Tackling the Imbalance for GNNs

Rui Wang, Weixuan Xiong, Qinghu Hou, Ou Wu

Center for Applied Mathematics, *Tianjin University*, Tianjin, China

wr_wangrui@foxmail.com, weixuan@tju.edu.cn, qh_hou@tju.edu.cn, wuou@tju.edu.cn

Abstract—Different from deep neural networks for non-graph data classification, graph neural networks (GNNs) leverage the information exchange between nodes (or samples) when representing nodes. The category distribution shows an imbalance or even a highly-skewed trend on nearly all existing benchmark GNN data sets. The imbalanced distribution will cause misclassification of nodes in the minority classes, and even cause the classification performance on the entire data set to decrease. This study explores the effects of the imbalance problem on the performances of GNNs and proposes new methodologies to solve it. First, a node-level index, namely, the label difference index (LDI), is defined to quantitatively analyze the relationship between imbalance and misclassification. The less samples in a class, the higher the value of its average LDI; the higher the LDI of a sample, the more likely the sample will be misclassified. We define a new loss and propose four new methods based on LDI. Experimental results indicate that the classification accuracies of the three among our proposed four new methods are better in both transductive and inductive settings. The LDI can be applied to other GNNs.

Index Terms—graph neural networks, imbalanced problem, label difference index

I. INTRODUCTION

As an increasing number of applications involve graph data, researchers have designed graph neural networks (GNNs) to process graph data. GNNs generally use the connections between nodes to exchange information in neighborhoods to obtain a better representation for each node. Therefore, a unique problem faced by GNNs is misclassification caused by neighbor nodes. Nodes in minority classes are more likely to be neighbored by those in majority classes, resulting in low accuracies.

Real-world data usually conform to an imbalanced even a highly-skewed distribution, in which majority classes occupy most of the data proportion. Deep neural networks also suffer when the training data are highly imbalanced [1].

This study investigates the imbalance problem for GNNs, and reveals more details for the relationship between imbalance and misclassification on graph data. Furthermore, based on the state-of-the-art algorithms for dealing with the imbalance of non-graph data, new methodologies are proposed. First, we define the label difference index (LDI) to measure the likelihood of a node being misclassified based on the category distribution of its neighborhood. The larger the index, the more likely the node can be misclassified. The LDI of a node is determined by the category distribution of its

neighbor nodes. Nodes neighbored by heterophily nodes[1] are more likely to be with larger LDIs, and easy to be negatively affected, resulting in misclassification. In addition, LDI is also affected by the global category distribution of the entire data set, and statistics show that the average LDIs of the nodes in minority classes are usually large. Thus, the relationship between category imbalance and misclassification is established. Second, we define a new loss and propose four new methods, namely, improved focal loss (iFL), Graph Re-sampling (GRS), Graph Re-weighting (GRW), Graph Metric Learning (GML), and Graph Bilateral-branch Network (GBBN). Experiments on several graph benchmark data sets show that except for GML, the new methods are better than the original GNNs. Imbalanced sampling with LDI can help to further improve performance.

Our contributions are summarized as follows:

- A node-level index (i.e., LDI) is defined to characterize the neighborhood of a node (or a sample) in a graph. On the basis of LDI, the relationship between imbalance and misclassification is analyzed and useful observations are obtained.
- A new loss and four new methods are proposed based on the index LDI. Extensive experiments on benchmark data sets indicate that the proposed methodology achieves better results than existing imbalance learning methods on graphs. Further, several classical GNNs are improved by utilizing LDI.

II. RELATED WORK

Classical GNNs include GCN [6], SGC [15], GAT [13], and so on. Kipf et al. [6] expanded the traditional convolution neural network (CNN) on high-dimensional graph data to obtain GCN. GCN iteratively updated each node's representation through the message exchange with their neighbor nodes. Wu et al. [15] reduced the complexity of GCN by repeatedly eliminating the nonlinearity between GCN layers and folding the resulting function into a linear transformation to obtain SGC. Veličković et al. [13] proposed GAT based on the attention mechanism to classify graph data.

In the research of graph, researchers usually define indices (e.g., degree and centrality) to characterize the properties of a graph. Likewise, a number of indices are also defined to measure the properties of the involved graphs in the research of GNNs. For example, graph homophily was proposed by Pei

---

* Ou Wu is the corresponding author.

[1]Heterophily nodes are the samples (or nodes) that belong to different categories [19].

et al. [10] to characterize the degree that similar nodes connect together. Graph heterophily was proposed by Zhu et al. [20] to measure the graph homophily level. Two graph smoothness metrics, namely, feature smoothness and label smoothness, were proposed by Hou et al. [4] to help understand the use of graph information in GNNs. Almost all existing graph indices in GNNs are graph-level or category-level[2]. These indices are only used to analyze graph characteristics and the learning performances. They are not involved in the training process. To our knowledge, there is no node-level index. This study will define a node-level index which can be directly used in both characteristics analyses and model training.

Machine learning has been increasingly applied in recent years. The imbalance has become a hot research topic. Kang et al. [5] divided the learning of a classification model into two steps. Zhou et al. [18] proposed a new learning model with two branches, both of which involve re-sampling. Zong et al. [21] proposed GNN-XML to overcome the imbalance problem in extreme multi-label text classification. Using re-weighting, Cui et al. [1] used a better weighting design to achieve a better imbalance classifier. Liu et al. [8] created some virtual samples around the minority classes samples to increase the number of minority classes samples. Liu et al. [9] transferred the visual knowledge of the majority to the minority classes by learning a set of dynamic element vectors. Lin et al. [7] proposed a new loss function, namely, focal loss, to deal with the imbalance problem in object detection.

As to the imbalance of graph data, Min et al. [12] conducted a pilot study for this problem. They proposed DR-GCN, adopting a conditional adversarial training together with distribution alignment to learn node representations. They only investigated the transductive setting. Therefore, DR-GCN is only compared in the transductive setting in Section V-B.

III. EFFECT OF IMBALANCE FOR GNNS

A. Qualitative analysis.

We first analyze the interaction between the majority and the minority classes in an imbalanced data set of a non-graph classification task from the perspective of training loss. The loss can be expressed as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{majority} + \mathcal{L}_{minority}, \tag{1}$$

where $\mathcal{L}_{majority}$ is the loss of the majority classes and $\mathcal{L}_{minority}$ is the loss of the minority classes. Because the number of samples in the majority classes (majority samples for brevity) is much greater than that of samples in the minority classes (minority samples for brevity), the loss is composed mainly of majority samples, and the gradients of model optimization are determined more by majority samples. As a result, both the feature representation and the decision layers of the training model are optimized towards a high classification accuracy of the majority classes in each training

---

[2]Several indices first define a node-level metrics and using the average of all nodes as the graph-level or category-level index. The node-level index is then no longer used.

epoch, which may lead to the overfitting of the majority classes and ignores the minority classes while training. Ultimately, the performance of the overall model is damaged.

In graph classification tasks, because the loss function can still be expressed in the above form, the training for GNNs will also be affected by the imbalance, that is, the majority samples determine the gradients in training. However, because the majority and the minority nodes in the graph usually have direct edge links, in addition to the negative effects of the loss of coupling on the minority classes, the nodes in the majority classes may affect the feature representation of the nodes in the minority classes directly through edges. When node heterophily holds, this kind of effect between neighbor nodes is very likely to be negative.

Information exchange between nodes of the same category is beneficial, while information exchange between nodes of different categories is likely to be harmful in GNNs. Because there are fewer minority nodes, the proportion of nodes in minority classes with heterophily nodes in their neighborhoods may be relatively larger, which may have a greater negative effect. In other words, minority nodes are more likely to be misclassified.

To sum up, imbalance in non-graph data affects both the decision and the feature representation layers through the loss minimization; imbalance in graph data affects both the decision and the feature representation layers according to both the loss minimization and the direct information exchange between nodes. Therefore, imbalance in graph data can have a more serious impact on GNNs training.

B. Quantitative analysis.

Suppose a graph $G = (V, E)$ and let $x_i \in V$. The category distribution of the K-neighbor nodes of x_i is $P_{\mathcal{N}_i} = \{p_{i,1}, \cdots, p_{i,c}, \cdots, p_{i,C}\}$, where $\mathcal{N}_i$ denotes the K-neighborhood and $p_{i,c}$ represents the proportion of the cth category in $\mathcal{N}_i$. We use 1-neighborhood to reduce the computation complexity in this study. Meanwhile, if x_i belongs to the cth category, the distribution of itself is marked as a one-hot vector $P_{\mathcal{I}_i} = \{0, \cdots, 0, 1, 0, \cdots, 0\}$, where the cth element is 1.

We define a node-level index, label difference index (LDI), on graph to better characterize the relationship between imbalance and misclassification.
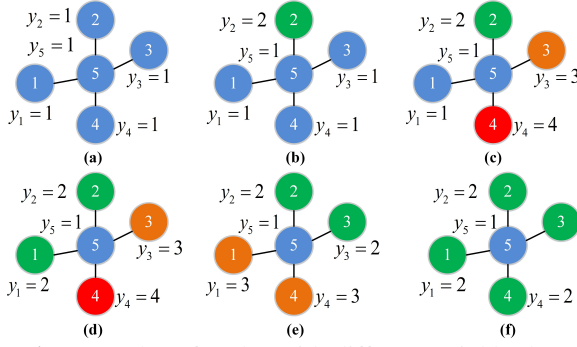
Label Difference Index (*LDI*): Given a node x_i, the LDI of x_i is

$$LDI_i = \frac{1}{\sqrt{2}}\|P_{\mathcal{N}_i} - P_{\mathcal{I}_i}\|_2. \tag{2}$$

The range of LDI_i is $[0, 1]$ with the constant $\frac{1}{\sqrt{2}}$. The higher the LDI value, the larger the difference between the two distributions. If x_i is an isolated node, then we define its LDI as the average LDI of nodes in the same category. We have the following proposition.

Proposition 1: Given a node x_i belonging to the cth category. When the proportion of homophily nodes in its neighborhood is fixed (i.e., $p_{i,c}$ is fixed), the less the number

Fig. 1: Examples of nodes with different neighborhoods. The LDI values of node 5 in (a)-(f) are 0, $\frac{1}{4}$, $\frac{\sqrt{6}}{4}$, $\frac{\sqrt{11}}{4}$, $\frac{\sqrt{3}}{2}$, and 1, respectively.

of categories of the heterophily nodes in its neighborhood, the larger the LDI of x_i.

Proof 1: Let $\sum_{c' \neq c} p_{i,c'} = 1 - p_{i,c} = \Delta$, then

$$LDI_i = \frac{1}{\sqrt{2}} \sqrt{(1 - p_{i,c})^2 + \sum_{c' \neq c} p_{i,c'}^2}. \quad (3)$$

By using Cauchy Inequality, we have

$$\frac{(\sum_{c' \neq c} p_{i,c'})^2}{C - 1} \leq \sum_{c' \neq c} p_{i,c'}^2 \leq (\sum_{c' \neq c} p_{i,c'})^2. \quad (4)$$

If and only if $\forall c'$, $p_{i,c'} = \frac{\Delta}{C-1}$, then

$$\frac{(\sum_{c' \neq c} p_{i,c'})^2}{C - 1} = \sum_{c' \neq c} p_{i,c'}^2 = \frac{\Delta^2}{(C - 1)^2}. \quad (5)$$

If and only if $p_{i,c''} = 1 - p_{i,c}$ and $p_{i,c} = 0 (c' \neq c'')$, then

$$\sum_{c' \neq c} p_{i,c'}^2 = (\sum_{c' \neq c} p_{i,c'})^2 = \Delta^2. \quad (6)$$

Accordingly,

$$\Delta \sqrt{\frac{1 + (C - 1)^2}{2(C - 1)^2}} \leq LDI_i \leq \Delta. \quad (7)$$

The upper bound is attained only when all the heterophily nodes belong to the same category. To sum up, the more concentrated the categories of heterophily nodes, the larger the LDI of x_i is. The proof ends.

Fig. 1 shows an illustrative example. The colors denote the categories. Assume that $P_{\mathcal{I}_5} = \{1, 0, 0, 0\}$. In Fig. 1(a), $P_{\mathcal{N}_5} = \{1, 0, 0, 0\}$, then $LDI_5 = 0$. In Fig. 1(b), $P_{\mathcal{N}_5} = \{\frac{3}{4}, \frac{1}{4}, 0, 0\}$, then $LDI_5 = \frac{1}{4}$. In Fig. 1(c), $P_{\mathcal{N}_5} = \{\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\}$, then $LDI_5 = \frac{\sqrt{6}}{4}$. In Fig. 1(d), $P_{\mathcal{N}_5} = \{0, \frac{1}{2}, \frac{1}{4}, \frac{1}{4}\}$, then $LDI_5 = \frac{\sqrt{11}}{4}$. In Fig 1(e), $P_{\mathcal{N}_5} = \{0, \frac{1}{2}, \frac{1}{2}, 0\}$, then $LDI_5 = \frac{\sqrt{3}}{2}$. In Fig. 1(f), $P_{\mathcal{N}_5} = \{0, 1, 0, 0\}$, then $LDI_5 = 1$. In other words, node 5 in Fig. 1(f) is the most easiest to be misclassified. The six LDI values in Fig. 1 indicate that our LDI definition is reasonable and Proposition 1 holds.

LDI is a node-level index that can be easily applied to concrete GNN algorithms. Different from the homophily index mentioned in [10], LDI utilizes the distribution of neighbors,

and is more fine-grained. Proposition 1 verifies this. The indices of node 5 in Figs. 1 (d), (e), and (f) are the same (equal to 0) if the homophily index in [10] is used. However, LDI can distinguish these three cases better.

Based on LDI, we conduct analyses in terms of the differences among categories, the differences among samples within specific categories, the relationship between LDI and accuracies, and the relationship between LDI and misclassification under different numbers of layers ($\#layers$).

Average *LDI*s of different categories. We calculate the average LDIs of different categories on five benchmark graph data sets, as shown in Fig. 2. Although the trend of the category distribution (Fig. 2 (up)) is not completely the opposite of the trend of the LDI distribution (Fig. 2 (middle)) on each data set, the overall trend is that the LDIs of minority classes are large, indicating that the proportions of heterophily nodes around the nodes in the minority classes are relatively high. In Fig. 2 (down), the categories with large average LDIs usually have low accuracies.
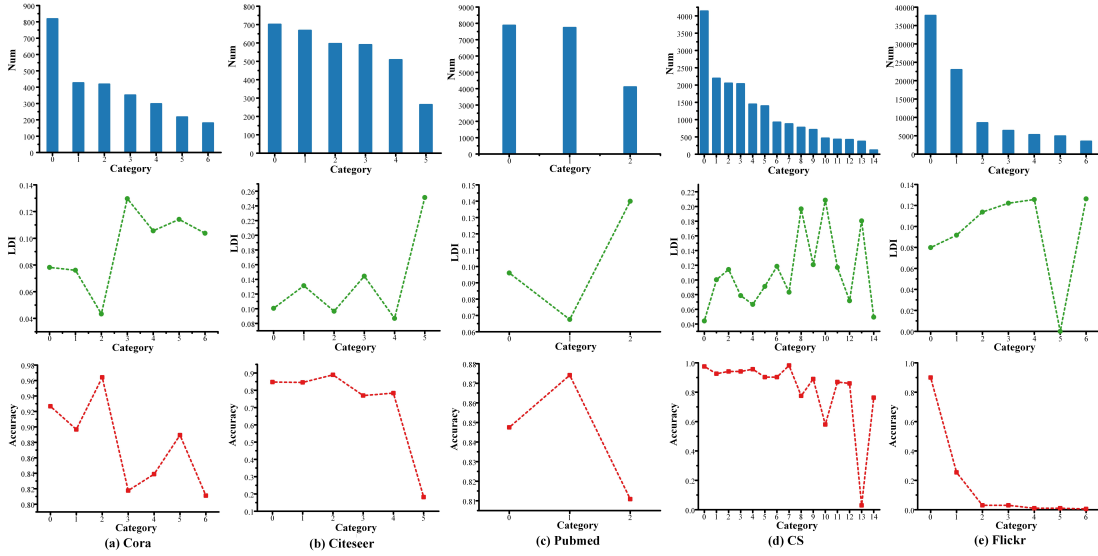
***LDI* distributions within specific categories.** Within a category, the LDIs of different samples are also distinct. We arbitrarily select a majority class and a minority class from the graph data set Citeseer [12]. As shown in Fig. 3, samples in the majority class are more concentrated in the small LDI intervals. Contrarily, samples in the minority class not only concentrate in the large LDI intervals. That is, samples with large LDIs can be found in the majority classes and samples with small LDIs can be found in the minority classes.

Average accuracies of different *LDI* intervals. Furthermore, we analyze the average accuracies of nodes in different LDI intervals. As shown in Fig. 4, nodes concentrate in small LDI intervals, and the classification accuracy decreases as the LDI increases on different data sets. Nodes with large LDIs are more likely to exchange information with heterophily nodes in their neighborhoods, resulting in low accuracies. As shown in Fig. 3, within a specific category, the average accuracies of different LDI intervals show that samples with large LDI values in the majority classes can also be more likely to be misclassified. Thus establishing a reasonable node-level index to supplement the existing category-level index is necessary.
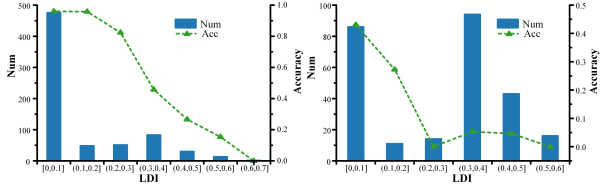
Relationship between *LDI* and misclassification under different #layers. The performance of GNNs decreases with the increase of layers (e.g., $\#layers > 4$), which is mainly caused by oversmoothing. We explore the relationship between LDI and the misclassification caused by oversmoothing. Let the correctly predicted sample set be R_n when the number of layers equals to n. We calculate the ratio of the average LDI of the newly wrongly predicted samples (i.e., $\overline{R_n} \cap R_{n-1}$) when the layers increase by one to the average LDI of the correctly predicted samples before the layer increases, that is,

$$r_n = \frac{LDI_{avg}(\overline{R_n} \cap R_{n-1})}{LDI_{avg}(R_{n-1})}, \quad (8)$$

where $LDI_{avg}(\Omega)$ represents the average LDI of set Ω.

Fig. 2: Category distributions (up), average LDIs (middle), and average accuracies (down) of the five different data sets used in this study. Categories with small average LDIs always have high accuracies. On Flickr, Category 5 has a small average LDI but a small average accuracy. This is mainly because loss minimization and feature exchange between nodes negatively affect the accuracy of minority classes, and loss minimization plays a main role.

Fig. 3: Sample distributions and average accuracies for different LDI intervals of the majority (left) and the minority class (right) of Citeseer. In the majority class, samples concentrate in small LDI intervals. In the minority class, samples disperse more uniformly.

In Fig. 5, as the $\#layers$ increases, the value of r_n is greater than 1 in nearly all cases. It shows that as the number of layers increases, samples with large LDIs are more likely to be misclassified.

To sum up, the category distribution of neighbor nodes that LDI relies on is related closely to the global category distribution as shown in Fig. 2. LDI reflects the possibility of each node being misclassified as shown in Fig. 4. Therefore, on the basis of the global distribution and LDI, we proposed several new methods to tackle the imbalance for GNNs in the succeeding section.

IV. METHODOLOGY

A new loss and four new methods, namely, iFL, GRS, GRW, GML, and GBBN, are proposed in this section.

A. Improved Focal Loss (iFL).

Focal loss is designed to solve the imbalance problem in object classification. For binary classification, the focal loss can be defined as

$$\mathcal{L}_{FL} = -\sum_i \alpha_{y_i}(1 - p_{i,y_i})^\gamma \log(p_{i,y_i}), \tag{9}$$

where p_{i,y_i} is the probability that node i is correctly classified; α_{y_i} and γ are hyperparameters. For binary classification, the value of α_{y_i} is easy to set. Nevertheless, the number of categories for GNN benchmark data sets is ususally large, resulting in that it is difficult to set α_{y_i} properly. Therefore, to avoid the manually setting for α_{y_i}, an improved focal loss is defined as follows:

$$\mathcal{L}_{iFL} = -\sum_i (1 - \overline{p}_{y_i})^\gamma \log(p_{i,y_i}), \tag{10}$$

where $\overline{p}_{y_i}$ can be seen as the average probability that nodes in the category y_i are correctly classified. For a category with a low classification accuracy, the value of $(1 - \overline{p}_{y_i})^\gamma$ will be large.
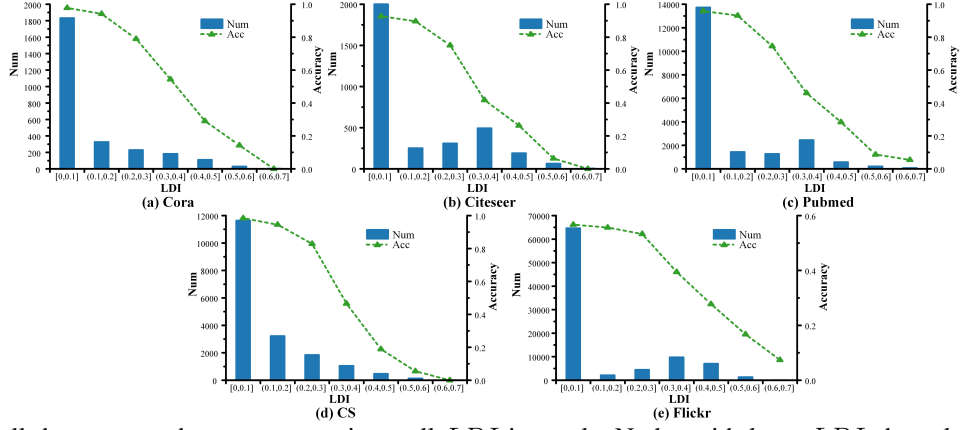
B. Graph Re-sampling (GRS) and Graph Re-weighting (GRW).

Re-sampling/re-weighting essentially resamples/reweights training samples during network training to strengthen the learning of the minority classes.

Our proposed GRS and GRW are based on re-sampling and re-weighting, respectively. In this study, re-sampling and re-weighting are performed according to the numbers of samples in different categories, and give more weights to the minority classes with smaller sample sizes. LDI is also utilized. Suppose the category of x_i is c. N_c is the number of samples of category c. N is the total number of samples of the data set. For x_i, there are three weighting strategies:

W1: Label-based weighting:

$$w_i^L = \frac{N}{N_c}. \tag{11}$$

Fig. 4: For all data sets, nodes concentrate in small LDI intervals. Nodes with large LDIs have low accuracies.

Fig. 5: Variations in r (solid lines) and accuracy (dot lines) with increasing #layers using GCN and GAT under insductive (left) and transductive (right) settings on the Citeseer data set.

W2: LDI-based weighting[3]:

$$w_i^D = e^{LDI_i}. \qquad (12)$$

W3: Combination of label and LDI:

$$w_i^{LD} = w_i^L \cdot w_i^D. \qquad (13)$$

C. Graph Metric Learning (GML).

Metric learning aims to learn a feature space in which the distances between homophily samples are close and the distances between heterophily samples are far. Dong et al. [2] applied a triplet loss-based metric learning to deal with the imbalance problem in image classification.

Triplet loss is used in this study. The key to triplet loss-based metric learning is to select the appropriate triplets as training samples. When choosing a triplet, the anchor point is first selected; the positive and negative sample pairs corresponding to it are then selected from other samples. According to the probability of being easy to be classified correctly, anchor points are divided into easy ones and hard ones. Because the hard triplets are not conducive to the learning of the model, Wang et al. [14] proposed that only choosing the easy anchors would obtain better performance.

In this study, LDI is used to measure the hardness of a sample. The top 10% samples with the highest LDI values are removed (i.e. the top 10% hardest samples are removed). For the remaining samples, the larger the LDI of a sample, the more likely it is to be selected. For each anchor, in its K-neighborhood, the negative samples are selected according to the closeness of distances. Outside its K-neighborhood,

---

[3]Because LDIs of some samples are zero, to prevent the weights from being zero, e^{LDI_i} is used as the weight.

---

the positive samples are selected according to the further distances. The candidates are shown in Fig. 6. Let the tth triplet be (x_t, x_{t+}, x_{t-}), its score can be calculated by

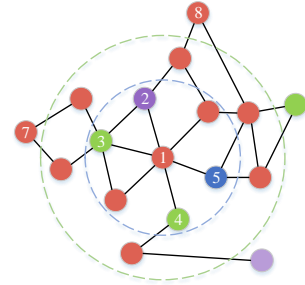$$score_t = max(0, m + d(x_t, x_{t+}) - d(x_t, x_{t-})), \qquad (14)$$

where m refers to margin, and $d(\cdot)$ represents the distance between two feature vectors. Assume that T is the set of selected triplets with relatively high scores. The triplet loss can be represented as

$$\mathcal{L}_{ML} = \frac{1}{|T|} \sum_t score_t. \qquad (15)$$

The triplet loss and the focal loss form the final loss function. Following the setting of Zhou et al. [18], we also adopt a cumulative learning strategy to combine the above two losses:

$$\mathcal{L}_{GML} = f(e)\mathcal{L}_{ML} + \mathcal{L}_{iFL}, \qquad (16)$$

where $f(e)$ refers to the weight in the current epoch.

Fig. 6: Node 1 is the selected anchor. Let $K = 2$, nodes 2, 3, 4, and 5 represent negative candidate nodes, and the outermost two red nodes 7 and 8 are positive candidate nodes.

D. Graph Bilateral-Branch Network (GBBN).

Our GBBN is based on the bilateral-branch network proposed by Zhou et al. [18] to solve the imbalance in visual recognition tasks. The network consists of two branches: the conventional learning branch and the re-balancing branch. The conventional branch samples uniformly from the original data, maintaining the original data distribution for feature learning. The re-balancing branch is a reverse sampling, which aims to

increase the probability of minority classes and reduce data imbalance [3].

Fig. 7 shows the structure of GBBN. The conventional (upper) branch uses conventional models such as GCN and GAT. The re-balancing (lower) branch no longer adopts the same structure as the upper branch, but uses DropEdge GNN [11] to reduce information exchange between nodes. In the experiments of this study, we drop all edges to reduce the training complexity. In addition, the proposed iFL is used rather than the standard cross entropy loss. Comparing with GNN-XML proposed in [21] which aims to overcome the imbalance for text classification, the settings of two branches and the loss functions are different.

The upper branch learns the representations of the nodes according to uniform sampling. The sampled nodes also present an imbalanced distribution. The reverse sampling of the lower branch takes the label and LDI of a node into account. For samples with higher LDIs, the sampling probabilities are greater, which not only balances the data from the category-level, but also focuses on the samples that are easily misclassified from the node-level. For this branch, this study leverages the three weighting strategies defined in Section IV-B, namely, W1, W2, and W3.

The graph is entered into the two branches. Two logit outputs of each sample under GNN and DropEdge GNN are obtained and denoted as x^G and x^{DE}, respectively. Then, a weighted logit vector is obtained by controlling the adaptive weight parameter α. The final output logit vector can be obtained by using the following formula:

$$z = \alpha x^G + (1 - \alpha)x^{DE}, \quad (17)$$

where $z \in \mathbb{R}^C$ refers to the final logic vector. The prediction is $\hat{p} = \text{Softmax}(z)$.

The classification loss of GBBN can be expressed as follows:

$$\mathcal{L}_{GBBN} = \alpha\mathcal{L}_{iFL}(\hat{p}, y^G) + (1 - \alpha)\mathcal{L}_{iFL}(\hat{p}, y^{DE}). \quad (18)$$

The network training also adopts the cumulative learning strategy. By changing the parameter α, the output logit and the loss also vary, resulting in that different training stages have different learning emphases. In the early stage, the emphasis is on the upper branch. Then, the emphasis gradually shots to the minority categories and the lower branch. This is reasonable. With the increase of training epoch, minority nodes (especially nodes with large LDIs) should rely on their own representations more than neighborhoods because their neighborhoods contain more heterophily nodes than others. In our experiments, we use linear decay, i.e. $\alpha = 1 - e/Epoch$, where e is the current epoch and $Epoch$ is the expected max epoch.

V. EXPERIMENT

We design extensive experiments to evaluate the effectiveness of the four proposed methods. Both transductive and inductive settings are tested[4].

[4]The code can be seen in https://github.com/wasdfghjklr/Tackling-the-Imbalance-for-GNNs

A. Experimental setting.

Data sets. Five[5] benchmark data sets in Fig. 2 are used: Cora [12], Citeseer, Pubmed [17], CoauthorCS (CS) [17], and Flickr [16]. For Cora and Citeseer, the train, validation, and test division ratio is 2:4:4; for others, the division ratio is 0.5:4.5:5. Previous transductive studies on the above data sets do not shuffle training data in their experiments. Data sets with and without shuffle are considered in the experiments. When data shuffle is adopted, the average results of five randomly shuffle are recorded. The data statistics are shown in Table I. All five graphs are class-imbalanced as shown in Fig. 2.

TABLE I: Data statistics of the five Graph data sets.

Name	#Node	#Edge	#Features	#Class
Cora	2,708	5,429	1,433	7
Citeseer	3,327	4,732	3,703	6
Pubmed	19,717	44,338	500	3
CoauthorCS	18,333	81,894	6,805	15
Flickr	89,250	899,756	500	7

Hyperparameters. For GCN, SGC, and GAT, the #hidden units of Cora, Citeseer, and Pubmed, dropout rate, and $l2$ regularization penalty settings are the same as [17]. For Flickr, the #hidden unit is set the same as [16]. The max epoch is set to 1500. For GML, K is set to 2, the margin is set to 0.2, ϵ is set to 0.001, and ρ is set to 0.3. For GBBN, α is set to 0.5 in the validation and testing phrases. In the comparison with DG-GCN, the settings in [12] are followed.

B. Results.

Comparison with DR-GCN. DR-GCN [12] only focuses on the transductive setting. Therefore, only the weighting strategy W1 is used in GRS, GRW, and the lower branch of GBBN. The reason why W2 and W3 are not used is because the input data under transductive setting is the entire graph, when calculating the LDIs of the training set, the labels of the validation set and the test set will be referred to, resulting in label leak in the validation and test sets. The CS data set is not involved because this data set is not used in the DR-GCN study. The results in Table II show that our methods are better than DR-GCN on Cora, Citeseer, and Pubmed data sets. The accuracies of DR-GCN are directly from [12].

Results on the transductive setting without shuffle. For the same reason for why W2 and W3 are not used in the transductive setting, GML is also not utilized in this experiment. The accuracy, G-mean, and macro F1-score are used as the evaluation metrics. The results are shown in Tables III and IV.

The following observations can be obtained: (1) Overall, GRS, GRW, and GBBN performed better as compared to the

TABLE II: Accuracy of the competing methods.

Data set	Cora	Citeseer	Pubmed
DR-GCN [12]	0.741	0.677	0.817
GRS(GCN, W1)	0.745	0.685	0.820
GRW(GCN, W1)	**0.780**	0.657	0.817
GBBN(GCN, W1)	0.756	**0.703**	**0.827**

[5]The rest three data sets Reddit [16], Amazon-Computer [16], and Amazon-Photo [16] are relatively large. A subgraph sampling-based strategy (e.g., GraphSAINT [16]) should be employed. We leave it as our future work.
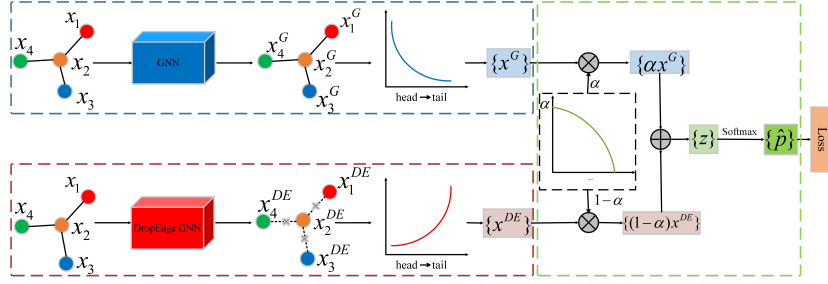
Fig. 7: The network structure of GBBN.

TABLE III: Accuracy and G-mean on transductive setting without shuffle.

Data set	Cora		Citeseer		Pubmed		CS	
	Acc	G-mean	Acc	G-mean	Acc	G-mean	Acc	G-mean
GCN	0.849	0.903	0.748	0.814	0.841	0.878	0.918	0.938
GRS(GCN,W1)	0.854	0.905	0.744	0.812	0.840	0.878	0.917	0.936
GRW(GCN,W1)	0.845	0.901	0.748	0.819	0.828	0.871	0.882	0.928
GBBN(GCN,W1)	0.839	0.898	0.737	0.812	0.831	0.891	0.929	0.939
SGC	0.847	0.894	0.736	0.802	0.816	0.858	0.881	0.880
GRS(SGC,W1)	0.848	0.894	0.739	0.804	0.816	0.858	0.904	0.924
GRW(SGC,W1)	0.713	0.864	0.706	0.795	0.775	0.834	0.111	0.522
GBBN(SGC,W1)	0.796	0.860	0.736	0.802	0.842	0.884	0.906	0.914
GAT	0.855	0.904	0.748	0.810	0.846	0.881	0.914	0.934
GRS(GAT,W1)	0.849	0.905	0.743	0.810	0.845	0.880	0.914	0.936
GRW(GAT,W1)	**0.857**	**0.907**	**0.754**	**0.824**	**0.858**	**0.892**	0.894	**0.940**
GBBN(GAT,W1)	0.835	0.890	0.735	0.804	0.855	0.875	**0.930**	0.934

original models. (2) GRW performed better than GRS and GBBN overall. (3) The effectiveness of these three methods indicates that our label-based weighting strategy (W1) is effective.

Results on the inductive setting with shuffle. In this part, we compare the effect of the model considering whether the LDI is used in the inductive setting. Accordingly, all the weighting strategies (W1, W2, and W3) are leveraged. The macro F1-score and accuracy/G-mean based on SGC with shuffle are shown in Tables V and VI, respectively.

The following observations can be obtained: (1) The overall performances of our three methods (i.e., GRS, GRW, and GBBN) are better than the conventional model SGC. (2) When with shuffle, GBBN achieved the best performance. GML did not improve the performance of its base models. The reason is discussed in the next subsection. (3) Among the three weighting strategies, the combination of label and LDI weighting strategy (i.e., W3) performed best overall, indicating that the index LDI does contain useful cues for training.

C. Discussion.

The better performances of our three methods (GML is poor) over existing methods indicate the importance of considering the distribution characteristics of the imbalance. The combination of label and LDI weighting strategy W3 is better in most cases under the inductive setting, indicating that LDI contains meaningful training cues. Our future research will study the significance of this index in terms of larger sizes of the neighborhood (only 1-neighborhood is considered in this study) and more layers.

Further, we discuss why GML failed in the experiment. The principle of metric learning is to decrease the distance between samples of the same category and increase the distance between samples of different categories. However, in the graph data, edge connections exist among samples. The network in training exchanges information among the feature representations of connected samples during training, which will cause metric learning to reduce the distance between an anchor and a positive sample. However, it is also likely to reduce the distance between the anchor and the heterophily samples in the positive sample's neighbors. The feature coupling caused by data connection causes the metric learning strategy proposed in this study to be unsuitable for graph data sets. As far as we know, there is currently no research on metric learning specifically for GNNs, and exploring more sophisticated methods in the future would be a worthwhile step.

VI. CONCLUSION

This study focuses on the imbalanced distribution in learning with GNNs for graph data. Most benchmark data sets used in GNN studies exhibit an imbalanced distribution over categories. A node-level index called LDI is defined to establish the connection between the imbalance and misclassification. Initial qualitative and quantitative analyses between LDI and GNNs performances are conducted to reveal that samples with high LDI values are more likely to be misclassified when layers increase. A new loss and four new methods (i.e. iFL, GRS, GRW, GML, and GBBN) are proposed based on LDI. Comparative experiment results show that the proposed methods are better than DR-GCN. Overall, the proposed GRW with iFL performs better than the other methods when without shuffle on the transductive setting. The proposed GBBN with iFL and LDI performs better than the other methods when with shuffle on the inductive setting. The effectiveness of W3

TABLE IV: Macro F1-score on transductive setting without shuffle.

Dataset	Cora	Citeseer	Pubmed	CS
GCN	0.836	0.696	0.835	0.895
GRS(GCN,W1)	0.841	0.692	0.835	0.893
GRW(GCN,W1)	0.833	0.706	0.824	0.858
GBBN(GCN,W1)	0.827	0.694	0.827	0.901
SGC	0.834	0.672	0.811	0.813
GRS(SGC,W1)	0.835	0.676	0.811	0.874
GRW(SGC,W1)	0.724	0.673	0.774	0.158
GBBN(SGC,W1)	0.771	0.673	0.841	0.867
GAT	0.843	0.683	0.840	0.889
GRS(GAT,W1)	0.837	0.685	0.839	0.892
GRW(GAT,W1)	**0.845**	**0.714**	**0.854**	0.871
GBBN(GAT,W1)	0.820	0.678	0.852	**0.902**

TABLE V: Macro F1-score on the inductive setting based on SGC with shuffle.

Data set	Cora	Citeseer	Pubmed	CS	Flickr
SGC	0.329	0.439	0.606	0.164	0.124
GML	0.330	0.436	0.607	0.166	0.124
GRS(W1)	0.439	0.495	0.631	0.327	0.140
GRS(W2)	0.321	0.478	0.608	0.154	0.126
GRS(W3)	0.447	0.498	0.633	0.358	0.138
GRW(W1)	0.458	0.508	0.640	0.403	0.118
GRW(W2)	0.183	0.202	0.499	0.072	0.091
GRW(W3)	0.455	0.459	0.642	0.382	0.103
GBBN(W1)	0.637	0.639	0.837	0.849	0.140
GBBN(W2)	0.631	0.626	0.836	0.811	0.142
GBBN(W3)	**0.662**	**0.644**	**0.838**	**0.853**	**0.197**

TABLE VI: Accuracy and G-mean on the inductive setting based on SGC with shuffle.

Data set	Cora		Citeseer		Pubmed		CS		Flickr	
	Acc	G-mean	Acc	G-mean	Acc	G-mean	Acc	G-mean	Acc	G-mean
SGC	0.450	0.529	0.504	0.633	0.642	0.692	0.346	0.386	0.440	0.372
GML	0.449	0.530	0.503	0.633	0.643	0.693	0.347	0.387	0.440	0.372
GRS(W1)	0.482	0.643	0.529	0.670	0.648	0.710	0.469	0.530	0.338	0.385
GRS(W2)	0.454	0.528	0.527	0.658	0.643	0.694	0.334	0.375	0.440	0.373
GRS(W3)	0.502	0.635	0.527	0.670	0.650	0.711	0.490	0.568	0.329	**0.417**
GRW(W1)	0.518	0.641	0.542	0.679	0.652	0.716	0.508	0.598	0.353	0.369
GRW(W2)	0.367	0.448	0.332	0.488	0.598	0.637	0.291	0.336	0.424	0.353
GRW(W3)	0.493	0.653	0.468	0.645	0.652	0.719	0.462	0.611	0.178	0.400
GBBN(W1)	0.666	0.762	0.675	0.770	0.835	0.878	0.894	0.905	0.448	0.384
GBBN(W2)	0.675	0.757	0.669	0.761	0.835	0.876	0.882	0.881	0.441	0.380
GBBN(W3)	**0.687**	**0.784**	**0.683**	**0.774**	**0.837**	**0.879**	**0.896**	**0.908**	**0.456**	0.389

indicates that the index LDI does contain useful cues for training.

VII. ACKNOWLEDGMENT

REFERENCES

[1] Y. Cui, M. Jia, T. Lin, Y. Song and S. Belongie, "Class-Balanced Loss Based on Effective Number of Samples," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9268–9277, 2019.

[2] Q. Dong, S. Gong and X. Zhu, "Imbalanced Deep Learning by Minority Class Incremental Rectification," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 6, pp. 1367–1381, 2019.

[3] G. V. Horn and P. Perona, "The Devil is in the Tails: Fine-grained Classification in the Wild," arXiv:1709.01450, pp. 1–22, 2017.

[4] Y. Hou et al., "Measuring and Improving the Use of Graph Information in Graph Neural Networks," International Conference on Learning Representations (ICLR), pp. 1–16, 2020.

[5] Bingyi Kang et al., "Decoupling Representation and Classifier for Long-Tailed Recognition," International Conference on Learning Representations (ICLR), pp. 1–16, 2020.

[6] T. N. Kipf and M. Welling, "Semi-Supervised Classification with Graph Convolutional Networks," International Conference on Learning Representations (ICLR), pp. 1–16, 2017.

[7] T. Lin, P. Goyal, R. Girshick, K. He and P. Dollar, "Focal loss for dense object detection." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2980–2988, 2017.

[8] J. Liu, Y. Sun, C. Han, Z. Dou and W. Li, "Deep Representation Learning on Long-Tailed Data: A Learnable Embedding Augmentation Perspective," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2970–2979, 2020.

[9] Z. Liu, Z. Miao, X. Zhan, J. Wang, B. Gong and S. X. Yu, "Large-Scale Long-Tailed Recognition in an Open World," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2537–2546, 2019.

[10] H. Pei, B. Wei, K. C. Chang, Y. Lei and B. Yang, "Geom-gcn: Geometric graph convolutional networks," International Conference on Learning Representations (ICLR), pp. 1–12, 2020.

[11] Y. Rong, W. Huang, T. Xu and J. Huang, "DropEdge: Towards Deep Graph Convolutional Networks on Node Classification," International Conference on Learning Representations (ICLR), pp. 1–18, 2020.

[12] M. Shi, Y. Tang, X. Zhu, D. Wilson and J. Liu, "Multi-Class Imbalanced Graph Convolutional Network Learning," Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI), pp. 2879–2885, 2020.

[13] P. Veličković et al., "Graph Attention Networks," International Conference on Learning Representations (ICLR), pp. 1–12, 2018.

[14] Y. Wang, W. Gan, J. Yang, W. Wu and J. Yan, "Dynamic Curriculum Learning for Imbalanced Data Classification," Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 5017–5026, 2019.

[15] Felix Wu et al., "Simplifying Graph Convolutional Networks," Proceedings of the 36th International Conference on Machine Learning (ICML), pp. 6861–6871, 2019.

[16] H. Zeng, H. Zhou, A. Srivastava, R. Kannan and V. Prasanna, "Graph-SAINT: Graph sampling based inductive learning method," International Conference on Learning Representations (ICLR), pp. 1–19, 2020.

[17] L. Zhao and L. Akoglu, "PairNorm: Tackling Oversmoothing in GNNs." International Conference on Learning Representations (ICLR), pp. 1–17, 2020.

[18] B. Zhou, Q. Cui, X. Wei and Z. Chen, "BBN: Bilateral-Branch Network With Cumulative Learning for Long-Tailed Visual Recognition," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9719–9728, 2020.

[19] Jiong Zhu et al., "Graph Neural Networks with Heterophily," Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), pp. 11168–11176, 2021.

[20] Jiong Zhu et al., "Beyond Homophily in Graph Neural Networks: Current Limitations and Effective Designs," Advances in Neural Information Processing Systems, vol. 33, pp. 7793–7804, 2020.

[21] D. Zong and S. Sun, "GNN-XML: Graph Neural Networks for Extreme Multi-label Text Classification," arXiv:2012.05860, pp. 1–11, 2020.