[AAAI 2020] Multi-Stage Self-Supervised Learning for Graph Convolutional Networks on Graphs with Few Labels. [paper]

Node/Graph Tasks: Node classification

Training Type: self-training. Here we define another training method for self-supervised learning, self-training, which is to apply the original supervised learning to generate pseudo labels for nodes without annotation. These generated pseudo labels are combined with the original ground-truth labels and feed together to supervised learning model again to make the prediction. The whole procedure is performed iteratively.

Pretext task data: graph topology, node features
For self-training, there is no specific pretext task.

Initial short summary here:
To handle the challenge of short of labeled data in training GNN models, this paper leverages the multi-stage training framework to utilize the information of the pseudo labels generated by predictions in the previous iterations in the future training. Besides, the DeepCluster method is applied to generate clusters which are then used to correct the pseudo labels generated in the multi-stage training framework.

The multi-stage training algorithm repeatedly adds the most confident predictions of each class to the label set and re-utilize these pseudo label data to train the GNN model. Furthermore, a self-checking mechanism based on DeepCluster is proposed to guarantee the precision of labeled data. In GNNs setting, the DeepCluster takes a set of embedding vectors produced by GNN based model F as input and groups them into k distinct clusters represented by a $d \times k$ centroid matrix C by solving:

$$\min_{C \in \mathbb{R}^{d \times k}} \frac{1}{N} \sum_{n=1}^{N} \min_{y_n \in \{0,1\}^k} ||F(x_n) - Cy_n||_2^2, \quad s.t. \quad y_n^{\mathrm{T}} 1_k = 1 \tag{51}$$

After applying DeepCluster to cluster nodes into multiple clusters, aligning mechanism is used to assign nodes in each cluster to their corresponding class. For each cluster l in unlabled data, the computation of aligning mechanism is:

$$c^{(l)} = \arg\min_m ||| v_l - \mu_m ||^2, \tag{52}$$

where μ_m denotes centroids of class m in labled data, v_l denotes the centroid of cluster l in unlabeled data and $c^{(l)}$ represents the aligned class that has the closest distance from v_l among all centroids of class in the original labeled data. Note that we can perform self-checking way naively by simply comparing the distance of each unlabeled node to centroids of classes in labeled data. However, when the number of clusters is equivalent to the amount of all unlabeled nodes, the self-checking mechanism performed here is the same as the naive way. Therefore to reduce the computational load, we don't directly checking using naive way.

By performing the node classification on citation networks, all self-training methods outperform the original GCNs with a large margin, espically when the graph has low label rate. The margin decreases as the labeled data increases. Sensitivity analysis of number of clusters suggest that class tends to be more balanced as the number of clusters increases, facilitating the final performance of M3S training algorithm.

Bibtex:
@inproceedingssun2020multi, title=Multi-stage self-supervised learning for graph convolutional networks on graphs with few labeled nodes, author=Sun, Ke and Lin, Zhouchen and Zhu, Zhanxing, booktitle=Proceedings of the AAAI Conference on Artificial Intelligence, volume=34, number=04, pages=5892–5899, year=2020