STRUCTURE-AWARE MIXUP FOR GRAPH NEURAL NETWORKS: ENHANCING GRAPH REPRESENTATION VIA STRUCTURAL SIMILARITY

ZI-YU CUI[1], XIN YUAN[2], LI-XIANG XU[1,3,*], BIN LUO[3], XIN NIU[1], YUAN-YAN TANG[4]

[1]School of Artificial Intelligence and Big Data, Hefei University, Hefei 230601, China
[2]School of Electrical and Electronic Engineering, the University of Adelaide, Adelaide, SA 5005, Australia
[3]School of Artificial Intelligence, Hefei Institute of Technology, Hefei 238076, China
[4]Zhuhai UM Science and Technology Research Institute, FST University of Macau, Macau
*Corresponding Author's E-MAIL: xlxiang@ustc.edu.cn

Abstract:

Graph Neural Networks (GNNs) have achieved remarkable performance in semi-supervised node classification tasks. However, they often suffer from the under-reaching problem, where sparse labeled nodes limit the propagation of supervision signals, negatively affecting representation learning and generalization. To address this issue, we propose a novel Structure-aware Mixup strategy for graph data augmentation, which leverages structural similarity based on node degree and PageRank scores to guide interpolation between node pairs. This approach dynamically generates mixup weights, ensuring augmented samples are both semantically consistent and structurally coherent. Extensive experiments on five benchmark datasets demonstrate that our approach significantly improves classification performance under low-label conditions while achieving a good balance between augmentation quality and model stability. These results highlight the effectiveness and generalizability of the proposed structure-aware mixup strategy in graph data augmentation.

Keywords:

Graph Neural Networks; Semi-supervised Learning; Graph Data Augmentation; Node Classification

1. Introduction

With the rapid emergence of graph-structured data in domains such as social networks, bioinformatics, and recommender systems, Graph Neural Networks (GNNs) [?] have become one of the core approaches for modeling graph data. By aggregating information from neighboring nodes, GNNs can effectively capture the underlying structural and semantic relationships within graphs. They have been widely applied to tasks such as node classification, link prediction, and graph classification. In particular, under semi-supervised node classification settings, GNNs are capable of achieving strong representation and classification performance even with only a small number of labeled nodes.

However, existing GNN models still face several challenges in practical applications. One key issue is the sparse distribution of labeled nodes in the graph. Due to limitations in graph structure and the model's propagation range, many unlabeled nodes struggle to receive effective supervision signals, leading to the under-reaching phenomenon [?]. This issue becomes more pronounced when the model is shallow or the graph is sparsely connected, significantly hindering the model's generalization ability under low-label-rate conditions.

To address the above issues, data augmentation has been introduced into Graph Neural Networks in recent years, aiming to improve model learning capability and robustness by generating new training samples. Among these methods, Mixup-based approaches create virtual samples by interpolating the features and labels of different nodes, and have shown promising results in graph classification tasks. However, applying Mixup to node classification still faces two major challenges: First, traditional Mixup often ignores the structural relationships between nodes in the graph and performs interpolation solely in the feature space, which may lead to structural-semantic conflicts. Second, the augmented samples may lack topological consistency within the graph, making them susceptible to introducing noisy interference [?].

To tackle the above challenges, this paper proposes a Structure-aware Mixup method that leverages graph structural features to guide the data augmentation process, aiming to generate virtual training samples with greater structural coherence and semantic consistency. Specifically, a structure-aware mechanism is introduced during sample generation by combining local node degree and global PageRank scores to compute the structural similarity between node pairs. This similarity is then used to dynamically adjust the mixup weights. Compared to random mixing strategies, the structure-aware Mixup can effectively preserve the relational structure within the graph, thereby improving the quality of augmented samples and enhancing the model's generalization performance.

The main contributions of this work are as follows:

(1) We propose a structure-aware Mixup method that leverages structural similarity between node pairs to guide feature and label interpolation, improving consistency and mitigating the under-reaching issue in GNNs.

(2) A structural similarity function combining node degree and PageRank is designed to adaptively generate mixup weights, balancing semantic coherence with sample diversity.

(3) Extensive evaluations on five benchmark node classification datasets show that our method achieves state-of-the-art or competitive performance across multiple GNN architectures, demonstrating its effectiveness and generalizability.

2. Related Work

Graph Neural Networks (GNNs) have emerged as a core method for processing graph-structured data and are widely applied in domains such as social networks, bioinformatics, and recommendation systems. In addition to message-passing based GNNs, graph kernel techniques have also been explored as an alternative or complementary mechanism for graph representation learning [?, ?]. Extensive research has been conducted on their modeling capacity, label propagation, and augmentation strategies, mainly covering the following areas:

2.1. Challenges in GNNs and Supervision Propagation

Classic GNNs like GCN, GAT [?], and GraphSAGE [?] learn node representations via neighbor aggregation and perform well in semi-supervised classification. However, shallow models have limited receptive fields, while deeper ones suffer from over-smoothing and over-squashing, impairing representation learning for distant nodes.

2.2. Data Augmentation and Mixup Extensions

To improve model robustness, various data augmentation strategies have been introduced into graph learning, including feature perturbation, structural perturbation (e.g., DropEdge), and subgraph sampling. Inspired by Mixup techniques in computer vision, methods like GraphMix [?] have applied interpolation strategies to graph classification tasks. However, most approaches in node classification fail to fully leverage structural information, resulting in augmented samples with poor topological consistency.

2.3. Under-reaching and Cross-node Enhancement

The under-reaching problem refers to unlabeled nodes far from labeled ones being unable to receive sufficient supervision signals. Some approaches address this by introducing implicit skip connections or constructing virtual augmented samples to propagate information and enhance performance. However, these methods often rely on random pairing and lack structural guidance, limiting their effectiveness.

2.4. Consistency Regularization and Multi-granularity Alignment

Consistency learning methods such as DGI [?], GRACE [?], and BGRL [?] aim to maintain stable node representations under perturbations. IMCN [?] further introduces multi-granularity consistency regularization across node-level, class-level, and distribution-level representations, improving generalization. Nonetheless, its augmentation strategy does not incorporate structure-aware mechanisms.

2.5. Structure-aware Augmentation Methods and Challenges

Structure-aware augmentation methods, such as DropEdge and GraphDiffusion [?], aim to preserve essential structural information. However, in Mixup-based augmentation, most approaches still rely on random weights

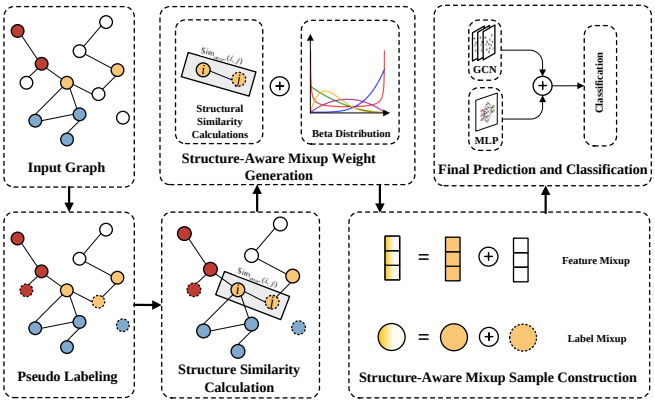

FIGURE 1. Overview of the proposed structure-aware Mixup framework.

or feature space distance, overlooking structural indicators. This paper proposes incorporating structural similarity into the weight computation of Mixup to enhance semantic-topological consistency and training stability of augmented samples.

3. Method

This section presents the design and technical details of our proposed structure-aware Mixup augmentation strategy. Unlike traditional Mixup, we perform interpolation not only in the feature and label spaces but also incorporate graph structural information to guide the generation of Mixup weights. This enables structure-aware virtual sample generation. As illustrated in FIGURE 1, the pipeline includes four modules: pseudo-label generation, structural similarity computation, weight fusion sampling, Mixup sample construction, followed by GNN feature fusion and supervised training.

3.1. Problem Definition and Notations

Given an undirected graph $G = (V, E, X)$, where V denotes the set of nodes with $|V| = N$, $E \subseteq V \times V$ represents the set of edges, and $X \in \mathbb{R}^{N \times d}$ is the node feature matrix where each node has a d-dimensional feature vector. The label matrix is denoted as $Y \in \mathbb{R}^{N \times C}$, where C is the number of classes, and only a small subset of nodes are labeled. Let $D_l \subset V$ be the labeled node set and $D_u = V \setminus D_l$ be the unlabeled node set. This work focuses on semi-supervised node classification, aiming to learn a prediction function $f_\theta : X \to Y$ that generalizes well on the entire graph. To enhance generalization under limited supervision, we propose a structure-aware Mixup strategy that utilizes graph structural information to guide the interpolation between labeled and pseudo-labeled nodes.

3.2. Pseudo Labeling and Node Pair Construction

To extend the supervision signal beyond the limited labeled nodes, we adopt a pseudo-labeling strategy to incorporate high-confidence unlabeled nodes into the Mixup augmentation process. Specifically, we use the current graph neural network model f_θ to perform forward prediction on each unlabeled node $x_j \in D_u$, obtaining a soft label:

$$\hat{y}_j = \text{softmax}(f_\theta(x_j))$$

where $\hat{y}_j \in \mathbb{R}^C$ denotes the predicted probability distribution over all classes. To ensure the reliability of pseudo labels, we introduce a confidence threshold $\gamma \in (0, 1)$ and filter out pseudo-labeled nodes whose maximum predicted probability is below this threshold:

$$D_{pl} = \{(x_j, \hat{y}_j) \mid \max(\hat{y}_j) \geq \gamma\}$$

Then, we sample a labeled node $(x_i, y_i) \in D_l$ and a pseudo-labeled node $(x_j, \hat{y}_j) \in D_{pl}$ to form a candidate node pair. We only retain the pair if their predicted classes are consistent, i.e., $\arg\max(y_i) = \arg\max(\hat{y}_j)$, ensuring class consistency in the interpolation. This pairing strategy guarantees semantic consistency in the augmented samples and mitigates the negative impact of noisy pseudo labels during training.

3.3. Structural Similarity Computation

In graph-structured data, nodes are not only associated with feature attributes but also embedded within complex topological relationships. To improve the structural rationality of the Mixup samples, we introduce a structural similarity measure to quantify the structural consistency between node pairs. We design a similarity function that combines both local connectivity and global importance, by integrating node degree and PageRank. The structural similarity between node i and node j is defined as:

$$\text{Sim}_{struct}(i, j) = w_1 \cdot \text{DegSim}(i, j) + w_2 \cdot \text{PRSim}(i, j)$$

where $w_1 + w_2 = 1$, and $w_1, w_2 \in [0, 1]$ are weighting coefficients balancing the two components. The degree-based similarity is defined as:

$$\text{DegSim}(i, j) = 1 - \frac{|d_i - d_j|}{\max(d_i, d_j, 1)}$$

where d_i denotes the degree of node i, indicating the number of its adjacent edges. This component evaluates how similar two nodes are in terms of local connectivity. The PageRank-based similarity is defined as:

$$\text{PRSim}(i, j) = 1 - \frac{|\text{PR}_i - \text{PR}_j|}{\max(\text{PR}_i, \text{PR}_j, \epsilon)}$$

where PR_i represents the PageRank score of node i, reflecting its global importance in the graph structure, and ϵ is a small positive constant to prevent division by zero. A higher structural similarity indicates that the node pair plays a more similar role in the graph, making them more suitable candidates for Mixup. This structural similarity function guides the sample interpolation process and prevents structural distortion and semantic inconsistency, thereby improving the quality and discriminability of the augmented data.

3.4. Structure-Aware Mixup Weight Generation and Sample Construction

After constructing node pairs and computing their structural similarity, we generate structure-aware Mixup weights to create augmented samples. Unlike traditional Mixup, which samples interpolation coefficients from a Beta distribution without structural context, we fuse structural similarity with Beta randomness to compute the final weight. The final Mixup weight is computed as follows:

$$\lambda_{ij} = \alpha \cdot \text{Sim}_{struct}(i, j) + (1 - \alpha) \cdot \lambda_{\text{rand}}, \quad \lambda_{\text{rand}} \sim \text{Beta}(\alpha, \alpha)$$

where $\text{Sim}_{struct}(i, j)$ is the structural similarity defined in Section 3.3, λ_{rand} is a randomly sampled coefficient from the Beta distribution, and $\alpha \in (0, 1)$ is a hyperparameter that controls the balance between structural guidance and randomness. A higher structural similarity leads to a Mixup weight more biased toward preserving structural consistency, while lower similarity encourages greater interpolation diversity.

Using the generated weight λ_{ij}, we perform linear interpolation in both the feature and label space to construct the augmented sample:

$$\tilde{x}_{ij} = \lambda_{ij} \cdot x_i + (1 - \lambda_{ij}) \cdot x_j$$

$$\tilde{y}_{ij} = \lambda_{ij} \cdot y_i + (1 - \lambda_{ij}) \cdot \hat{y}_j$$

Here, $\tilde{x}_{ij} \in \mathbb{R}^d$ denotes the structure-aware interpolated feature, and $\tilde{y}_{ij} \in \mathbb{R}^C$ is the corresponding soft label. This construction ensures that the augmented samples preserve both semantic smoothness and structural coherence, thus avoiding feature drift and enhancing generalization under limited supervision.

3.5. Final Prediction and Classification

To capture both structural and individual characteristics of augmented samples, we employ a dual-branch encoder integrating a Graph Convolutional Network (GCN) for neighborhood structure and a Multi-Layer Perceptron (MLP) for intrinsic node features. This architecture mitigates over-smoothing and enhances representation diversity and discriminability.

Specifically, the structure-aware augmented node $\tilde{x}_{ij}$ is fed into both the GCN and MLP branches to obtain a structure-dependent representation h_{ij}^{gcn} and an individual representation h_{ij}^{mlp}, respectively. The final representation is formed by a weighted fusion of the two:

$$h_{ij} = \lambda \cdot h_{ij}^{\text{mlp}} + (1 - \lambda) \cdot h_{ij}^{\text{gcn}}$$

where $\lambda \in [0, 1]$ is a tunable coefficient that controls the balance between individual-specific and structure-aware features.

The fused representation h_{ij} is then passed into a classifier for prediction:

$$\hat{y}_{ij}^{\text{pred}} = \text{softmax}(\text{MLP}_{cls}(h_{ij}))$$

To optimize the model, we define a joint loss function that includes three components: supervised loss on labeled samples, Mixup loss on structure-aware augmented samples, and a multi-level consistency loss. The total loss is formulated as:

$$\mathcal{L}_{total} = \mathcal{L}_{sup} + \lambda_1 \cdot \mathcal{L}_{mixup} + \lambda_2 \cdot \mathcal{L}_{cons}$$

Here, $\mathcal{L}_{sup}$ denotes the standard cross-entropy loss computed on labeled nodes, $\mathcal{L}_{mixup}$ supervises the structure-aware interpolated samples, and $\mathcal{L}_{cons}$ enforces multi-granularity consistency, including node-level, class-level,

or distribution-level alignment. The coefficients λ_1 and λ_2 control the relative contribution of each component.

Through dual-branch feature modeling and multi-objective optimization, our framework not only performs structure-aware augmentation in the feature space but also improves learning stability and generalization capability, making it well-suited for semi-supervised node classification under limited label scenarios.

4. Experiments

In this section, we conduct extensive experiments on multiple node classification benchmarks to evaluate the effectiveness of the proposed structure-aware augmentation framework, SAMix-GNN. The experiments cover citation networks, co-purchase networks, and co-authorship graphs to comprehensively examine the performance, robustness, and generalization capability of our method under low-label scenarios.

4.1. Datasets

We evaluate SAMix-GNN on five benchmark datasets: Cora, Citeseer, PubMed, Photo, and CS. These datasets cover diverse graph types—citation, co-purchase, and co-authorship—and vary in node scale, class count, feature dimension, and sparsity, offering challenging settings for testing GNN generalization and robustness. Detailed statistics are shown in TABLE 1.

TABLE 1. Statistics of benchmark datasets

Dataset	Type	Nodes	Edges	Classes	Features
Cora	Citation	2,708	5,429	7	1,433
Citeseer	Citation	3,327	4,732	6	3,703
PubMed	Citation	19,717	44,338	3	500
Photo	Co-purchase	7,650	119,081	8	745
CS	Co-authorship	18,333	81,894	15	6,805

TABLE 2. Classification accuracy (%) comparison on five benchmark datasets

Method	Cora	Citeseer	PubMed	Photo	CS
GCN	81.9 ± 0.4	70.9 ± 0.2	78.8 ± 0.2	91.5 ± 0.6	91.3 ± 0.3
GAT	79.6 ± 0.7	72.2 ± 0.8	78.6 ± 0.5	90.4 ± 0.7	90.1 ± 0.6
SGC	81.1 ± 0.1	72.6 ± 0.0	78.9 ± 0.2	91.7 ± 1.2	92.7 ± 0.0
IMCN	84.6 ± 0.4	76.5 ± 0.3	81.2 ± 0.8	93.1 ± 0.5	93.3 ± 0.2
HCPL	84.2 ± 0.6	74.4 ± 0.7	82.4 ± 0.7	92.3 ± 0.5	93.2 ± 0.2
SMGCL	85.2 ± 0.4	73.8 ± 0.3	81.4 ± 0.3	–	93.2 ± 0.2
RNCGLN	84.5 ± 0.4	74.0 ± 0.3	81.2 ± 0.6	92.2 ± 0.5	93.3 ± 0.3
SAMix-GNN	85.3 ± 0.5	76.6 ± 0.4	82.6 ± 0.5	93.2 ± 0.5	93.4 ± 0.2

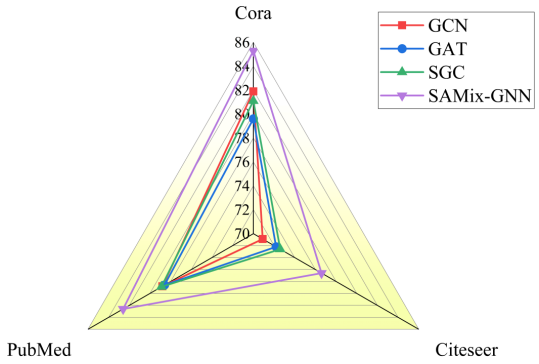

FIGURE 2. Model-wise Accuracy Comparison on Citation Datasets.

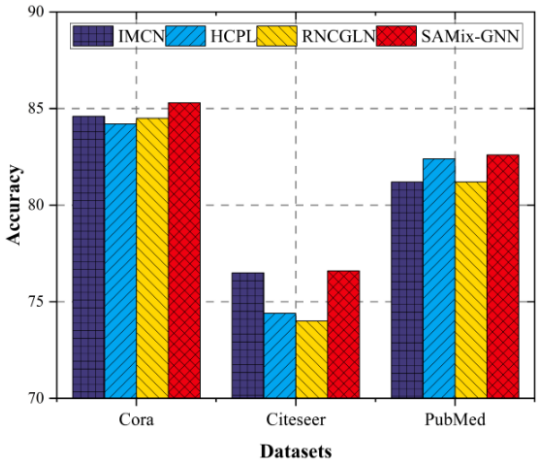

FIGURE 3. Performance of SAMix-GNN vs. Prior Augmentation Methods.

4.2. Comparison with Baselines and Results Analysis

We compare SAMix-GNN with several representative baselines, including standard GNN models such as GCN, GAT, and SGC, as well as recent advanced augmentation or regularization-based methods including IMCN, HCPL [?], SMGCL [?], and RNCGLN [?]. These baselines represent different directions in graph learning, from basic aggregation mechanisms to consistency regularization and mixup-based strategies.

As shown in TABLE 2, SAMix-GNN consistently achieves the best or competitive accuracy across all five datasets. Notably, on Citeseer, PubMed, Photo, and CS, our method significantly outperforms other approaches, demonstrating the superiority of structure-aware mixup for node augmentation. Compared to basic models,

SAMix-GNN achieves an average improvement of more than 4%, especially under limited labeled data. Compared with recent augmentation methods, SAMix-GNN benefits from structurally guided interpolation that generates semantically meaningful and structurally coherent virtual samples, which effectively improves the model's discrimination ability on boundary nodes and enhances its robustness.

To provide a more intuitive understanding of cross-dataset performance, we further present two types of visualization: radar charts and bar plots. FIGURE 2 shows the radar plot of GCN, GAT, SGC, and SAMix-GNN on Cora, Citeseer, and PubMed. SAMix-GNN forms the largest outer polygon, indicating its stable and superior performance under diverse graph structures. FIGURE 3 displays bar charts comparing IMCN, HCPL, RNCGLN, and SAMix-GNN on the same datasets. In Citeseer and PubMed, SAMix-GNN outperforms the second-best method by over 1%, showing the effectiveness of structure-aware interpolation in improving label propagation and sample quality.

In summary, SAMix-GNN consistently outperforms existing models across different datasets and categories, demonstrating strong generalization and stability. The results validate the practical value and adaptability of our structure-aware augmentation framework for graph neural networks.

5. Conclusions

We propose SAMix-GNN, a structure-aware Mixup framework designed to mitigate the under-reaching issue in graph neural networks under limited supervision. By incorporating structural similarity into the Mixup weight generation, our method ensures semantic and topological consistency in the augmented samples. A dual-branch encoder is employed to balance structure modeling and individuality preservation, while confidence-based pseudo-labeling enables high-quality intra-class pair construction. Extensive experiments on benchmark datasets demonstrate that SAMix-GNN achieves superior accuracy, robustness, and generalization compared to existing methods. Future work will explore inter-class structure-aware Mixup, dynamic pair sampling, and integration with contrastive learning to further improve scalability and adaptability.

Acknowledgements

This work was financially supported by National Natural Science Foundation of China (62176085, 62172458) and Industry-University-Research Cooperation Project (GP/026/2020 and HF-010-2021) Zhuhai City, Guangdong Province, China.