ON THE EQUIVALENCE OF DECOUPLED GRAPH CONVOLUTION NETWORK AND LABEL PROPAGATION

A NOTE

Zepeng Zhang

July 8, 2022

This paper [1] proved that a decoupled GCN is essentially the same as a two-step label propagation: 1) propagating the known labels along the graph to generate pseudo-labels for the unlabeled nodes, and 2) training a neural network predictor on the augmented pseudo-labeled data. Formally, the model is optimized with the following objective function:

$$L(\theta) = \ell\left(f_\theta(X), \bar{A}Y\right),$$

where $\bar{A} = \sum \beta_k \hat{A}^k$ is determined by the propagation strategy and Y represent the one-hot labels for nodes (initial labels for unlabeled nodes are zero vectors). Using the cross-entropy loss, we have

$$\begin{aligned} L(\theta) &= -\sum_{i\in\mathcal{V}, k\in C}\left(\sum_{j\in\mathcal{V}_l}\bar{a}_{ij}y_{jk}\right)\log f_{ik} \\ &= -\sum_{i\in\mathcal{V}, j\in\mathcal{V}_l}\bar{a}_{ij}\sum_{k\in C}y_{jk}\log f_{ik} \\ &= \sum_{i\in\mathcal{V}, j\in\mathcal{V}_l}\bar{a}_{ij}\mathrm{CE}\left(f_i, \boldsymbol{y}_j\right), \end{aligned}$$

where C is the label set and $\mathcal{V}_l$ is the set of labeled nodes. The reformulated loss can be interpreted as using node j's label to train the node i and weighting the loss with $\bar{a}_{ij}$. However, the static weight $\bar{a}_{ij}$ maybe unsatisfying and this paper extends it to a more general framework with flexible weighting strategies:

$$L_{PT} = L(\theta) = \sum_{i\in\mathcal{V}, j\in\mathcal{V}_l} w_{ij}\mathrm{CE}\left(f_i, \boldsymbol{y}_j\right),$$

where $w_{ij} = g(f(X), A)$ represents a general weighting strategy that is controlled by the model prediction and graph structure with a specific function g.

Lemma 1. *Training a decoupled GCN is equivalent to performing Propagation then Training with dynamic weight* $w_{ij} = \frac{\bar{a}_{ji}f_{i,h(j)}}{\sum_{q\in\mathcal{V}}\bar{a}_{jq}f_{q,h(j)}}$ *for each pseudo-labeled data* $(x_i, \mathbf{y}_j)$*, where* $h(j)$ *means the label ID of node* j*, i.e.,* $y_{j,h(j)} = 1$*.*

Proof. The lemma can be proved by comparing the gradients of decoupled GCN and PT. The loss of decoupled GCN is

$$L_{DGCN} = \ell\left(\bar{A}f_\theta(X), Y\right) = -\sum_{j\in\mathcal{V}_l, k\in C} y_{jk}\left(\log\sum_{i\in\mathcal{V}}\bar{a}_{ji}f_{ik}\right),$$

whose gradient is

$$\begin{aligned}
\nabla_\theta L_{DGCN} &= -\sum_{j\in\mathcal{V}_l,k\in C} y_{jk}\nabla_\theta\left(\log\sum_{i\in\mathcal{V}}\bar{a}_{ji}f_{ik}\right)\\
&= -\sum_{j\in\mathcal{V}_l,k\in C} y_{jk}\frac{\sum_{i\in\mathcal{V}}\bar{a}_{ji}\nabla_\theta f_{ik}}{\sum_{q\in\mathcal{V}}\bar{a}_{jq}f_{qk}}\\
&= -\sum_{j\in\mathcal{V}_l} y_{j,h(j)}\frac{\sum_{i\in\mathcal{V}}\bar{a}_{ji}\nabla_\theta f_{i,h(j)}}{\sum_{q\in\mathcal{V}}\bar{a}_{jq}f_{q,h(j)}}\\
&= -\sum_{i\in\mathcal{V},j\in\mathcal{V}_l}\frac{\bar{a}_{ji}f_{i,h(j)}}{\sum_{q\in\mathcal{V}}\bar{a}_{jq}f_{q,h(j)}}y_{j,h(j)}\frac{\nabla_\theta f_{i,h(j)}}{f_{i,h(j)}}\\
&= \sum_{i\in\mathcal{V},j\in\mathcal{V}_l}\frac{\bar{a}_{ji}f_{i,h(j)}}{\sum_{q\in\mathcal{V}}\bar{a}_{jq}f_{q,h(j)}}\nabla_\theta\mathsf{CE}\left(f_i,\boldsymbol{y}_j\right).
\end{aligned}$$

Note that the gradients of the PT w.r.t. θ is

$$\nabla_\theta L_{PT} = \sum_{j\in\mathcal{V}_l,i\in\mathcal{V}} w_{ij}\nabla_\theta\mathrm{CE}\left(f_i,\boldsymbol{y}_j\right).$$

By comparing the two gradients the proof is completed. □

From the above lemma, we can see that the weights for each nodes are normalized to be unity, i.e., $\sum_{i\in\mathcal{V}} w_{ij} = 1$. To make the model more robust to label noise, this paper remove the normalization of weights to let different labeled data exert varying impact on model training and propose PTA. The weight without normalization can be written as

$$w_{ij} = \bar{a}_{ji}f_{i,h(j)}.$$

Then the accumulated weight is

$$S_j = \sum_{i\in\mathcal{V}}\bar{a}_{ji}f_{i,h(j)}.$$

When $f_{i,h(j)}$ is small, it implies the labels of neighbors of node j are predicted different from $h(j)$, which means the labeled data may be contaminated by noises. The PTA naturally assigns small weights to this unreliable data and make the model more robust to label noise. An adaptive weighting strategy is further used to reduce the sensitivity to model initialization:

$$w_{ij} = \bar{a}_{ji}f_{i,h(j)}^{\gamma},\quad \gamma = \log(1+e/\epsilon).$$

PTA gradually enlarges the impact of model prediction on the weighting of pseudo-labeled data to increase the model's robustness to noise.

References

[1] Hande Dong, Jiawei Chen, Fuli Feng, Xiangnan He, Shuxian Bi, Zhaolin Ding, and Peng Cui. On the equivalence of decoupled graph convolution network and label propagation. In *Proceedings of the Web Conference 2021*, pages 3651–3662, 2021.