

# GraphMixup: Improving Class-Imbalanced Node Classification on Graphs by Self-supervised Context Prediction

Lirong Wu<sup>1,2,3</sup>, Haitao Lin<sup>1,2</sup>, Zhangyang Gao<sup>1,2,3</sup>, Cheng Tan<sup>1,2</sup>, Stan.Z.Li<sup>1,2,†</sup>

<sup>1</sup> AI Lab, School of Engineering, Westlake University, Hangzhou 310024, Zhejiang Province, China

<sup>2</sup> Institute of Advanced Technology, Westlake Institute for Advanced Study, Hangzhou 310024, Zhejiang Province, China

<sup>3</sup> Zhejiang University, Hangzhou 310058, Zhejiang Province, China

{wulirong, linhaitao, gaozhangyang, tancheng, stan.zq.li}@westlake.edu.cn

## Abstract

Recent years have witnessed great success in handling node classification tasks with Graph Neural Networks (GNNs). However, most existing GNNs are based on the assumption that node samples for different classes are balanced, while for many real-world graphs, there exists the problem of class imbalance, i.e., some classes may have much fewer samples than others. In this case, directly training a GNN classifier with raw data would under-represent samples from those minority classes and result in sub-optimal performance. This paper presents GraphMixup, a novel mixup-based framework for improving class-imbalanced node classification on graphs. However, directly performing mixup in the input space or embedding space may produce out-of-domain samples due to the extreme sparsity of minority classes; hence we construct semantic relation spaces that allows the *Feature Mixup* to be performed at the semantic level. Moreover, we apply two context-based self-supervised techniques to capture both local and global information in the graph structure and then propose *Edge Mixup* specifically for graph data. Finally, we develop a *Reinforcement Mixup* mechanism to adaptively determine how many samples are to be generated by mixup for those minority classes. Extensive experiments on three real-world datasets show that GraphMixup yields truly encouraging results for class-imbalanced node classification tasks.

## Introduction

Recently, the emerging Graph Neural Networks (GNNs) have demonstrated their powerful capability to handle the task of semi-supervised node classification: inferring unknown node labels by using the graph structure and node features with partially known node labels. Despite all these successes, existing works are mainly based on the assumption that node samples for different classes are roughly balanced. However, in many real-world applications, there exists the serious class-imbalanced problem, i.e., some classes may have significantly fewer samples for training than other classes. For example, the majority of users in a transaction fraud network are benign users, while only a small portion of them are bots. Similarly, topic classification for citation networks also suffers from this problem, as the papers for some topics may be scarce, comparing to those on-trend topics.

The class-imbalanced problems have been well studied in the image domain, and data-level algorithms can be summarized into two groups: down-sampling and over-sampling

(More 2016). The down-sampling methods sample a representative sample set from the majority class to make its size close to the minority class, but this inevitably entails a loss of information. In contrast, the over-sampling methods aim to generate new samples for minority classes, which have been found to be more effective and stable. However, directly applying existing over-sampling strategies to graph data may lead to sub-optimal results due to the non-Euclidean property of graphs. Three key problems for mitigating the class-imbalanced problem on graphs by over-sampling are: (1) *How to generate new nodes and their features for minority classes?* (2) *How to capture the connections between the generated node and the existing nodes in the graph?* (3) *How to determine the upsampling scale for each minority class?*

Mixup (Zhang et al. 2017; Verma et al. 2019) is an effective method to solve *Problem (1)*, which performs feature interpolation for minority classes to generate new samples. However, most existing mixup methods are performed either in the input space or embedding space, which may generate out-of-domain samples, especially for those minority classes due to their extreme sparsity. To alleviate this problem, disentangled semantic spaces are constructed in this paper to allow the *Feature Mixup* to be performed at the semantic level. To solve *Problem (2)*, GraphSMOTE (Zhao, Zhang, and Wang 2021) proposes to train an edge generator through the task of adjacency matrix reconstruction and then applies it to predict the existence of edges between generated nodes and existing nodes. However, MSE-based matrix reconstruction completely ignores local and global structural information, making the edge generator overemphasize the connections between nodes with similar features while neglecting the long-range dependencies between nodes. Therefore, we design two context-based self-supervised tasks to consider both local and global information in the graph structure. Finally, unlike heuristic estimation for *Problem (3)*, we develop a reinforcement mixup mechanism to adaptively determine the upsampling scale for each minority class.

Our main contributions are summarized as follows:

- Disentangled semantic spaces are constructed to perform *Semantic Feature Mixup* at the semantic level.
- Propose *Contextual Edge Mixup* specifically for graphs and apply two context-based self-supervised techniques to consider both local and global structure information.

† Corresponding author. Manuscript is under review.

- Develop a reinforcement mixup mechanism instead of heuristic hyperparameters to adaptively determine the up-sampling ratio for each minority class.
- Extensive experiments on three real-world datasets show that GraphMixup outperforms other leading methods covering the full spectrum of low-to-high imbalance ratios.

## Related Work

**Class-Imbalanced Problem.** The class-imbalanced problem is common in real-world scenarios and has become a popular research topic (Johnson and Khoshgoftaar 2019; Rout, Mishra, and Mallick 2018). The mainstream algorithms can be divided into two categories: algorithm-level and data-level. The algorithm-level methods (Ling and Sheng 2008; Zhou and Liu 2005; Parambath, Usunier, and Grandvalet 2014) seek to directly increase the importance of minority classes with suitable penalty functions. Instead, the data-level methods usually adjust class sizes through down-sampling or over-sampling. In this paper, we mainly focus on solving the class-imbalanced problem for graph data with *oversampling-like algorithms*. The vanilla oversampling is replicating existing samples, which reduces the class imbalance but can lead to over-fitting as no extra information is introduced. SMOTE (Chawla et al. 2002) solves this problem by generating new samples by feature interpolation between samples of minority classes and their nearest neighbors, and many of its variants (Han, Wang, and Mao 2005; Bunkhumpornpat, Sinapiromsaran, and Lursinsap 2009) have been proposed with promising results. However, most previous efforts focused on the image domain, and few attempts have been made on class-imbalanced problems for non-Euclidean graph data. GraphSMOTE (Zhao, Zhang, and Wang 2021) is the first work to consider the problem of node-class imbalance on graphs, but their contribution is only to extend SMOTE to graph settings without making full use of the semantic feature information and local/global structural information embedded in graph data.

**Disentanglement Learning.** The disentanglement aims to decompose an entity, such as a feature vector, into several independent components to better capture semantic information. Most recent works are based on the autoencoder architecture, where the latent features generated by the encoder are constrained to be independent in each dimension. The works of DisenGCN (Ma et al. 2019) and IPGDN (Liu et al. 2020), as pioneering attempts, achieve node-level disentanglement through neighbor routines that divide the neighbors of a node into several mutually exclusive parts. FactorGCN (Yang et al. 2020), on the other hand, performs relation disentanglement by taking into account global topological semantics. The semantic disentanglement method proposed in this paper is similar to FactorGNN in that the disentangled semantic features are learned for each node by considering higher-order semantic relations between nodes.

**Graph Self-Supervised Learning (SSL).** The primary goal of Graph SSL is to learn transferable prior knowledge from abundant unlabeled data with well-designed pretext tasks and then generalize the learned knowledge to downstream tasks. The existing graph SSL methods can be divided into three categories: contrastive, generative, and predictive (Wu

et al. 2021). The contrastive methods contrast the views generated from different augmentation by mutual information maximization. Instead, the generative methods focus on the (intra-data) information embedded in the graph, generally based on pretext tasks such as reconstruction. Moreover, the predictive methods generally self-generate labels by some simple statistical analysis or expert knowledge and then perform prediction-based tasks based on self-generated labels. In this paper, we mainly focus on context-based self-supervised prediction since it takes full account of the contextual information in the graph structure, both local and global, allowing us to better capture connections between generated nodes and existing nodes.

## Methodology

### Problem Statement

Given an input graph  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ , where  $\mathcal{V}$  is the set of nodes with features  $\mathcal{X} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N) \in \mathbb{R}^{N \times F}$  and  $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$  is the set of edges. Each node  $v \in \mathcal{V}$  is associated with an features vector  $x_v \in \mathcal{X}$ , and each edge  $e_{u,v} \in \mathcal{E}$  denotes a connection between node  $u$  and node  $v$ . The graph structure can also be represented by an adjacency matrix  $\mathbf{A} \in [0, 1]^{N \times N}$  with  $A_{u,v} = 1$  if  $e_{u,v} \in \mathcal{E}$  and  $A_{u,v} = 0$  if  $e_{u,v} \notin \mathcal{E}$ . We first define the concepts and notions about node class-imbalance ratio:

**Definition 1** Suppose there are  $m$  classes of nodes  $\mathcal{C} = \{C_1, \dots, C_m\}$  in the graph  $\mathcal{G}$ , where  $|C_i|$  is the number of samples belong to  $i$ -th class. *Class-Imbalance Ratio*  $h = \frac{\min_i(|C_i|)}{\max_i(|C_i|)}$  is the ratio of the size of the largest majority class to the smallest minority class in the graph  $\mathcal{G}$ .

Node classification is a typical node-level task where only a subset of node  $\mathcal{V}_L$  with corresponding features  $\mathcal{X}_L$  and labels  $\mathcal{Y}_L$  are known, and we denote the labeled set as  $\mathcal{D}_L = (\mathcal{V}_L, \mathcal{X}_L, \mathcal{Y}_L)$  and unlabeled set as  $\mathcal{D}_U = (\mathcal{V}_U, \mathcal{X}_U, \mathcal{Y}_U)$ . The purpose of GraphMixup is to perform feature, label and edge mixups for minority classes  $\mathcal{C}_S \subseteq \mathcal{C}$  to generate a synthetic set  $\mathcal{D}_S = (\mathcal{V}_S, \mathcal{X}_S, \mathcal{Y}_S)$  and its corresponding edge set  $\mathcal{E}_S = \{e_{v',u} | v' \in \mathcal{V}_S, u \in \mathcal{V}\}$ . Then the synthesized set  $\mathcal{D}_S$  is moved into the labeled set  $\mathcal{D}_L$  to obtain an updated labeled set  $\mathcal{D}_N = \mathcal{D}_L \cup \mathcal{D}_S$ . Similarly, we can obtain an updated edge set  $\mathcal{E}_N = \mathcal{E}_L \cup \mathcal{E}_S$  as well as its corresponding adjacency matrix  $\mathbf{A}_N$ , where  $\mathbf{A}_N[N, : N] = \mathbf{A}$ . Let  $\Phi : \mathcal{V} \rightarrow \mathcal{Y}$  be a graph network trained on labeled data  $\mathcal{D}_N$  so that it can be used to infer the labels  $\mathcal{Y}_U$  of unlabeled data.

In this paper, we present the details of the proposed GraphMixup framework, with an overview shown in Fig. 1. The main idea of GraphMixup is to perform feature mixup to generate synthetic minority nodes in disentangled semantic spaces by a *Semantic Feature Mixup* module. Next, two context-based self-supervised pretext tasks are applied to train a *Contextual Edge Mixup* module that captures both local and global connections between generated nodes and existing for synthetic edge generation. Finally, we detail the *Reinforcement Mixup* mechanism, which can adaptively determine the number of samples to be generated (upsampling scale) by mixup for minority classes.

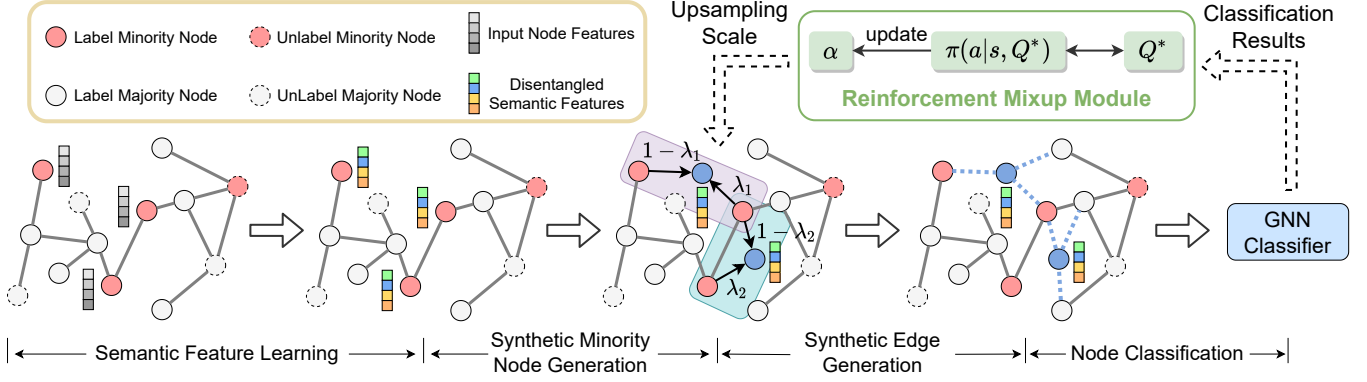


Figure 1: Illustration of the GraphMixp framework, which consists of the following steps: (1) learning disentangled semantic features by constructing semantic relation spaces; (2) generate synthetic minority nodes by semantic-level feature mixup; (3) generate synthetic edges by performing edge mixup with an edge predictor trained on two well-designed context-based self-supervised tasks; (4) Classify using a GNN classifier and feed the results back to the RL agent to update the upsampling scale.

### Semantic Feature Mixup

One effective way to generate minority nodes is to apply feature mixup directly in the input space or embedding space. However, this may lead to sub-optimal results since samples of minority classes are usually quite scarce, resulting in a sparse distribution of samples in the input and embedding space, which in turn produces out-of-domain samples during the interpolation process. Therefore, we consider higher-order relations between samples to learn disentangled semantic features through a *semantic feature extractor*, and thus perform semantic-level feature mixup. To this end, we first construct several semantic relation spaces, represented by semantic relation graphs. Then, we perform feature aggregation and transformation in each semantic space separately, and finally merge the semantic features from each space into a concatenated disentangled semantic feature.

**Semantic Relation Learning.** Specifically, we first transform the input nodes to a low-dimensional space, done by multiplying the features of nodes with a parameter matrix  $\mathbf{W}_h \in \mathbb{R}^{F_h \times F}$ , that is  $\mathbf{h}'_i = \mathbf{W}_h \mathbf{x}_i$ . The transformed features are then used to generate a semantic relation graph with respect to semantic relation  $k$  ( $1 \leq k \leq K$ ) as follows

$$G_{k,i,j} = \sigma(\Omega_k(\mathbf{h}'_i, \mathbf{h}'_j)) \quad (1)$$

where  $\sigma = \tanh(\cdot)$  is an activation function, and  $\Omega_k(\cdot)$  is a function that takes the concatenated features of node  $i$  and node  $j$  as input and takes the form of an one-layer MLP in our implementation. However, without any other constraints, some of the generated relation graphs may contain similar structures. More importantly, it is not easy to directly maximize the gap between various semantic relation graphs due to the non-Euclidean property of graph structure. Therefore, we first derive a graph descriptor  $\mathbf{d}_k$  for each relation graph  $G_k$ ,

$$\mathbf{d}_k = f(\text{Readout}(\mathcal{A}(G_k, \mathbf{H}')))) \quad (2)$$

where  $\mathcal{A}(\cdot)$  is a two-layer graph autoencoder (Kipf and Welling 2016b) which takes  $\mathbf{H}' = \{\mathbf{h}'_1, \mathbf{h}'_2, \dots, \mathbf{h}'_N\}$  as inputs, and generates new features for each node,  $\text{Readout}(\cdot)$

performs global average pooling for all nodes, and  $f(\cdot)$  is a fully connected layer. Note that all semantic relation graphs share the same node features  $\mathbf{H}'$ , making sure that the information discovered by the feature extractor comes only from the differences between graph structures rather than node features. The loss used to train the extractor is defined as

$$\mathcal{L}_{dis} = \sum_{i=1}^{K-1} \sum_{j=i+1}^K \frac{\mathbf{d}_i \cdot \mathbf{d}_j^T}{\|\mathbf{d}_i\| \|\mathbf{d}_j\|} \quad (3)$$

**Disentangled Semantic Feature Learning.** Once the semantic relation learning is completed, the disentangled semantic-specific features can be learned by taking the weighted sum of its neighbors for  $l$ -th ( $1 \leq l \leq L$ ) layer,

$$\mathbf{h}_{i,k}^{(l)} = \sigma\left(\sum_{j \in \mathcal{N}_{i,k}} G_{k,i,j} \mathbf{W}^{(l,k)} \mathbf{h}_j^{(l-1)}\right) \quad (4)$$

where  $\mathbf{h}_j^{(0)} = \mathbf{x}_j$  and  $\mathbf{h}_{i,k}^{(l)}$  represents the semantic feature of node  $i$  with respect to relation  $k$  in  $l$ -th layer. In the semantic relation graph  $G_k$ ,  $\mathcal{N}_{i,k}$  is the neighbours of node  $i$ ,  $G_{k,i,j}$  is the weighting coefficient from node  $i$  to node  $j$ , and  $\mathbf{W}^{(l,k)} \in \mathbb{R}^{F_h \times F_h}$  is a parameter matrix. Finally, the learned features from different semantic relation space can be merged to produce disentangled node features, as follows

$$\mathbf{h}_i^{(l)} = \parallel_{k=1}^K \mathbf{h}_{i,k}^{(l)} \quad (5)$$

**Synthetic Minority Node Generation.** After obtaining the disentangled semantic features for each node by semantic feature extractor, we can perform semantic-level feature mixup to generate new samples for minority classes. Specifically, we perform interpolation on sample  $v$  from one target minority class with its nearest neighbor  $nn(v)$ , as follows

$$\begin{aligned} \mathbf{h}_{v'}^{(L)} &= (1 - \delta) \cdot \mathbf{h}_v^{(L)} + \delta \cdot \mathbf{h}_{nn(v)}^{(L)} \\ nn(v) &= \underset{u \in \{\mathcal{V}/v\}, y_u = y_v}{\text{argmin}} \left\| \mathbf{h}_u^{(L)} - \mathbf{h}_v^{(L)} \right\| \end{aligned} \quad (6)$$

where  $\delta$  is a random variable, following uniform distribution in the range  $[0, 1]$ . Since node  $v$  and  $nn(v)$  belong to the

same class and are very close to each other, the generated node  $v'$  should also belong to the same class. In this way, the *label mixup* can be simplified to directly assign the same label as the source node  $v$  to the newly synthesized node  $v'$ .

### Contextual Edge Mixup

Now we have generated synthetic node  $\mathcal{V}_S$ , node feature  $\mathcal{X}_S$ , and label  $\mathcal{Y}_S$  by means of feature mixup and label mixup described above. However, these new synthetic nodes are still isolated from the raw graph  $\mathcal{G}$  and do not have any links with the nodes in the raw node set  $\mathcal{V}$ . Therefore, we introduce *edge mixup* to capture the connections between generated nodes and existing nodes. To this end, we design an edge prediction that is trained on the raw node set  $\mathcal{V}$  and edge set  $\mathcal{E}$  and then used to predict relation connectivity between generated nodes in the set  $\mathcal{V}_S$  and existing nodes in the set  $\mathcal{V}$ . Specifically, we implement the edge predictor as:

$$\widehat{\mathbf{A}}_{v,u} = \sigma(\mathbf{z}_v \cdot \mathbf{z}_u^T); \mathbf{z}_u = \overline{\mathbf{W}}\mathbf{h}_u^{(L)}, \mathbf{z}_v = \overline{\mathbf{W}}\mathbf{h}_v^{(L)} \quad (7)$$

where  $\widehat{\mathbf{A}}_{v,u}$  refers to the predicted relation connectivity between node  $v$  and  $u$ , and  $\overline{\mathbf{W}} \in \mathbb{R}^{F_h \times F_h}$  is the parameter matrix. The loss function for training the edge predictor is

$$\mathcal{L}_{rec} = \|\widehat{\mathbf{A}} - \mathbf{A}\|_F^2 \quad (8)$$

Since the above MSE-based matrix reconstruction only considers the connectivity between nodes based on feature similarity, it may ignore important information of the graph structure, so we employ two additional context-based self-supervised prediction tasks to capture both local and global structural information for a better edge predictor.

**Context-based Self-supervised Prediction.** The first pretext task *Local-Path Prediction* is to predict the shortest path length between different node pairs. To prevent very noisy ultra-long pairwise distances from dominating the optimization, we truncate the shortest path longer than 4, which also *forces the model to focus on the local structure*. Specifically, it first randomly samples a certain amount of node pairs  $\mathcal{S}$  from all node pairs  $\{(v,u)|v,u \in \mathcal{V}\}$  and calculates the pairwise node shortest path length  $d_{v,u} = d(v,u)$  for each node pair  $(v,u) \in \mathcal{S}$ . Furthermore, it groups the shortest path lengths into four categories:  $C_{v,u} = 0, C_{v,u} = 1, C_{v,u} = 2$ , and  $C_{v,u} = 3$  corresponding to  $d_{v,u} = 1, d_{v,u} = 2, d_{v,u} = 3$ , and  $d_{v,u} \geq 3$ , respectively. The learning objective is then formulated as a multi-class classification problem, as follows

$$\mathcal{L}_{local} = \frac{1}{|\mathcal{S}|} \sum_{(v,u) \in \mathcal{S}} \ell\left(f_\omega^{(1)}(|\mathbf{z}_v - \mathbf{z}_u|), C_{v,u}\right) \quad (9)$$

where  $\ell(\cdot)$  denotes the cross-entropy loss and  $f_\omega^{(1)}(\cdot)$  linearly maps the input to a 4-dimension value.

The second pretext task *Global-Path Prediction* pre-obtains a set of clusters from raw node set  $\mathcal{V}$  and then guides the model to *preserve global topology information* by predicting the shortest path from each node to the anchor nodes associated with cluster centers. Specifically, it first partitions the graph into  $T$  clusters  $\{M_1, M_2, \dots, M_T\}$  by applying unsupervised graph partition algorithm (Karypis and Kumar 1998). Inside each cluster  $M_t$  ( $1 \leq t \leq T$ ), the node with

the highest degree is taken as corresponding cluster center, denoted as  $m_t$ . Then it calculates the distance  $\mathbf{l}_i \in \mathbb{R}^T$  from node  $v_i$  to cluster centers  $\{m_k\}_{k=1}^T$ . The learning objective is then formulated as a regression problem, defined as

$$\mathcal{L}_{global} = \frac{1}{|\mathcal{V}|} \sum_{v_i \in \mathcal{V}} \left\| f_\omega^{(2)}(\mathbf{z}_i) - \mathbf{l}_i \right\|^2 \quad (10)$$

where  $f_\omega^{(2)}(\cdot)$  linearly maps the input to  $K$ -dimension values. The total loss to train the edge predictor is defined as

$$\mathcal{L}_{edge} = \mathcal{L}_{rec} + \mathcal{L}_{local} + \mathcal{L}_{global} \quad (11)$$

Context-based self-supervised methods have been proposed in other work (Jin et al. 2020; Peng et al. 2020) as auxiliary tasks to help feature extraction. However, we apply self-supervised tasks for learning a better edge predictor rather than for learning transferable knowledge on unlabeled data. More importantly, the two self-supervised tasks described above capture both local and global information in the graph structure, which makes them *more beneficial for edge prediction as opposed to the task of feature extraction*.

**Synthetic Edge Generation.** With the learned edge predictor, we can perform *Edge Mixup* in two different ways. The first scheme is to directly use *continuous edges*, that is

$$\mathbf{A}_N[v', u] = \widehat{\mathbf{A}}_{v',u} \quad (12)$$

where  $v' \in \mathcal{V}_S$  and  $u \in \mathcal{V}$ . The second scheme is to obtain the *binary edges* by setting a threshold value, as follows

$$\mathbf{A}_N[v', u] = \begin{cases} 1, & \text{if } \widehat{\mathbf{A}}_{v',u} > \eta \\ 0, & \text{otherwise} \end{cases} \quad (13)$$

The above two strategies are both implemented in this paper denoted as GraphMixup<sub>C</sub> and GraphMixup<sub>B</sub> respectively, and their performance are compared in the experiment part.

### Reinforcement Mixup Mechanism

The upsampling scale, i.e., the number of synthetic samples to be generated by mixup, is important for model performance. A too large scale may introduce redundant and noisy information, while a too small scale is not efficient enough to alleviate the class-imbalanced problem. Therefore, instead of setting the upsampling scale  $\alpha$  as a fixed hyperparameter for all minority classes and then estimating it heuristically, we use a novel reinforcement learning algorithm that adaptively updates the upsampling scale for each minority class. We model the updating process as a Markov Decision Process (MDP) (White III and White 1989). Formally, the state, action, transition, reward, and termination are defined as:

- **State.** For minority class set  $\mathcal{C}_S$ , the state  $s_e$  at epoch  $e$  is represented by the number of new samples for each minority class, that is  $s_e = \{|C_i| \cdot \alpha_i\}_{C_i \in \mathcal{C}_S}$ , where  $\alpha_i = \alpha_i^{init} + \kappa_i$ .
- **Action.** RL agent updates  $\{\kappa_i\}_{C_i \in \mathcal{C}_S}$  by taking action  $a_e$  based on reward. We define the action  $a_e$  as add or minus a fixed value  $\Delta\kappa$  from  $\{\kappa_i\}_{C_i \in \mathcal{C}_S}$  at each epoch  $e$ .
- **Transition.** We generate  $|C_i| \cdot \alpha_i$  new synthetic nodes as defined in Eq. (6) for each minority class in the next epoch.
- **Reward.** Due to the black-box nature of GNN, it is hard to sense its state and cumulative reward. So we define a discrete reward function  $\text{reward}(s_e, a_e)$  for each action  $a_e$  at state  $s_e$  directly based on the classification results, as follows

$$\text{reward}(s_e, a_e) = \begin{cases} +1, & \text{if } cla_e > cla_{e-1} \\ 0, & \text{if } cla_e = cla_{e-1} \\ -1, & \text{if } cla_e < cla_{e-1} \end{cases} \quad (14)$$

where  $cla_e$  is the macro-F1 score at epoch  $e$ . Eq. (14) indicates that if the macro-F1 with action  $a_e$  is higher than the previous epoch, the reward for  $a_e$  is positive, and vice versa.

- **Termination.** If the change of  $\{\kappa_i\}_{C_i \in \mathcal{C}_S}$  among twenty consecutive epochs is no more than  $\Delta\kappa$ , the RL algorithm will stop, and  $\{\kappa_i\}_{C_i \in \mathcal{C}_S}$  will remain fixed during the next training process. The terminal condition is formulated as:

$$\text{Range}(\{\kappa_i^{e-20}, \dots, \kappa_i^e\}) \leq T_\kappa, \quad C_i \in \mathcal{C}_S \quad (15)$$

The  $Q$ -learning (Watkins and Dayan 1992) is applied to learn the above MDP.  $Q$ -learning is an off-policy reinforcement learning algorithm that seeks to find best actions given the current state. It fits the Bellman optimality equation,

$$Q^*(s_e, a_e) = \text{reward}(s_e, a_e) + \gamma \arg \max_{a'} Q^*(s_{e+1}, a') \quad (16)$$

where  $\gamma \in [0, 1]$  is a discount factor of future reward. We adopt a  $\varepsilon$ -greedy policy with an explore probability  $\varepsilon$ :

$$\pi(a_e | s_e; Q^*) = \begin{cases} \text{random action} & \text{w.p. } \varepsilon \\ \arg \max_{a_e} Q^*(s_e, a) & \text{otherwise} \end{cases} \quad (17)$$

This means that the RL agent explores new states by selecting an action at random with probability  $\varepsilon$  instead of only selecting actions based on the max future reward. The RL agent and other modules can be trained jointly in an end-to-end manner. The results in the experiment part verify the effectiveness of the reinforcement mixup mechanism.

### Optimization Objective & Training Strategy

Let  $\mathbf{P}$  be a new embedding matrix by concatenating the semantic embedding  $\mathbf{H}^{(L)}$  of real nodes  $\mathcal{V}$  with the semantic embedding  $\mathbf{H}_S^{(L)}$  of the synthetic nodes  $\mathcal{V}_S$ . Then we can obtain label prediction for node  $v$  with a *node classifier*,

$$\mathbf{h}_v^{(L+1)} = \sigma(\widetilde{\mathbf{W}}^{(1)} \cdot \text{CONCAT}(\mathbf{h}_v^{(L)}, \mathbf{P} \cdot \widehat{\mathbf{A}}[:, v])) \quad (18)$$

$$\widehat{\mathbf{y}}_v = \text{softmax}(\widetilde{\mathbf{W}}^{(2)} \cdot \mathbf{h}_v^{(L+1)})$$

where  $\widetilde{\mathbf{W}}^{(1)} \in \mathbb{R}^{F_h \times F_h}$  and  $\widetilde{\mathbf{W}}^{(2)} \in \mathbb{R}^{m \times F_h}$  are parameter matrices. The above node classifier is optimized using cross-entropy loss on the updated labeled set  $\mathcal{V}_N = \mathcal{V} \cup \mathcal{V}_S$  as:

$$\mathcal{L}_{\text{node}} = \sum_{v \in \mathcal{V}_N} \sum_c (\mathbb{1}(y_v = c) \cdot \log(\widehat{\mathbf{y}}_v[c])) \quad (19)$$

As the model performance is dependent on the quality of embedding space and generated edges, to make training phrase more stable, we adopt a two-stage training paradigm. Let  $\theta, \gamma, \phi$  be the parameters for semantic feature extractor, edge predictor, and node classifier respectively. Firstly, the semantic feature extractor and edge predictor are pre-trained with loss  $\mathcal{L}_{dis}$  and  $\mathcal{L}_{edge}$ , then the pre-trained parameters  $\theta_{init}$  and  $\gamma_{init}$  are used as the initialization. At the fine-tuning stage, the pre-trained encoder  $\theta_{init}(\cdot)$  with a node classifier is trained under the supervision of  $\mathcal{L}_{node}$ . The learning objective is defined as

$$\theta^*, \phi^* = \arg \min_{(\theta, \phi)} \mathcal{L}_{node}(\theta, \gamma, \phi) \quad (20)$$

with initialization  $\theta_{init}, \gamma_{init} = \arg \min_{(\theta, \gamma)} \mathcal{L}_{dis}(\theta) + \beta \mathcal{L}_{edge}(\gamma)$ , where  $\beta$  is the weight to balance these two losses. Since  $\mathcal{L}_{dis}$  and  $\mathcal{L}_{edge}$  are roughly on the same order of magnitude, without loss of generality we set  $\beta$  to 1.0

by default (hyperparametric search for  $\beta$  may yield better results, but this is not the focus of this paper). The pseudo code of the proposed GraphMixup is summarized in Algorithm 1.

---

#### Algorithm 1 Algorithm for the proposed GraphMixup

---

**Input:** Feature Matrix:  $\mathbf{X}$ ; Adjacency Matrix:  $\mathbf{A}$ .

**Output:** Predicted Labels.

- 1: Randomly initialize the semantic feature extractor, edge predictor and node classifier; Initialize upsampling scale  $\alpha_i^{init} = \frac{N}{m|C_i|}$  and  $\kappa_i = 0$  for minority class  $C_i \in \mathcal{C}_S$ ;
  - 2: Train the feature extractor and edge predictor until convergence, based on  $L_{dis}$  and  $L_{edge}$  defined in Eq. 3 and Eq. 11.
  - 3: **while** Not Converged **do**
  - 4:   # *Feature Mixup*
  - 5:   Obtain disentangled features  $\mathbf{H}^{(L)}$  by Eq. 4 and Eq. 5;
  - 6:   **for** class  $i$  in minority classes set  $\mathcal{C}_S$  **do**
  - 7:     Calculate upsampling scale  $\alpha_i = \alpha_i^{init} + \kappa_i$
  - 8:     **for**  $j \in \{0, 1, \dots, |C_i| * \alpha_i\}$  **do**
  - 9:       Generate new samples for class  $i$  by Eq. 6;
  - 10:    **end for**
  - 11:    **end for**
  - 12:    # *Edge Mixup*
  - 13:    Generate new adjacency matrix  $\mathbf{A}_N$  by Eq. 12 or Eq. 13;
  - 14:    Train feature extractor and classifier with  $L_{node}$  by Eq. 19;
  - 15:    # *RL process*
  - 16:    **if** E thenq. 15 is False
  - 17:      $\text{reward}(s_e, a_e) \leftarrow$  Eq. 14;
  - 18:      $a_e \leftarrow$  Eq. 17;
  - 19:      $\kappa_i \leftarrow a_e \cdot \Delta\kappa$  for  $C_i \in \mathcal{C}_S$ ;
  - 20:    **end if**
  - 21: **end while**
  - 22: **return** Predicted labels  $\mathcal{Y}_U$  for unlabeled nodes  $\mathcal{V}_U$ .
- 

## Experiments

In this section, we show the effectiveness of the proposed GraphMixup on three real-world datasets and provide extensive ablation studies and analysis on its various components. The experiments aim to answer the following five questions:

- **Q1.** How does GraphMixup perform in class-imbalance node classification on various real-world datasets?
- **Q2.** Is GraphMixup robust to different imbalance ratios?
- **Q3.** How does semantic feature extractor (bottleneck encoder) influence the performance of GraphMixup?
- **Q4.** How do the two context-based self-supervised prediction tasks influence the performance of GraphMixup?
- **Q5.** How does the reinforcement mixup mechanism work? What happens if the upsampling scale is fixed?

### Experimental setups

**Datasets.** The experiments are conducted on three widely used datasets, namely BlogCatalog (Tang and Liu 2009), Wiki-CS (Mernyei and Cangea 2020), and Cora (Sen et al. 2008) datasets. The first one is BlogCatalog dataset, where 14 classes with fewer than 100 samples are taken as minority classes. The second one is Wiki-CS dataset, where we consider classes with fewer than the average samples per class as minority classes. Finally, on the Cora dataset, we randomly selected three classes as minority classes and the rest as majority classes. All majority classes have a training set of 20 samples. For each minority class, the number is  $20 \times im\_ratio$  with  $im\_ratio$  being 0.5 by default, and we

Table 1: Performance comparison of different methods for class-imbalanced node classification.

Methods	Cora			BlogCatalog			Wiki-CS		
	Acc	AUC-ROC	Macro±F1	Acc	AUC-ROC	Macro±F1	Acc	AUC-ROC	Macro±F1
Origin	0.718±0.002	0.919±0.002	0.715±0.003	0.208±0.005	0.583±0.004	0.067±0.002	0.767±0.001	0.940±0.002	0.735±0.001
Over-Sampling	0.731±0.007	0.927±0.006	0.728±0.008	0.202±0.004	0.592±0.003	0.072±0.003	0.779±0.002	0.948±0.002	0.744±0.002
Re-weight	0.728±0.009	0.925±0.005	0.724±0.006	0.204±0.005	0.785±0.004	0.069±0.002	0.761±0.002	0.939±0.002	0.738±0.002
SMOTE	0.732±0.010	0.925±0.007	0.729±0.005	0.206±0.004	0.795±0.003	0.073±0.001	0.780±0.004	0.945±0.003	0.745±0.003
Embed-SMOTE	0.722±0.006	0.918±0.003	0.721±0.004	0.202±0.006	0.781±0.004	0.070±0.003	0.750±0.005	0.943±0.003	0.721±0.004
GraphSMOTE	0.742±0.003	0.930±0.002	0.739±0.002	0.247±0.004	0.644±0.005	0.123±0.002	0.785±0.003	0.955±0.004	0.752±0.003
GraphMixup <sub>B</sub>	0.761±0.001	0.934±0.002	0.758±0.002	0.255±0.003	0.663±0.003	0.126±0.002	0.792±0.002	0.958±0.002	0.764±0.002
GraphMixup <sub>C</sub>	<b>0.775±0.003</b>	<b>0.942±0.002</b>	<b>0.773±0.001</b>	<b>0.268±0.003</b>	<b>0.673±0.001</b>	<b>0.132±0.002</b>	<b>0.804±0.002</b>	<b>0.964±0.003</b>	<b>0.775±0.001</b>

have varied  $im\_ratio$  to evaluate the performance of GraphMixup under different imbalanced ratios in the following.

**Baselines.** To demonstrate the power of GraphMixup to handle class-imbalance problems, we compare it with six baselines: (1) *Origin*: original implementation without additional tricks; (2) *Over-Sampling*: repeat samples directly from minority classes; (3) *Re-weight*: assign higher loss weights to samples from minority classes (Yuan and Ma 2012); (4) *SMOTE*: generate synthetic samples by interpolating in the input space, and the edges of newly generated nodes are set to be the same as the source nodes; (5) *Embed-SMOTE*: an extension of SMOTE by interpolating in the embedding space (Ando and Huang 2017); (6) *GraphSMOTE*: an extension of Embed-SMOTE by linking generated nodes to existing nodes through a well-trained edge generator. Basing on strategies for setting edges, two variants of GraphMixup are tested: (7) *GraphMixup<sub>B</sub>*: the generated edges are set to binary values by thresholding as Eq. (13); (8) *GraphMixup<sub>C</sub>*: the generated edges are set as continuous values as Eq. (12).

**Evaluation Metrics.** Following existing works in evaluating imbalanced classification, three evaluation metrics are adopted in this paper: Accuracy ( $Acc$ ), AUC-ROC, and Macro-F1.  $Acc$  is calculated on all test samples at once and thus may underestimate those minority classes. In contrast, both AUC-ROC and Macro-F1 are calculated for each class separately and then non-weighted average over them, thus better reflecting the performance on minority classes.

**Hyperparameters.** The following hyperparameters are set for all datasets: Adam optimizer with learning rate  $lr = 0.001$  and weight decay  $decay = 5e-4$ ; Maximum Epoch  $E = 4000$ ; Layer number  $L = 1$  with hidden dimension  $d_F = 32$ ; Semantic Relation  $K = 4$ ; Loss weights  $\alpha = 1.0$ ; Threshold  $\eta = 0.5$ . In the reinforcement mixup module, we set  $\gamma = 1$ ,  $\varepsilon = 0.9$ ,  $\Delta\kappa = 0.05$ . Besides, the initial  $\kappa_i^{init}$  is set class-wise:  $\frac{N}{m|C_i|}$  for minority class  $C_i \in C_S$  on each dataset. Each set of experiments is run 5 times with different random seeds, and the average results are reported as performance metrics.

### Class-Imbalanced Classification (Q1)

To evaluate the effectiveness of GraphMixup in class-imbalanced node classification tasks, we compare it with the other six baselines on three datasets. Table. 1 shows that the improvements brought by GraphMixup are much larger than directly applying other over-sampling algorithms. For example, compared with GraphSMOTE, *GraphMixup<sub>C</sub>* shows an improvement of 3.3% in  $Acc$  score and 3.4% in Macro-F1 score. Moreover, both two variants of GraphSMOTE show significant improvements for imbalanced node classification, compared to almost all baselines on all datasets. Notably, we find that *GraphMixup<sub>C</sub>* exhibits slightly better per-

formance than *GraphMixup<sub>B</sub>*, which implies the advantage of soft continuous edges over thresholded binary edges.

### Influence of Imbalance Ratio (Q2)

The performance under different imbalance ratios is reported in Table. 2 to evaluate their robustness. Experiments are conducted in the Cora dataset by varying class imbalance ratio  $im\_ratio$  as  $\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6\}$ . The ROC-AUC scores in Table. 2 show that: (1) GraphMixup generalizes well to different imbalance ratios and achieves the best performance across all settings. (2) The improvement of GraphMixup is more significant when the imbalance ratio is more extreme. For example, when the imbalance ratio is 0.1, *GraphMixup<sub>C</sub>* outperforms SMOTE by 6.4%, and the gap reduces 1.5% when the imbalance ratio reaches 0.6.

Table 2: Performance under different imbalance ratios.

Methods	Class-Imbalanced Ratio					
	0.1	0.2	0.3	0.4	0.5	0.6
Origin	0.843	0.890	0.907	0.913	0.919	0.920
Over-Sampling	0.830	0.898	0.917	0.922	0.927	0.929
Re-weight	0.869	0.906	0.921	0.923	0.925	0.928
SMOTE	0.839	0.897	0.917	0.924	0.925	0.929
Embed-SMOTE	0.870	0.897	0.906	0.912	0.918	0.925
GraphSMOTE	0.887	0.912	0.923	0.927	0.930	0.932
GraphMixup <sub>B</sub>	0.898	0.915	0.923	0.932	0.934	0.935
GraphMixup <sub>C</sub>	<b>0.903</b>	<b>0.919</b>	<b>0.931</b>	<b>0.935</b>	<b>0.942</b>	<b>0.944</b>

### Influence of Bottleneck Encoder (Q3)

To analyze the effectiveness of the *Semantic Feature Extractor (SEM)* and the applicability of GraphMixup to different bottleneck encoders, we apply three other common encoders: GCN (Kipf and Welling 2016a), SAGE (Hamilton, Ying, and Leskovec 2017), and GAT (Veličković et al. 2017). Due to space limitations, only the performance of the AUC-ROC scores on the Cora dataset is reported. Table. 3 shows that GraphSMOTE works well with all four bottleneck encoders, achieving the best performance. Moreover, results with SEM as the bottleneck encoder are slightly better than the other three across all methods, indicating the benefits of constructing semantic relation spaces, extracting semantic features, and performing semantic-level mixup. Furthermore, Fig. 2 shows the correlation analysis of 128-dimensional latent features with  $K = 4$  semantic relations obtained from four different bottleneck encoders. We find that only the correlation map of SEM exhibits four clear diagonal blocks, which demonstrates its excellent capability to extract highly independent disentangled semantic features.

### RL Process Analysis (Q4)

To verify the importance of the reinforcement mixup mechanism, we remove it from GraphMixup to obtain a new variant - GraphMixup-Fix, which sets a *fixed* upsampling



Table 3: Performance with different bottleneck encoders.

Methods	Bottleneck Encoder			
	GCN	SAGE	GAT	SEM
Origin	0.909	0.897	0.912	<i>0.919</i>
Over-Sampling	0.916	0.907	0.923	<i>0.927</i>
Re-weight	0.917	0.904	0.919	<i>0.925</i>
SMOTE	0.917	0.907	0.919	<i>0.925</i>
Embed-SMOTE	0.914	0.906	0.916	<i>0.918</i>
GraphSMOTE	0.920	0.914	0.923	<i>0.930</i>
GraphMixup <sub>B</sub>	0.924	0.916	0.926	<i>0.934</i>
GraphMixup <sub>C</sub>	<b>0.926</b>	<b>0.919</b>	<b>0.932</b>	<b>0.942</b>

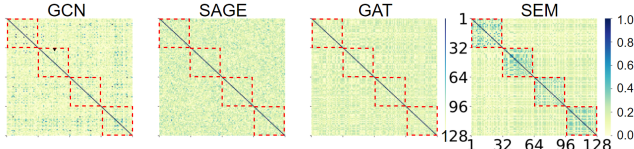


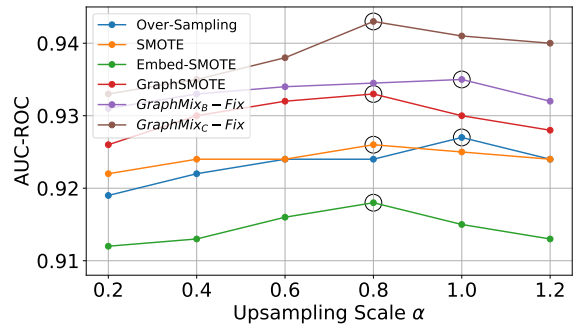
Figure 2: Feature correlation analysis on the Cora dataset.

scale for all minority classes. Then, we plot the performance curve of GraphMixup-Fix and four baselines under different (fixed) upsampling scales on the Cora dataset. As shown in Fig. 3(a), we find that generating more samples for minority classes helps achieve better performance when the upsampling scale is smaller than 0.8 (or 1.0). However, when the upsampling scale becomes larger, keep increasing it may result in the opposite effect, as too many new synthesis nodes will only introduce redundant and noisy information.

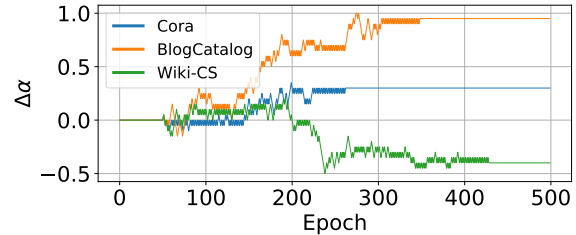
Since the RL algorithm is trained jointly with GNNs, its updating and convergence process is very important. In Fig. 3(b), we visualize the updating process of the cumulative change in upsampling ratio  $\alpha$ , e.g.,  $\Delta\alpha = \alpha_i - \alpha_i^{init}$ . Since other modules in the framework are updated together with the RL module, the RL environment is not very stable at the beginning, so the RL algorithm starts to run only after the first 50 epochs. When the framework gradually converges,  $\Delta\alpha$  bumps for several rounds and meets the terminal condition. From Fig. 3(b), we find that  $\Delta\alpha$  eventually converges to 0.3 on the Cora dataset, resulting in an upsampling scale  $\alpha_i = \Delta\alpha + \alpha_i^{init} = 0.8$  with initial value  $\alpha_i^{init} = \text{round}(\frac{N}{m|C_i|}) = 0.5$ . This corresponds to the result in Fig. 3(a) where GraphMixup<sub>C</sub> obtains the best performance when the upsampling scale is 0.8, which demonstrates the effectiveness of the reinforcement mixup mechanism, i.e., it adaptively determines suitable upsampling scale without the need for heuristic estimation like Fig. 3(a).

### Self-Supervised Prediction Analysis (Q5)

This evaluates the effectiveness of self-supervised prediction tasks in the proposed framework through four sets of experiments: the model without (A) Local-Path Prediction (*w/o LP*); (B) Glocal-Path Prediction (*w/o GP*); (C) both Local-Path and Global-Path Prediction (*w/o LP and GP*), and (D) the full model. Experiments are conducted on the Cora dataset, and ROC-AUC scores are reported as performance evaluation. After analyzing the reported results in Fig. 4, we can find that both Local-Path Prediction and Glocal-Path Prediction contribute to improving model performance. More importantly, applying these two tasks together can further improve performance on top of each of them, result-



(a) Performance under different (fixed) upsampling scale.



(b) Updating process of the cumulative change in  $\kappa$ .

Figure 3: Reinforcement mixup mechanism analysis.

ing in the best performance, which demonstrates the benefit of self-supervised prediction tasks on capturing local and global information embedded in the graph structure.

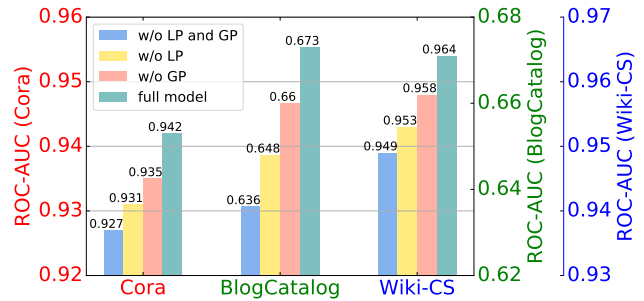


Figure 4: Ablation study with different self-supervised tasks.

## Conclusion

In this paper, we propose GraphMixup, a novel framework for improving class-imbalanced node classification on graphs. GraphMixup implements feature, label, and edge mixup simultaneously in a unified framework in an end-to-end manner. Specifically, GraphMixup performs semantic-level feature mixup by constructing semantic relation spaces and edge mixup with an edge predictor trained on two well-designed context-based self-supervised tasks; Moreover, a *Reinforcement Mixup* mechanism is applied to adaptively determine the number of samples to be generated (upsampling scale) by mixup for minority classes. Extensive experiments on three real-world datasets have shown that the proposed GraphMixup outperforms other leading methods on class-imbalanced node classification tasks.

## References

- Ando, S.; and Huang, C. Y. 2017. Deep over-sampling framework for classifying imbalanced data. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 770–785. Springer.
- Bunkhumpornpat, C.; Sinapiromsaran, K.; and Lursinsap, C. 2009. Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem. In *Pacific-Asia conference on knowledge discovery and data mining*, 475–482. Springer.
- Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; and Kegelmeyer, W. P. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 16: 321–357.
- Hamilton, W.; Ying, Z.; and Leskovec, J. 2017. Inductive representation learning on large graphs. In *Advances in neural information processing systems*, 1024–1034.
- Han, H.; Wang, W.-Y.; and Mao, B.-H. 2005. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In *International conference on intelligent computing*, 878–887. Springer.
- Jin, W.; Derr, T.; Liu, H.; Wang, Y.; Wang, S.; Liu, Z.; and Tang, J. 2020. Self-supervised learning on graphs: Deep insights and new direction. *arXiv preprint arXiv:2006.10141* .
- Johnson, J. M.; and Khoshgoftaar, T. M. 2019. Survey on deep learning with class imbalance. *Journal of Big Data* 6(1): 1–54.
- Karypis, G.; and Kumar, V. 1998. A fast and high quality multilevel scheme for partitioning irregular graphs. *SIAM Journal on scientific Computing* 20(1): 359–392.
- Kipf, T. N.; and Welling, M. 2016a. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* .
- Kipf, T. N.; and Welling, M. 2016b. Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308* .
- Ling, C. X.; and Sheng, V. S. 2008. Cost-sensitive learning and the class imbalance problem. *Encyclopedia of machine learning* 2011: 231–235.
- Liu, Y.; Wang, X.; Wu, S.; and Xiao, Z. 2020. Independence Promoted Graph Disentangled Networks. In *AAAI*, 4916–4923.
- Ma, J.; Cui, P.; Kuang, K.; Wang, X.; and Zhu, W. 2019. Disentangled graph convolutional networks. In *International Conference on Machine Learning*, 4212–4221.
- Mernyei, P.; and Cangea, C. 2020. Wiki-cs: A wikipedia-based benchmark for graph neural networks. *arXiv preprint arXiv:2007.02901* .
- More, A. 2016. Survey of resampling techniques for improving classification performance in unbalanced datasets. *arXiv preprint arXiv:1608.06048* .
- Parambath, S. P.; Usunier, N.; and Grandvalet, Y. 2014. Optimizing F-measures by cost-sensitive classification. In *Advances in Neural Information Processing Systems* 27.
- Peng, Z.; Dong, Y.; Luo, M.; Wu, X.-M.; and Zheng, Q. 2020. Self-supervised graph representation learning via global context prediction. *arXiv preprint arXiv:2003.01604* .
- Rout, N.; Mishra, D.; and Mallick, M. K. 2018. Handling imbalanced data: a survey. In *International Proceedings on Advances in Soft Computing, Intelligent Systems and Applications*, 431–443. Springer.
- Sen, P.; Namata, G.; Bilgic, M.; Getoor, L.; Galligher, B.; and Eliassi-Rad, T. 2008. Collective classification in network data. *AI magazine* 29(3): 93–93.
- Tang, L.; and Liu, H. 2009. Relational learning via latent social dimensions. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 817–826.
- Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; and Bengio, Y. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903* .
- Verma, V.; Lamb, A.; Beckham, C.; Najafi, A.; Mitliagkas, I.; Lopez-Paz, D.; and Bengio, Y. 2019. Manifold mixup: Better representations by interpolating hidden states. In *International Conference on Machine Learning*, 6438–6447. PMLR.
- Watkins, C. J.; and Dayan, P. 1992. Q-learning. *Machine learning* 8(3-4): 279–292.
- White III, C. C.; and White, D. J. 1989. Markov decision processes. *European Journal of Operational Research* 39(1): 1–16.
- Wu, L.; Lin, H.; Gao, Z.; Tan, C.; Li, S.; et al. 2021. Self-supervised on Graphs: Contrastive, Generative, or Predictive. *arXiv preprint arXiv:2105.07342* .
- Yang, Y.; Feng, Z.; Song, M.; and Wang, X. 2020. Factorizable Graph Convolutional Networks. *Advances in Neural Information Processing Systems* 33.
- Yuan, B.; and Ma, X. 2012. Sampling+ reweighting: Boosting the performance of AdaBoost on imbalanced datasets. In *The 2012 international joint conference on neural networks (IJCNN)*, 1–6. IEEE.
- Zhang, H.; Cisse, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2017. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412* .
- Zhao, T.; Zhang, X.; and Wang, S. 2021. GraphSMOTE: Imbalanced Node Classification on Graphs with Graph Neural Networks. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, 833–841.
- Zhou, Z.-H.; and Liu, X.-Y. 2005. Training cost-sensitive neural networks with methods addressing the class imbalance problem. *IEEE Transactions on knowledge and data engineering* 18(1): 63–77.