香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

香港科技大学(广州)
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

工业信息与智能研究所
TRIPLE-I INSTITUTE

NEURAL INFORMATION PROCESSING SYSTEMS

37th Conference on Neural Information Processing Systems

Deep Insights into Noisy Pseudo Labeling on Graph Data

- Introduction
- Theoretical analysis
- Experiments
- Conclusion and prospect

Presenter: Botao WANG

Facility: Hong Kong University of Science and Technology
Hong Kong University of Science and Technology (Guangzhou)

Authors: Jia Li, Yang Liu, Jianshun Cheng, Yu Rong, Wenjia Wang, Fugee Tsung

1. Introduction

Pseudo Labeling (PL) is a popular self-supervised learning approaches to tackle the label sparsity problem by iterative self-labeling. However, there is a trade-off between the benefit of PL and the effect of mislabeled samples.

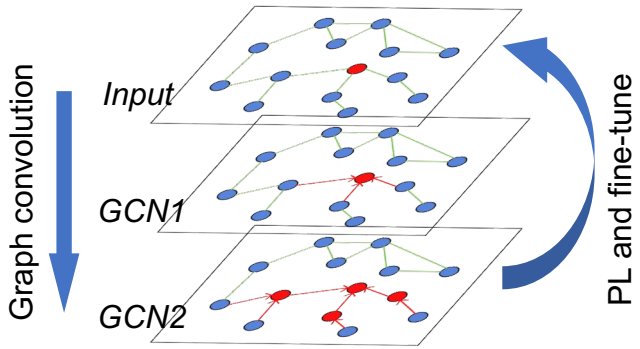

Fig.1 . Noise of PL

In a two-layer graph neural network, false information can influence their 2-hop neighbors, and accumulate during iterations.

Toy experiment (Fig.2) shows that the base model can be **improved, degraded or collapsed** by PL in different situations.

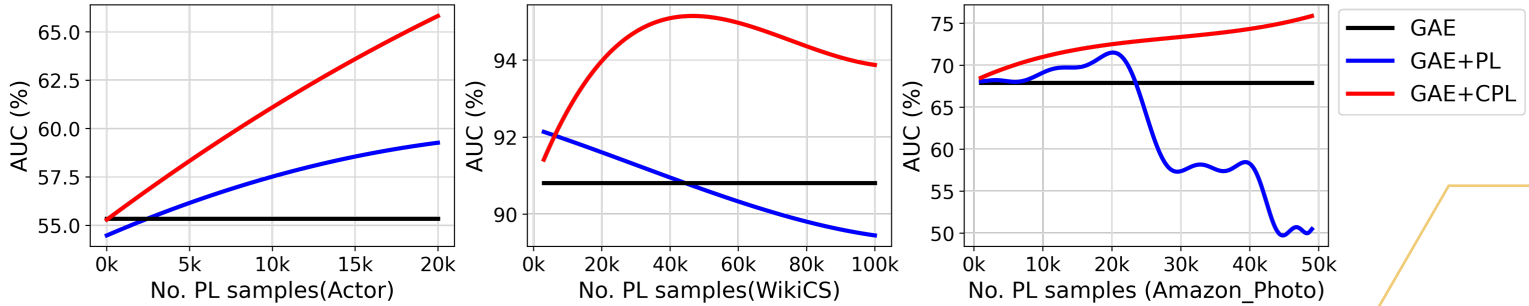

Fig.2 Toy experiment on the comparison of PL strategy in graph learning (link prediction)

Challenges of PL in graph neural networks (GNNs), Fig.1**:**

- For non-i.i.d. data such as graph, the message aggregation would amplify the noises of incorrect labels introduced by PL.
- The PL on the link can affect the inputs of GNN in the following iterations, which implies that the noises can accumulate to damage the base model's performance.

2. Theoretical analysis

香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Assumption 1: Graph invariant property. It guarantees the variation of the output confidence is linearly bounded by the degree of graph perturbation (C-Lipschitz condition).

Given a graph G and its perturbation $\hat{G} = G(X \odot, A \odot M_a)$ by the random feature masks $M_x \in \{1,0\}^{N\times F}$ and adjacent matrix mask $M_a \in \{1,0\}^{N\times N}$ satisfying:

$$\frac{1}{N \cdot F}\|\mathbf{1}^{N\times F} - M_x\|_2^2 + \frac{1}{N^2}\|\mathbf{1}^{N\times N} - M_a\|_2^2 < \epsilon$$

the GNN $g(\cdot)$ has GPI property if there exists a constant $C > 0$ such that the perturbed prediction confidence satisfies $\left\|g(\hat{G}) - g(G)\right\|_2^2 < C\epsilon$.

$\odot$ is element-wise product, $\|\cdot\|_2$ is the 2-norm of the vector or matrix.

Assumption 2: Additive expansion property. It guarantees the continuity of the p_f in the neighborhood of the local optimal subset U.

Define a local optimal subset $U \subset Y$, whose probability is higher than a threshold $p_f(y) > 1 - q, y \in U$, and its perturbation set $U_\varepsilon = \{\hat{y} = g(G): \|y - \hat{y}\|_2 < C\varepsilon, y \in U\}$, where $G \in \{\hat{G}\}$ is the space of the perturbed graph. Then, there exists $\alpha > 0, \eta > 0$, s.t. the probability measure p_f satisfying additive expansion property:

$$p_{\alpha f}(U_{\hat{\varepsilon}} \setminus U) \geq p_{\alpha f}(U) + \eta \cdot \alpha.$$

Correctness of the PL samples is discrete, we can apply multi-view augmentations, reparameterizing the p_f to be continuous.

2. Theoretical analysis

香港科技大學 THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Theorem: Prediction error measurement

Let $q > 0$ be a given threshold. For the GNN in the teacher model g_φ, if its corresponding density measure satisfies additive expansion, the error of the student predictor g_ψ is bounded by:

$$Err(g) = 2\left[q + \mathcal{A}\left(g_\psi\right)\right]$$

where $\mathcal{A}\left(g_\psi\right) = \mathbb{E}_{Y_{test}} \mathbf{1}\left(\exists g\left(\hat{G}\right) \neq g(G)\right)$ measures the inconsistency over differently augmented inputs, Y_{test} is the test set for evaluation.

- If q is small, the PL threshold $1 - q$ approaches 1, leading to a smaller lower bound of error.
- For random PL, confidence threshold $q = 0.5$, then the maximum theoretical error rate is 1.
- A small value of $\mathcal{A}$ indicates consistent prediction across different views. In such cases, we have more confidence in the predictions, leading to a smaller error bound.

Theorem: Convergence analysis

The PL sample selection strategy $\mathcal{T}$ influences the covariance term derived from the empirical loss, then affects the convergence property:

$$L_{\mathcal{T}}^{(t+1)} \leq \beta Cov\left[\mathcal{T}, ce\left(g_\psi, Y\right)\right] + L_{\mathcal{T}}^{(t)}$$

where β is a positive constant, $ce(\cdot)$ is the element-wise cross entropy.

- The effect of PL strategy is decoupled and encapsulated in the covariance term. If the covariance term is negative, then the loss function would be non-increasing.
- For random PL, $\mathcal{T}$ would be independent with g_ψ and the covariance becomes 0.

2. Theoretical analysis

香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Proposed model

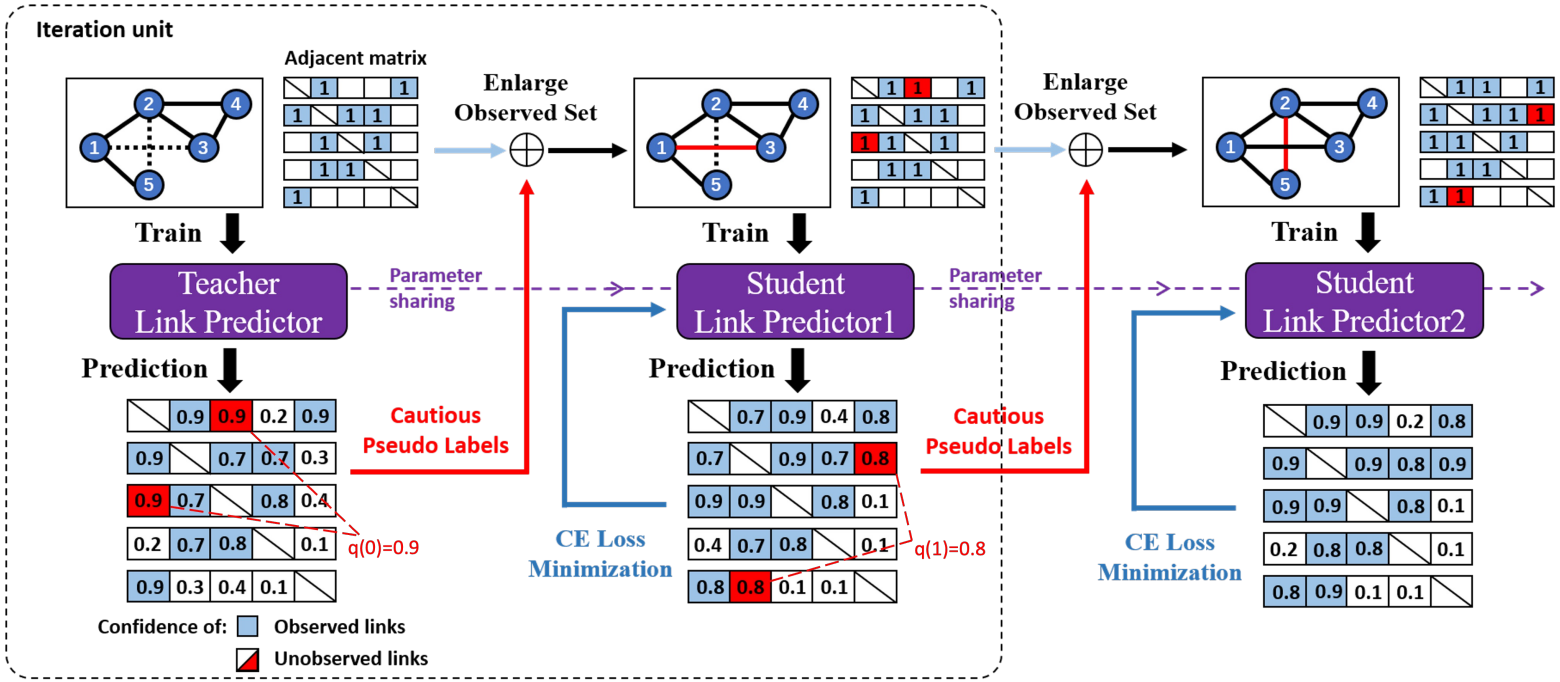

Fig.3 Main scheme of Cautious Pseudo Label (CPL) in link prediction

- We calculate the averaged confidence of the multi-view augmentation.
- We select PL samples in unobserved set with the top-k confident samples.

3. Experiments

香港科技大學 THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

3.1 Overall performance

Table 1: Performance (AUC/%) comparison on link prediction

	Model	Citeseer	Actor	WikiCS	TwitchPT	Amazon_Photo
AUC(%)	GAE	71.10 ± 0.56	55.34 ± 0.57	90.81 ± 0.69	74.48 ± 3.03	67.92 ± 1.31
	GAE+CPL	**72.45 ± 0.24**	**65.58 ± 1.04**	**95.56 ± 0.24**	**79.67 ± 3.77**	**76.30 ± 1.84**
	node2vec	52.03 ± 0.60	53.30 ± 0.59	88.82 ± 0.28	79.46 ± 0.77	89.32 ± 0.21
	node2vec+CPL	**55.22 ± 1.63**	**65.11 ± 2.31**	**91.99 ± 0.26**	**84.76 ± 3.52**	**89.53 ± 0.30**
	SEAL	63.60 ± 0.01	73.41 ± 0.02	86.01 ± 0.04	87.80 ± 0.01	76.96 ± 0.17
	SEAL+CPL	**64.33 ± 0.14**	**73.54 ± 0.01**	**86.83 ± 0.07**	**87.87 ± 0.01**	**78.86 ± 0.01**
AP(%)	GAE	72.12 ± 0.63	53.60 ± 1.06	90.58 ± 0.71	69.73 ± 5.06	67.06 ± 0.99
	GAE+CPL	**73.54 ± 0.20**	**67.65 ± 1.06**	**95.58 ± 0.29**	**79.09 ± 5.48**	**75.52 ± 4.23**
	node2vec	52.90 ± 0.36	55.43 ± 0.62	92.54 ± 0.51	83.37 ± 0.52	91.46 ± 0.18
	node2vec+CPL	**56.19 ± 1.60**	**68.33 ± 2.85**	**93.66 ± 0.29**	**85.87 ± 2.15**	**91.47 ± 0.21**
	SEAL	64.38 ± 0.01	73.17 ± 0.12	83.63 ± 0.16	87.69 ± 0.01	73.72 ± 0.56
	SEAL+CPL	**64.94 ± 0.14**	**73.44 ± 0.02**	**86.72 ± 0.12**	**87.75 ± 0.02**	**80.36 ± 0.09**

- CPL distinctively **improves the performance** of baseline models in nearly all cases in link prediction. The performance gain under the circumstances of **both high and low performance**

Table 2: Performance (AUC%) comparison on node classification

Model		Cora	CiteSeer	PubMed	Amazon_Photo	LastFMAsia
GCN	Raw	80.74 ± 0.27	69.32 ± 0.44	77.72 ± 0.46	92.62 ± 0.45	78.53 ± 0.60
	M3S	80.92 ± 0.74	72.7 ± 0.43	79.36 ± 0.64	93.07 ± 0.25	79.49 ± 1.42
	DR-GST	83.54 ± 0.81	72.04 ± 0.53	77.96 ± 0.25	92.89 ± 0.16	79.31 ± 0.55
	Cautious	**83.94 ± 0.42**	**72.96 ± 0.22**	**79.98 ± 0.92**	**93.15 ± 0.24**	**79.92 ± 0.61**
GraphSAGE	Raw	81.12 ± 0.32	69.80 ± 0.19	77.52 ± 0.38	92.46 ± 0.17	80.23 ± 0.28
	M3S	83.02 ± 0.49	70.98 ± 2.14	79.12 ± 0.25	92.41 ± 0.14	81.48 ± 0.56
	DR-GST	81.02 ± 1.99	72.28 ± 0.35	76.96 ± 0.43	92.58 ± 0.14	81.10 ± 0.30
	Cautious	**84.62 ± 0.19**	**73.14 ± 0.21**	**79.72 ± 0.72**	**92.90 ± 0.20**	**82.25 ± 0.25**
GAT	Raw	81.28 ± 0.87	71.18 ± 0.43	77.34 ± 0.34	93.26 ± 0.31	81.12 ± 0.58
	M3S	82.28 ± 0.95	71.7 ± 0.72	79.20 ± 0.21	93.71 ± 0.16	81.82 ± 0.93
	DR-GST	83.32 ± 0.31	72.64 ± 0.97	78.28 ± 0.32	93.60 ± 0.13	81.86 ± 0.50
	Cautious	**83.86 ± 0.22**	**73.02 ± 0.37**	**79.62 ± 0.31**	**93.72 ± 0.29**	**82.89 ± 0.56**
APPNP	Raw	82.52 ± 0.69	70.82 ± 0.24	79.96 ± 0.50	93.05 ± 0.29	82.40 ± 0.50
	M3S	82.54 ± 0.40	72.58 ± 0.45	79.98 ± 0.14	93.21 ± 0.59	83.55 ± 0.71
	DR-GST	82.46 ± 0.87	72.64 ± 0.54	80.00 ± 0.48	93.12 ± 0.32	82.88 ± 0.35
	Cautious	**84.20 ± 0.42**	**74.22 ± 0.24**	**80.62 ± 0.24**	**93.48 ± 0.23**	**83.56 ± 0.53**

- CPL consistently improves base models' performance and outperforms the others. Other PL strategy may be ineffective or degrade the base model.

3. Experiments

香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

3.2 Ablation experiments

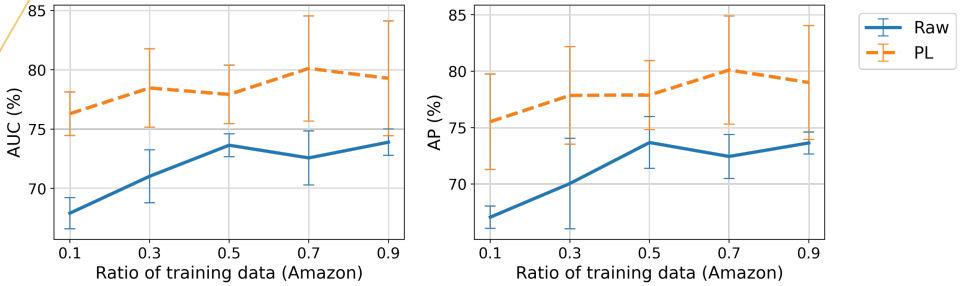

Fig.4 The effect of training ratio

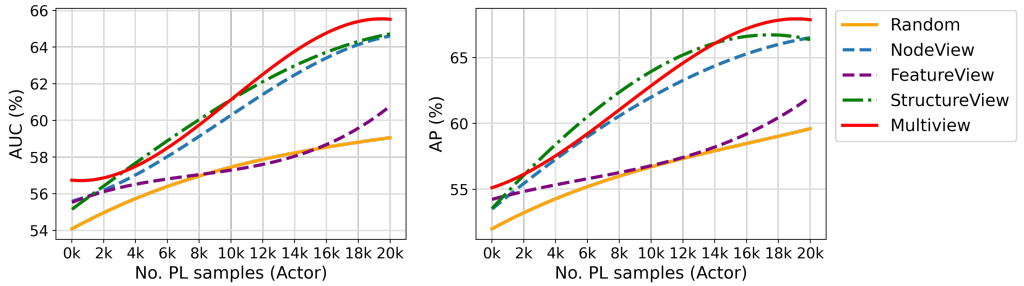

Fig.5 Impact of multi-view augmentation

- CPL consistently improves the performance of raw models even starting from a small training set.

- Multi-view augmentation contributes to a more robust graph learning and tends to obtain a consistent result.

Table 4: Case study on error analysis

	GCN	GraphSAGE	GAT	APPNP
Inconsistency $\mathcal{A}$ (%)	6.69	4.01	2.96	3.13
Confidence $1-q$ (%)	77.63	88.35	83.00	86.86
Theoretical Err_{th} (%)	58.12	31.32	39.92	32.54
Experimental Err_{exp} (%)	20.08	17.75	17.11	16.44
PL error (%)	7.78	6.43	8.18	6.02
M3S PL error (%)	65.63	27.00	65.00	27.49
DR-GST PL error (%)	26.31	14.10	13.38	29.09

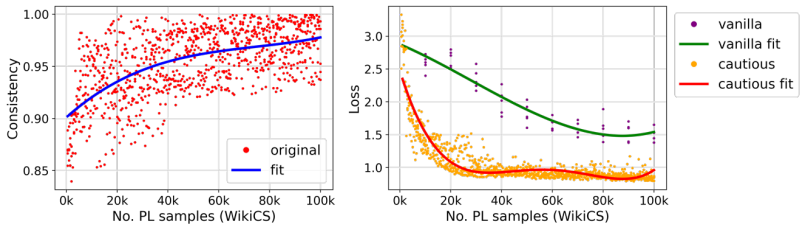

Fig.6 Case study on convergence analysis

- Case study on the error analysis shows the effectiveness of the error bond in the Theorem and CPL can improve the improve the convergence property.

4. Conclusion and prospect

香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Conclusion:

We provide deep insights into PL strategy:

- Offer theoretical explanations for the effect of PL strategies on prediction error and the convergence properties in graph learning.
- Introduce CPL strategy, a plug-in and practical technique that can be generally applied to various baseline models.
- The experiments demonstrate effectiveness and superiority of CPL.

Prospect

- We plan to explore a more reliable confidence measures as the PL criteria, such as informativeness in the multi-view network and prediction uncertainty.

INFORMS Annual Meeting 2023

ME48.Spatiotempora Decision Intelligence

Thanks for listening