation approach to estimate prediction uncertainty and employed the proposed framework with $k = 10$ for clustering, thereby subsampling 80% of the data with an emphasis on retaining more uncertain samples. Finally, the subsampled data were used to train the original three-layer GNN. Table 3 presents the results of the experiments. As shown in the table, subsampling using the proposed framework has improved the results across nearly all metrics.

Finding: The proposed subsampling framework facilitates the identification of samples that are more informative for training the network.

References

Vijay Prakash Dwivedi, Chaitanya K. Joshi, Anh Tuan Luu, Thomas Laurent, Yoshua Bengio, and Xavier Bresson. Benchmarking graph neural networks. *Journal of Machine Learning Research*, 24(43):1–48, 2023. URL http://jmlr.org/papers/v24/22-0567.html.

Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/5dd9db5e033da9c6fb5ba83c7a7ebea9-Paper.pdf.

Christopher Morris, Nils M. Kriege, Franka Bause, Kristian Kersting, Petra Mutzel, and Marion Neumann. Tudataset: A collection of benchmark datasets for learning with graphs. In *ICML 2020 Workshop on Graph Representation Learning and Beyond (GRL+ 2020)*, 2020. URL www.graphlearning.io.