FincGAN: A GAN Framework of Imbalanced Node Classification on Heterogeneous Graph Neural Network

Hung-Chun Hsu* [1] Ting-Le Lin* [2] Bo-Jun Wu [2] Ming-Yi Hong [1]
Che Lin [1] Chih-Yu Wang [3†]

[1] National Taiwan University, Taiwan [2] National ChengChi University, Taiwan
[3] Research Center for Information Technology Innovation, Academia Sinica, Taiwan

ABSTRACT

We introduce FincGAN, a GAN framework designed to address class imbalance in GNNs by enhancing minority sample synthesis and ensuring connectivity with sparsity-aware edge generators.

CONTRIBUTIONS

1. Proposed FincGAN, a GAN framework addressing class imbalance in graphs, enhancing minority sample synthesis.
2. Developed sparsity-aware edge generators, ensuring connectivity of synthetic nodes within heterogeneous graphs.
3. Demonstrated FincGAN's superiority in AUC-PRC and F-score against baselines, validating its efficacy in imbalanced datasets.

MOTIVATION

- **Graph's Class Imbalance Problem:** Using Vanilla Oversampling results in model overfitting to the dataset's minority classes.
- **Sparse Edge:** GraphSMOTE fails at sparse distributions; we offer adaptations for sparse edges.

PRELIMINARIES

- Denote $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ as a heterogeneous graph, with nodes $\nu \in \mathcal{V}$ mapped by $\phi(\nu): \mathcal{V} \to \mathcal{A}$ and edges $e \in \mathcal{E}$ mapped by $\psi(e): \mathcal{E} \to \mathcal{R}$. $\mathcal{A}$ and $\mathcal{R}$ denote node and edge type sets; $\mathcal{A}_T$ is the target node type.
- Training involves partial label information Y_L, defining training set labels $\mathcal{V}_L$ and addressing imbalance ratio $\frac{\min_i(|C_i|)}{\max_i(|C_i|)}$.
- *Goal:* A node classifier $\mathcal{F}$ trained on balanced graph $\mathcal{G}$, mapping $\mathcal{F}(\mathcal{V}, \mathcal{E}) \to Y$ to achieve good classification performance.

METHODOLOGY

FincGAN consists of four steps. We employ HGT as a (I) feature extractor to capture graph information. After steps (II) Node Generator and (III) Edge Generator, we train a (IV) GNN Classifier using the augmented graph $\widetilde{\mathcal{G}} = \{\widetilde{\mathcal{V}}, \widetilde{\mathcal{E}}\}$ with a balanced class distribution.

(II) Node Generator

We utilize DCGAN for node generation, generator G combines noise z from $\mathcal{N}(0,1)$ with node embedding h_s and label y_s to create synthetic $\widetilde{h}$, conditioned on node features, label, and noise.

$$\min_G \max_D V(D,G) = \mathbb{E}_{h \sim p_{\text{data}}}[\log D(h,y)] + \mathbb{E}_{\substack{z \sim N(0,1) \\ h_s \sim p_{\text{data}}}}[\log(1 - D(z, h_s, y_s))] \quad (1)$$

(III) Edge Generator

Adapting GraphSMOTE, we customize edge generators for both dense and sparse edges in heterogeneous graphs. Dense connections are modeled with weighted inner products and matrix S, while sparse edges utilize a Multilayer Perceptron (MLP). For a synthetic node v', if $E_{v',u} > \eta$, we connect v' and u; otherwise, they remain unconnected.

$$E_{v,u} = \sigma(h_v \cdot S \cdot h_u) \ \forall v, u \quad (2)$$

$$\mathcal{L}_{\text{D}_{\text{edge}}} = \|E - A\|_F^2, \quad (3)$$

$$\text{MLP}(\text{concat}(h_v, h_u)) \to E_{v,u} \quad \forall v, u \in \mathcal{V}, \quad (4)$$

$$\mathcal{L}_{\text{S}_{\text{edge}}} = -\sum (A_{v,u} \log(E_{v,u}) + (1 - A_{v,u}) \log(1 - E_{v,u})), \quad (5)$$

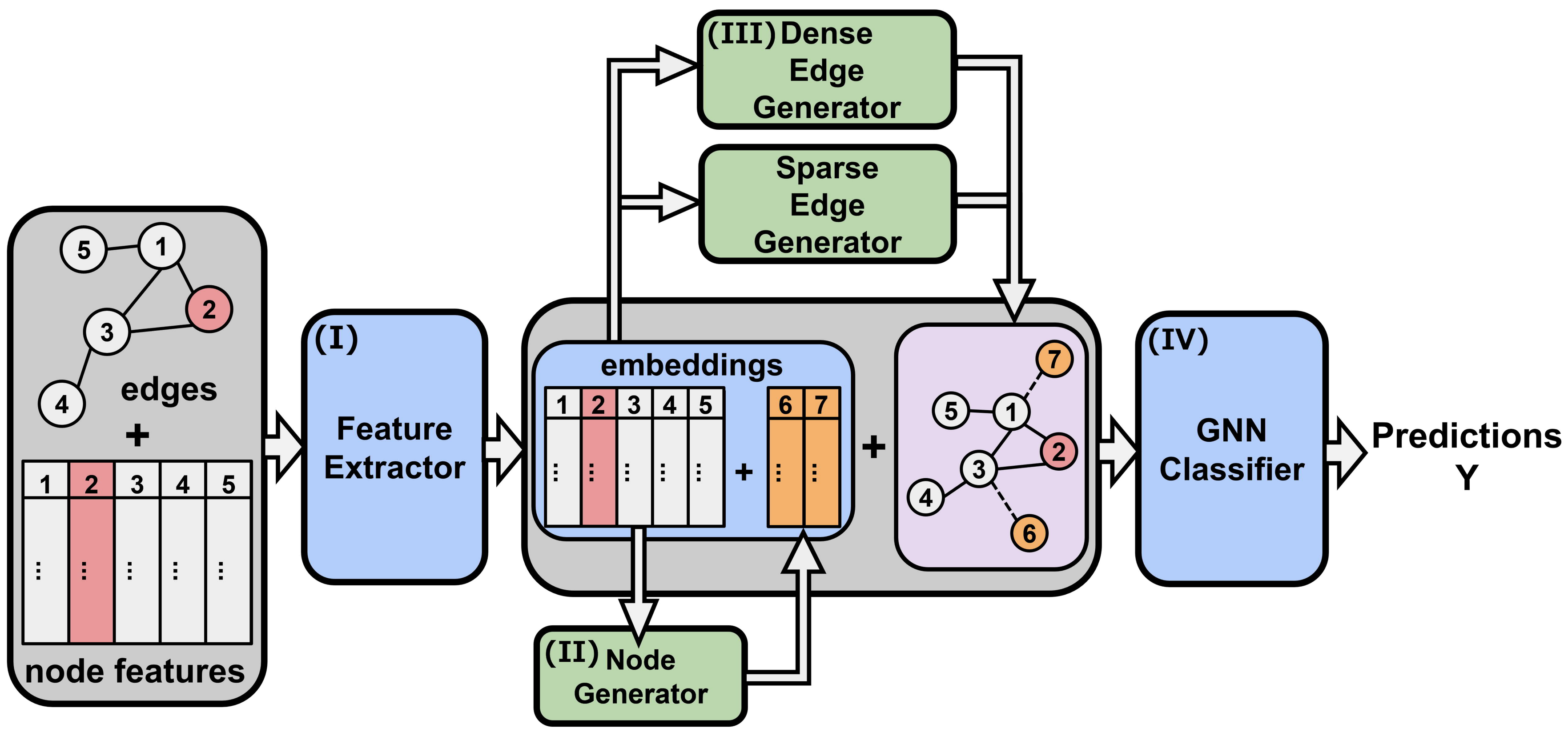


Figure 1. Overview of FincGAN

EXPERIMENTS

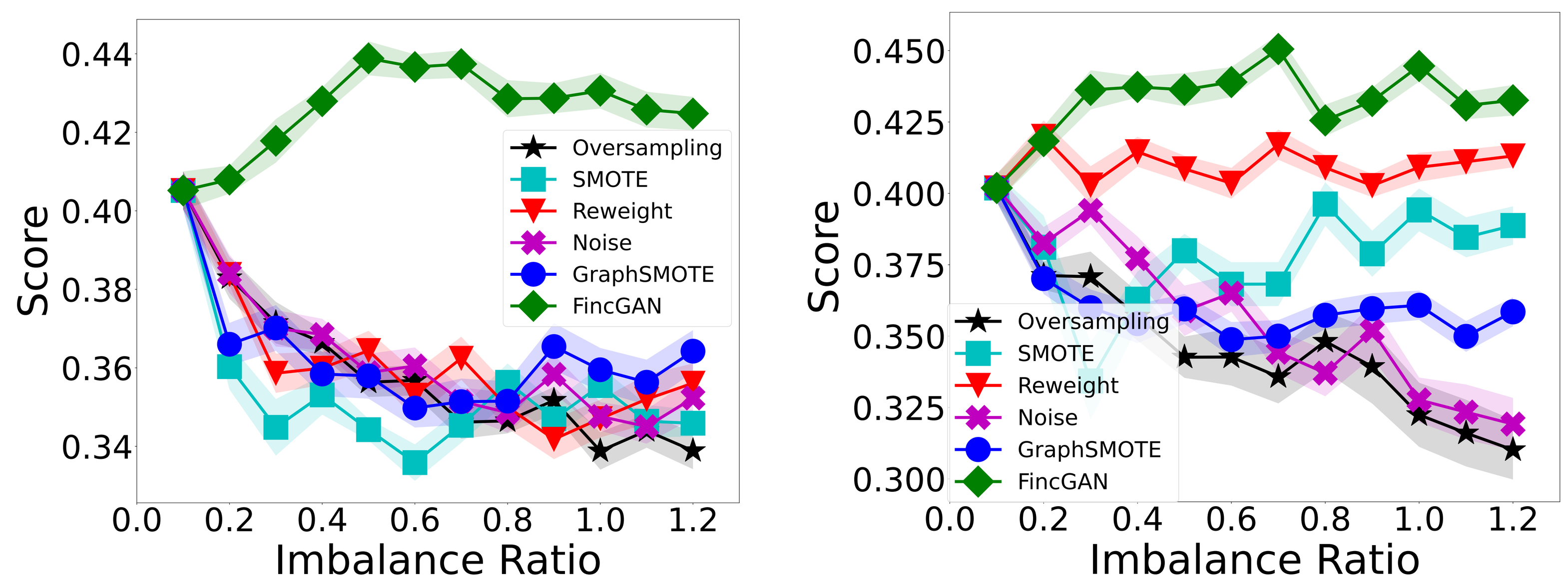
We employ FincGAN to Amazon and Yelp review datasets, constructing class imbalance graphs with diverse node and edge types. We assess FincGAN's performance by examining its effectiveness in addressing class imbalance, the impact of varying up-sampling scales, and the influence of the edge generator.

Table 1. Imbalance Classification Performance

Method	Amazon (Fraud 9.7%)		Yelp (Fraud 20.5%)	
	AUC-PRC	F-score	AUC-PRC	F-score
Original	0.4051	0.4018	0.5005	0.4541
Oversampling	0.3461	0.3357	0.4890	0.4083
SMOTE	0.3453	0.3682	0.4963	0.4537
Reweight	0.3624	0.4170	0.4859	**0.4991**
Noise	0.3514	0.3443	0.5132	0.4449
GraphSMOTE	0.3513	0.3499	0.5130	0.4505
ImGAN	0.2725	0.1812	0.4361	0.2800
PC-GNN	0.3381	0.2923	0.4860	0.4961
FincGAN	**0.4374**	**0.4505**	**0.5173**	0.4688

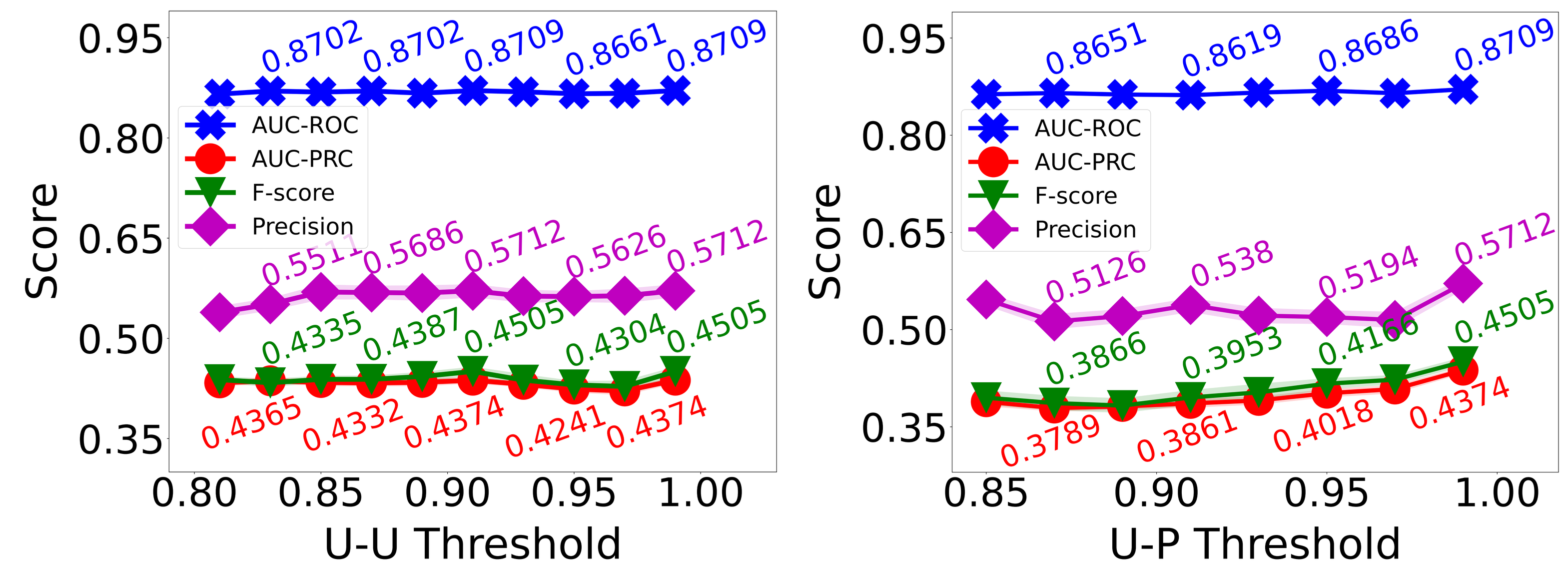
#Amazon: V_{User} *7k,* $V_{Product}$ *4.6k, U-U 0.53M, U-P 24k, P-P 0.1M*
#Yelp: V_U *13k,* V_R *25k,* V_P *570, U-U 14M, U-P 24k, U-R 25k, R-P 25k, P-P 77k*

Effects of up-sampling scale



(a) AUC-PRC (b) F-score

Influence of Edge Generators' Threshold



CONCLUSIONS

The proposed novel framework, FincGAN, addresses class imbalance in heterogeneous graphs by leveraging GAN's generative power to create synthetic nodes and integrate them into the graph, ensuring balanced training. Experimental results showcase FincGAN's effectiveness, surpassing baselines in key performance metrics.

†Corresponding author: cywang@citi.sinica.edu.tw *The first two authors contributed equally to this work.