Graph Classification via Reference Distribution Learning: Theory and Practice

Zixiao Wang Jicong Fan

The Chinese University of Hong Kong, Shenzhen, China

Contribution of This Work

- Propose a novel graph classification method GRDL that is both efficient and accurate.
- Provide theoretical guarantees, e.g. generalization error bounds, for GRDL.

Motivation

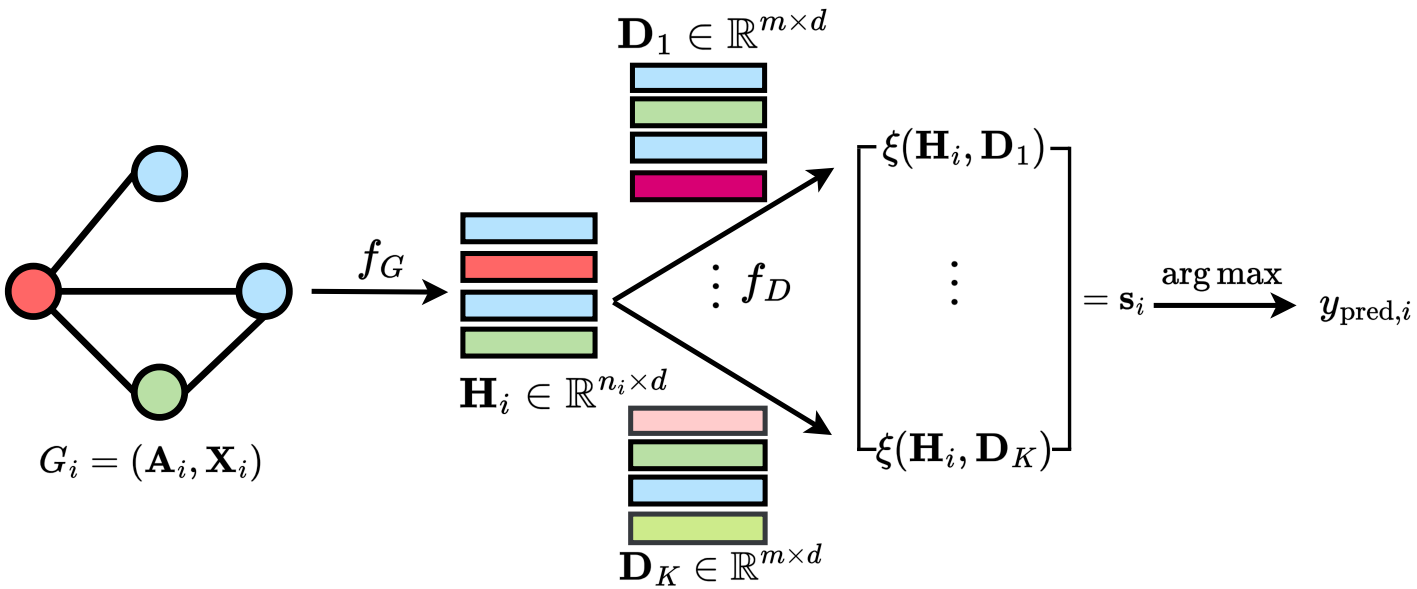
- Most of the global pooling methods are naive, often employing methods such as simple summation or averaging. These pooling methods collect only the first-order (statistics) information, leading to a loss of structural or semantic information.
- More sophisticated pooling operations retain more meaningful information, but still carry the inherent risk of information loss.

Proposed Model

GRDL is composed of two parts:

- f_G is a backbone GNN to transform each graph $G_i = (\mathbf{A}_i, \mathbf{X}_i)$ to a node embedding matrix $\mathbf{H}_i \in \mathbb{R}^{n_i \times d}$

$$\mathbf{H}_i = f_G(G_i) = f_G(\mathbf{A}_i, \mathbf{X}_i),$$



Proposed Model

GRDL is composed of two parts:

- f_G is a backbone GNN to transform each graph $G_i = (\mathbf{A}_i, \mathbf{X}_i)$ to a node embedding matrix $\mathbf{H}_i \in \mathbb{R}^{n_i \times d}$
- A reference layer f_D computes the similarity between each graph embedding $\mathbf{H}_i$ and reference distributions $\{\mathbf{D}_1, \ldots, \mathbf{D}_K\}$

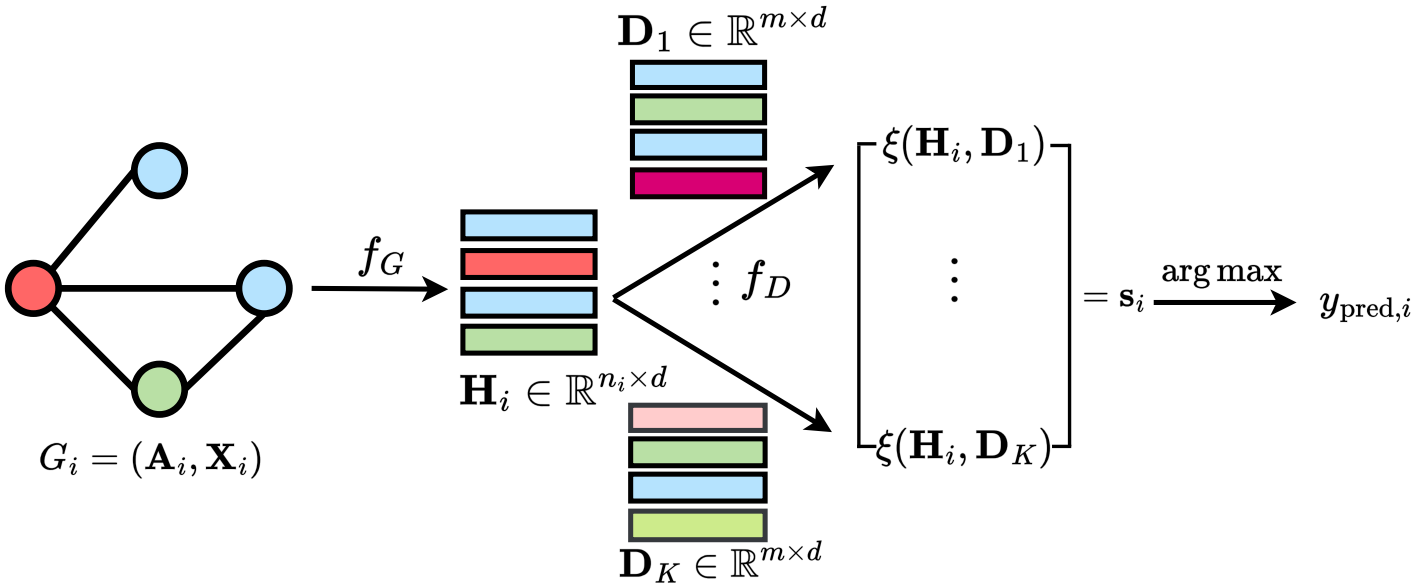
$$f_D(\mathbf{H}_i) = [s_{i1}, s_{i2}, \ldots, s_{iK}] = [\xi(\mathbf{H}_i, \mathbf{D}_1), \xi(\mathbf{H}_i, \mathbf{D}_2), \ldots, \xi(\mathbf{H}_i, \mathbf{D}_K)]^\top.$$

$\xi(\cdot, \cdot)$ is a similarity measure between two distributions, and is chosen to be the negative squared maximum mean discrepancy:

$$\xi(\mathbf{H}, \mathbf{D}) = \frac{2}{mn} \sum_{i=1}^{n} \sum_{j=1}^{m} k(\mathbf{h}_i, \mathbf{d}_j) - \frac{1}{n^2} \sum_{i=1}^{n} \sum_{i'=1}^{n} k(\mathbf{h}_i, \mathbf{h}_{i'}) - \frac{1}{m^2} \sum_{j=1}^{m} \sum_{j'=1}^{m} k(\mathbf{d}_j, \mathbf{d}_{j'}).$$

$k(\cdot, \cdot)$ is chosen to be the Gaussian kernel.

Optimization



$$\min_{f_G, f_D} -\frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} y_{ik} \log \frac{\exp(s_{ik})}{\sum_{j=1}^{K} \exp(s_{ij})} + \lambda \sum_{k} \sum_{k' \neq k} \xi(\mathbf{D}_k, \mathbf{D}_{k'})$$

Experiment Results on Graph Datasets

Method	Dataset								Average
	MUTAG	PROTEINS	NCI1	IMDB-B	IMDB-M	PTC-MR	BZR	COLLAB	
PATCHY-SAN	**92.6±4.2**	75.1±3.3	76.9±2.3	62.9±3.9	45.9±2.5	60.0±4.8	85.6±3.7	73.1±2.7	71.5
GIN	89.4±5.6	76.2±2.8	82.2±0.8	64.3±3.1	50.9±1.7	64.6±7.0	82.6±3.5	79.3±1.7	73.6
DropGIN	90.4±7.0	76.9±4.3	81.9±2.5	66.3±4.5	51.6±3.2	66.3±8.6	77.8±2.6	80.1±2.8	73.9
DIFFPOOL	89.4±4.6	76.2±1.4	80.9±0.7	61.1±3.0	45.8±1.4	60.0±5.2	79.8±3.6	80.8±1.6	71.8
SEP	89.4±6.1	76.4±0.4	78.4±0.6	**74.1±0.6**	51.5±0.7	68.5±5.2	86.9±0.8	**81.3±0.2**	75.8
GMT	89.9±4.2	75.1±0.6	79.9±0.4	**73.5±0.8**	50.7±0.8	**70.2**±6.2	85.6±0.8	80.7±0.5	75.7
MinCutPool	90.6±4.6	74.7±0.5	74.3±0.9	72.7±0.8	51.0±0.7	68.3±4.4	87.2±1.0	**80.9±0.3**	75.0
ASAP	87.4±5.7	73.9±0.6	71.5±0.4	72.8±0.5	50.8±0.8	64.6±6.8	85.3±1.3	78.6±0.5	73.1
WitTopoPool	89.4±5.4	80.0±3.2	79.9±1.3	72.6±1.8	**52.9±0.8**	64.6±6.8	87.8±2.4	80.1±1.6	75.9
OT-GNN	91.6±4.6	76.6±4.0	**82.9±2.1**	67.5±3.5	52.1±3.0	68.0±7.5	85.9±3.3	80.7±2.9	75.7
WEGL	91.0±3.4	73.7±1.9	75.5±1.4	66.4±2.1	50.3±1.0	66.2±6.9	84.4±4.6	79.6±0.5	73.4
FGW - ADJ	82.6±7.2	72.4±4.7	74.4±2.1	70.8±3.6	48.9±3.9	55.3±8.0	86.9±1.0	80.6±1.5	71.5
FGW - SP	84.4±7.3	74.3±3.3	72.8±1.5	65.0±4.7	47.8±3.8	55.5±7.0	86.9±1.0	77.8±2.4	70.6
WL	87.4±5.4	74.4±2.6	**85.6±1.2**	67.5±4.0	48.4±4.2	56.0±3.9	81.3±0.6	78.5±1.7	72.4
WWL	86.3±7.9	73.1±1.4	**85.7±0.8**	71.6±3.8	52.6±3.0	52.6±6.8	87.6±0.6	**81.4±2.1**	73.9
SAT	**92.6±4.3**	77.7±3.2	82.5±0.8	70.0±1.3	47.3±3.2	68.3±4.9	**91.7±2.1**	80.6±0.6	76.1
Graphormer	89.6±6.2	76.3±2.7	78.6±2.1	70.3±0.9	48.9±2.0	**71.4±5.2**	85.3±2.3	80.3±1.3	75.1
GRDL	**92.1±5.9**	**82.6±1.2**	80.4±0.8	**74.8±2.0**	**52.9±1.8**	68.3±5.4	**92.0±1.1**	79.8±0.9	**77.9**
GRDL-W	90.8±4.6	**82.1±0.9**	80.9±0.8	72.2±3.1	**53.1±0.9**	68.5±3.2	90.6±1.5	80.4±1.1	**77.3**
GRDL-S	90.6±5.7	**81.1±1.4**	81.2±1.5	72.4±3.3	52.5±1.1	64.2±3.2	**91.6±1.3**	78.6±1.3	76.5

Table: Classification accuracy (%). Bold text indicates the top 3 mean accuracy.

AUC-ROC Scores of Large Imbalanced Data Classification

Table: AUC-ROC scores of large imbalanced data classification. Bold text indicates the best.

METHOD	DATASET		
	PC-3	MCF-7	OGBG-MOLHIV
GIN	84.6±1.4	80.6±1.5	77.8±1.3
DIFFPOOL	83.2±1.9	77.2±1.3	73.7±1.8
PATCHY-SAN	80.7±2.1	78.9±3.1	70.2±2.1
GRDL	**85.1±1.6**	**81.4±1.3**	**79.8±1.0**

Visualization of Node Embedding & Reference Distributions

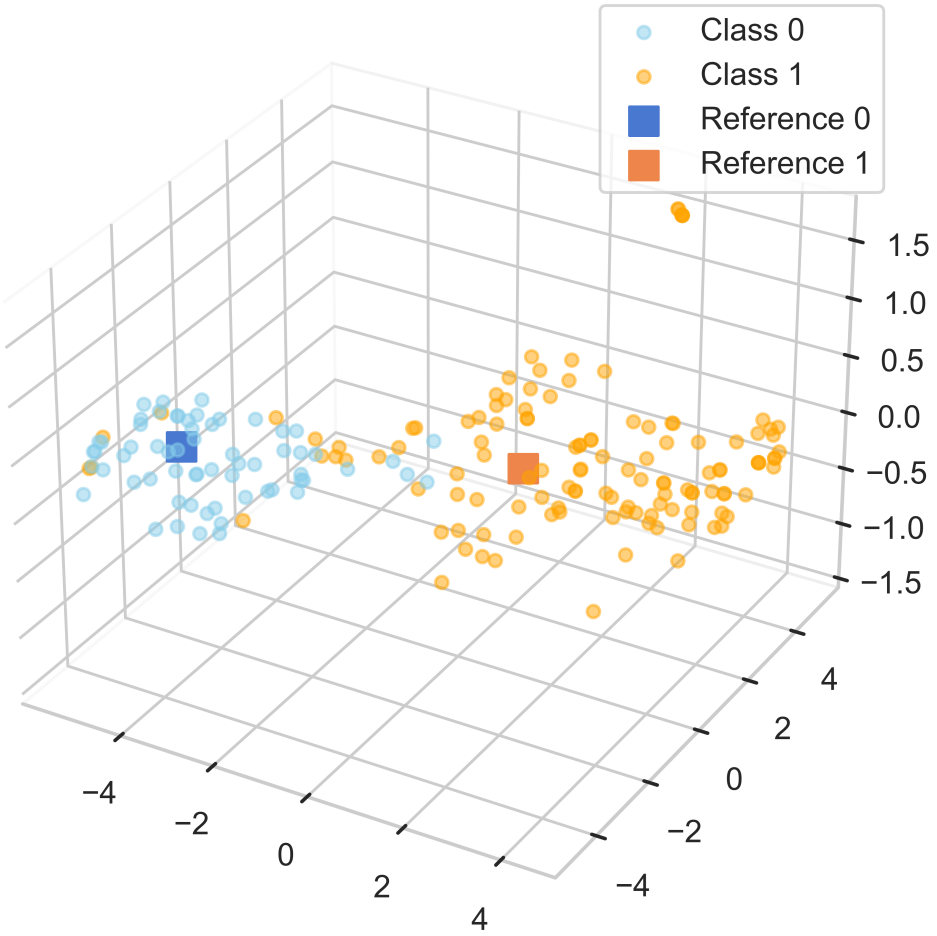


Figure: T-SNE visualization of MUTAG embeddings and reference distributions given by GRDL.

The End

More numerical results as well as the generalization bounds can be found in our paper.

Thanks for your attention!