End-to-End Open-Set Semi-Supervised Node Classification with Out-of-Distribution Detection

Tiancheng Huang, Donglin Wang, Yuan Fang, and Zhengyu Chen

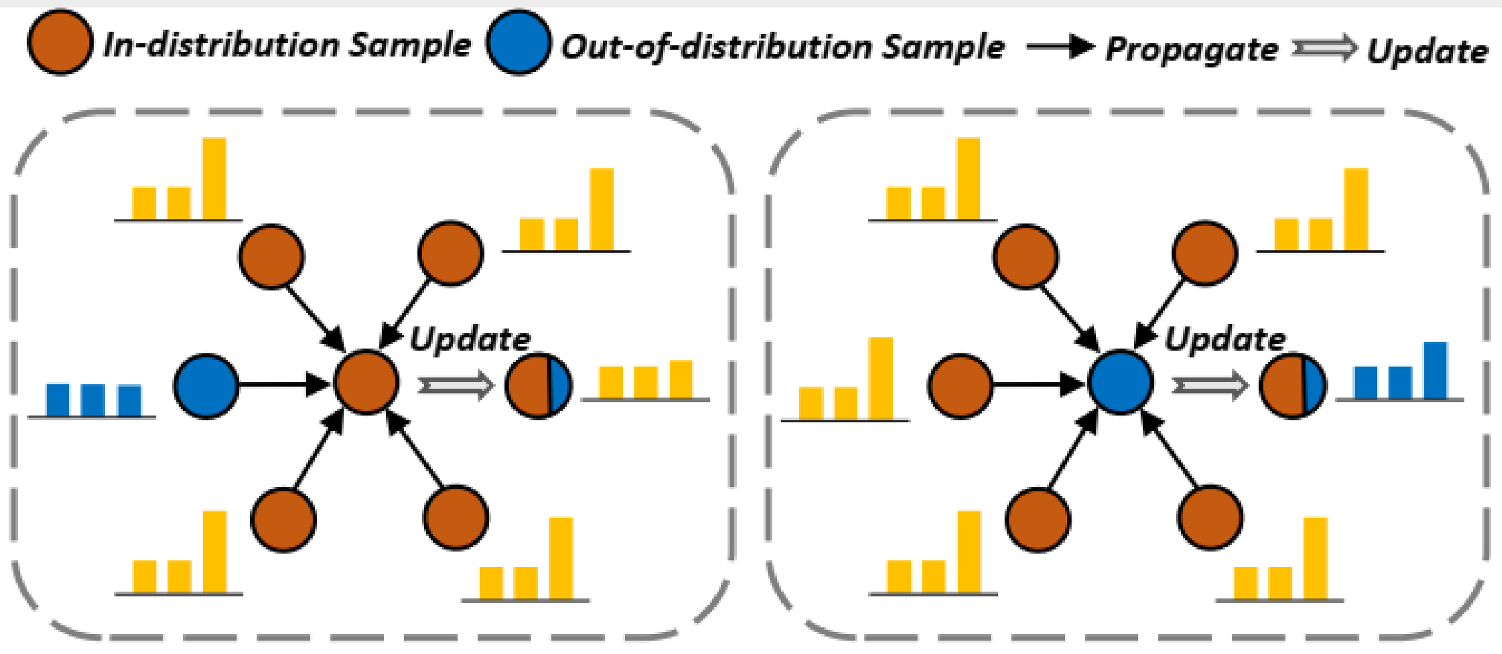
ZHEJIANG UNIVERSITY · WESTLAKE UNIVERSITY · SMU SINGAPORE MANAGEMENT UNIVERSITY

IJCAI ECAI VIENNA 22

A Crucial Question

How to train the GNN to perform open-set semi-supervised node classification with OOD detection in end-to-end manner ?

Across-distribution Mixture on Graphs



(a) Mixing to in-distribution (b) Mixing to OOD

Challenges

1) Lack of supervision on in-distribution or OOD; 2) Promote the information propagation but cause across-distribution mixture; 3) Over-fitting could be more severe (an additional model)

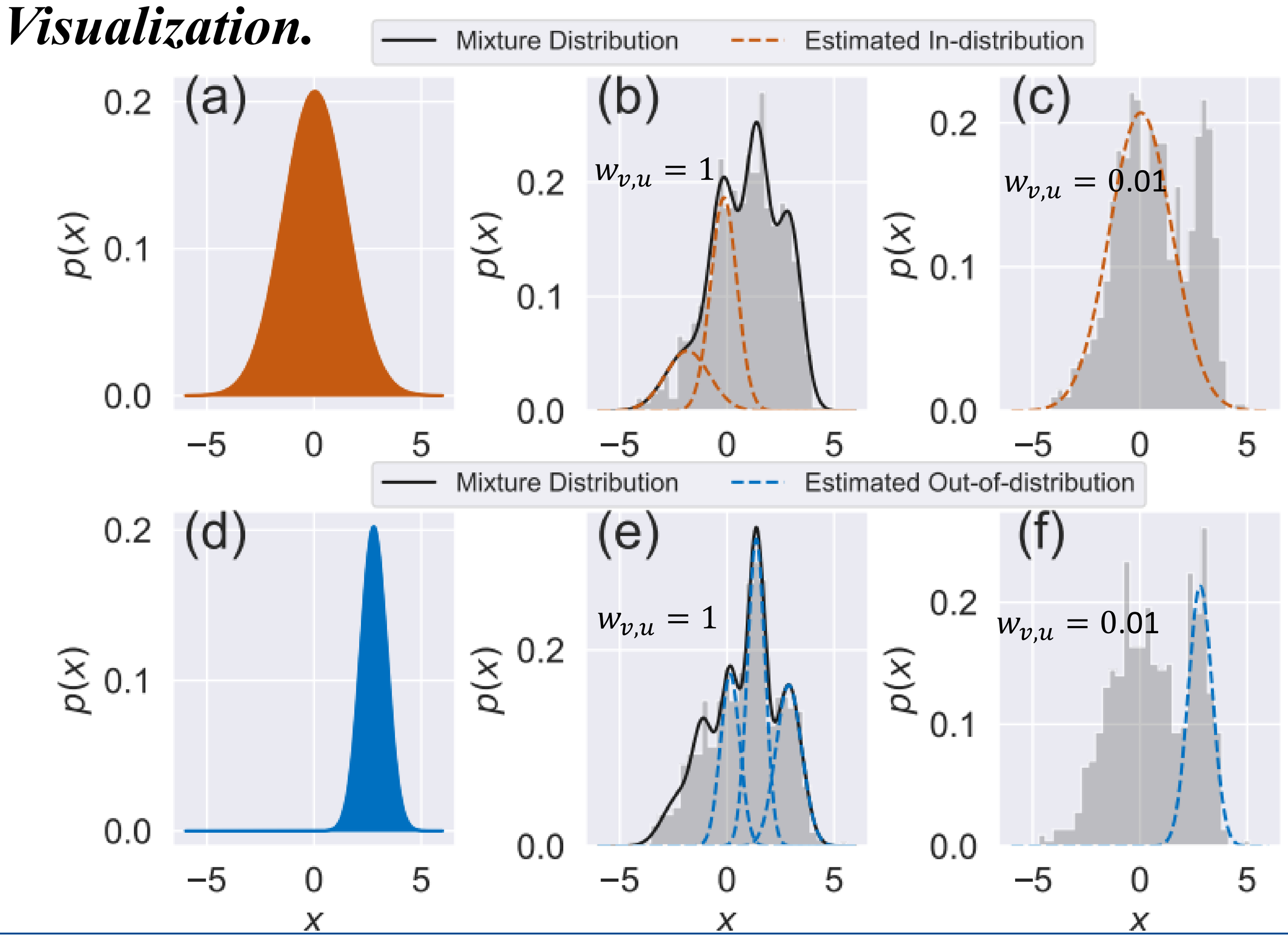
Analysis and Toy Visualization

Across-distribution Mixture on Graphs. Take one-layer aggregation of GNN as an example, the distribution mixture comes from the central node and its neighbors:

$$P_{mix}(\boldsymbol{x}_i) \sim \mathcal{N}(\mu_i, \sigma_i) + \sum_{v=1}^{|\mathbb{N}(u)|} w_{v,u}\mathcal{N}(\mu_{j_v}, \sigma_{j_v})$$

where $i, j_v \in \{1,2\}$, $w_{v,u}$ denotes weights between node u and v, and $\mathbb{N}(u)$ denotes neighbors of node u.

Visualization.



Methodology

Unified Learning Framework

The joint probability distribution of labels Y and O

$$P_\theta(Y, O|\mathbf{X}, \mathbf{A}) = P_\theta(Y|\mathbf{X}, \mathbf{A}, O)P(O|\mathbf{X}, \mathbf{A}),$$

where θ represents GNN parameter and the probability distribution of OOD detection can be described as $P\left(O|\mathbf{X}, \mathbf{A}\right)$. From the Bayesian perspective, the learning process includes:

i) Learning the GNN parameter by maximizing the likelihood

$$\log P_\theta(Y, O|\mathbf{X}, \mathbf{A}) = \log \sum_k P_\theta(Y|\mathbf{X}, \mathbf{A}, O_k) \cdot P(O_k|\mathbf{X}, \mathbf{A}),$$

ii) Inferring the following posterior of the latent variable O

$$P_\theta(O_k|\mathbf{X}, \mathbf{A}, Y) = \frac{P_\theta(Y|\mathbf{X}, \mathbf{A}, O_k)}{\sum_j P_\theta(Y|\mathbf{X}, \mathbf{A}, O_j)},$$

Variational Inference. We introduce a variational distribution to minimize Kullback-Leibler (KL) divergence:

$$\mathrm{KL}(Q_\phi \| P) = \mathbb{E}_Q\left[\log \frac{Q_\phi(O_k|\mathbf{X}, \mathbf{A})}{P(O_k)}\right].$$

Minimizing the following negative ELBO:

$$\mathcal{L}(\theta, \phi) = -\mathbb{E}_{Q_\phi}[\log P_\theta(Y|\mathbf{X}, \mathbf{A}, O)] + \mathrm{KL}(Q_\phi \| P),$$

Learning to Mix Neighbors (LMN)

Adopting a predictor to generates the weights to mix neighbors

$$Q_\phi(w_{v,u}|\mathbf{X}, \mathbf{A}) = \frac{1}{1 + exp(-\mathbf{W}^T[\mathbf{H}_v \| \mathbf{H}_u])},$$

The graph convolution based on message passing

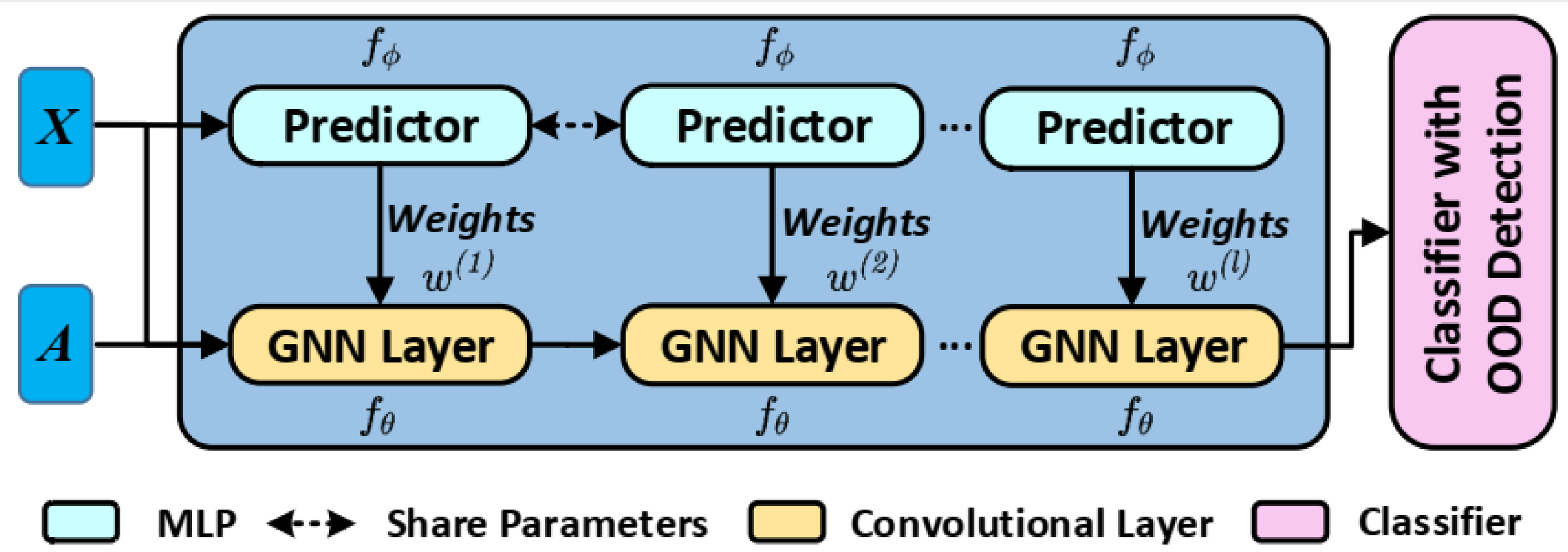
$$\mathbf{H}^{(l+1)} = f_{\theta^{(l)}}(\mathbf{H}^{(l)}, \mathbf{A}, w_{v,u}^{(l)}),$$

Bi-level Optimization

Updating Outer Loop $\min_\phi \mathcal{L}_{val}(\theta^*(\phi), \phi),$

Updating Inner Loop $s.t.\ \theta^*(\phi) = \arg\min_\theta \mathcal{L}_{train}(\theta, \phi),$

Framework of LMN



Experiments

Node Classification Task (Accuracy%)

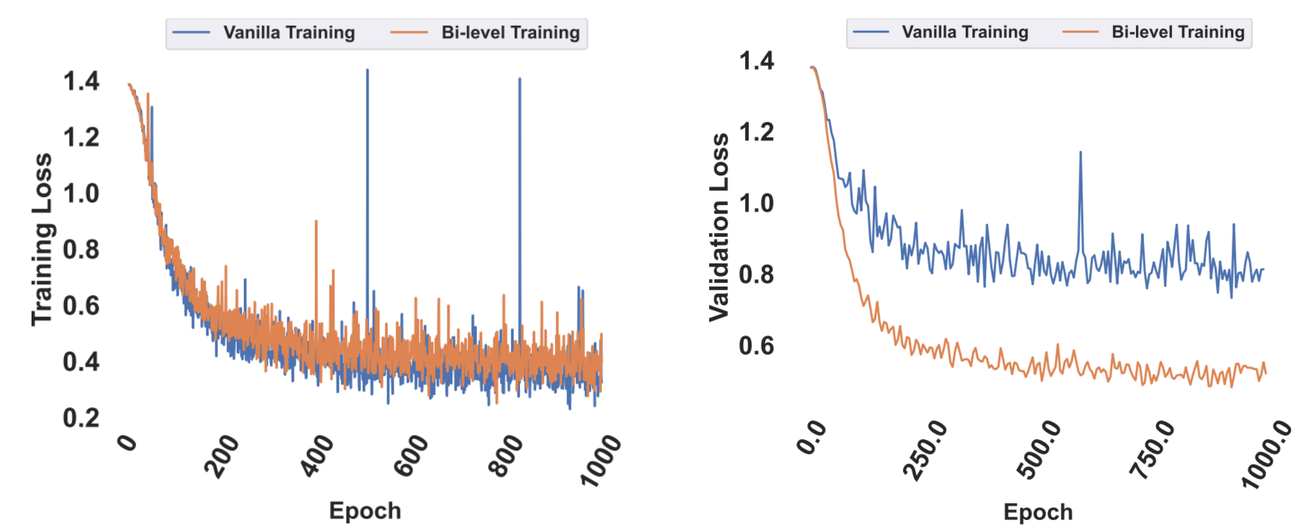
Methods	Cora	Citeseer	Pubmed	arXiv
GCN	87.4±0.3	66.0±0.6	89.0±0.2	47.4±0.6
ChebNet	85.6±0.4	65.0±0.6	88.4±0.3	46.5±0.4
GraphSAGE	85.3±1.2	65.8±0.7	89.6±0.6	46.8±0.9
GAT	88.7±0.6	69.6±0.6	90.6±0.9	49.8±1.5
SGC	87.2±0.3	69.2±0.2	91.5±0.6	40.5±2.6
JKNet	86.7±1.1	67.3±0.7	93.1±0.1	50.6±0.6
APPNP	88.2±0.4	68.3±0.5	93.2±0.1	51.3±0.9
SuperGAT	88.3±0.5	69.3±0.8	91.3±1.0	49.2±0.7
GCNII	88.7±0.3	69.4±1.4	93.0±0.7	51.6±1.7
DropEdge	88.9±0.7	69.6±1.2	93.0±0.9	51.7±2.7
LMN(Ours)	**89.7**±0.6	**71.1**±0.6	**93.4**±0.1	**54.1**±1.4

OOD Detection Task (AUROC%)

Methods	Cora	Citeseer	Pubmed	arXiv
GCN	77.8±0.4	73.1±2.2	63.3±1.4	56.1±0.5
ChebNet	73.5±1.3	69.7±4.0	62.2±1.2	57.1±0.8
GraphSAGE	75.6±1.8	72.8±3.1	59.5±2.0	56.9±1.0
GAT	80.2±1.4	77.9±3.1	61.6±4.2	58.0±1.0
SGC	70.0±0.8	75.5±2.3	61.4±1.8	51.8±1.5
JKNet	76.3±1.8	70.8±3.4	64.4±1.8	52.9±0.6
APPNP	77.8±0.5	72.3±2.7	64.3±0.8	53.7±0.3
SuperGAT	78.5±1.6	78.1±1.6	63.2±3.9	54.1±0.8
GCNII	78.0±1.3	72.4±2.1	65.2±3.9	56.3±2.2
DropEdge	79.3±0.9	75.2±3.5	63.0±2.1	57.9±0.9
LMN(Ours)	**80.5**±1.2	**78.5**±3.2	**68.7**±1.3	**60.4**±0.3

The Effect of Bi-level Optimization

➢ Vanilla Training vs. Bi-level Training (Ours)



Ablation Study (Accuracy%)

Methods	OOD Modules	Cora	Citeseer
GCNII	None	88.7±0.3	69.4±1.4
LMN	Mixing Neighbors	**89.7**±0.6	**71.1**±0.6