MG-GNN: Enhancing GNNs for Anomaly Detection via Minority Class Sample Generation

Ronghui Guo[1], Minghui Zou[1], Sai Zhang[1], Xiaowang Zhang[1,*] and Zhiyong Feng[1]

[1]College of Intelligence and Computing, Tianjin University, Tianjin, 300350, China

Abstract

Anomaly detection distinguishes anomalies from normals. In an anomaly graph, both anomalies and normals are represented as nodes, with their relationships denoted by edges. However, in graph anomaly detection, the number of anomalous nodes is typically far fewer than that of normal nodes. To address the issue of class imbalance, existing Graph Neural Networks (GNNs) tend to overlook anomalous (minority class) node samples, resulting in suboptimal performance. To solve this, we propose a method MG-GNN, which generates minority class samples for GNN in the hidden space, thereby improving the classification performance for anomalous nodes. Experiments have demonstrated the effectiveness of our method in solving this problem.

Keywords

Graph anomaly detection, Class imbalance, Graph neural networks

1. Introduction

Typically, the graph anomaly detection (GAD) task is treated as a semi-supervised binary node classification problem (normal vs. anomalous). However, in an anomaly graph, the number of anomalies is significantly lower than normals. Generally, efforts to adapt GNNs to class-imbalanced graphs can be broadly categorized into two types [1]: data-level and algorithm-level methods. Data-level methods typically attempt to balance class distribution by pre-processing the training samples using oversampling or undersampling techniques [2]. Algorithm-level methods consider misclassification costs to focus more on minority classes or to ignore majority classes, thereby mitigating the impact of class imbalance [3].

However, recent GAD methods struggle to adapt to this extreme class imbalance, leading to poor classification performance for anomalous nodes (minority class). Table 1 summarizes the distribution of the two classes of nodes in the YelpChi [4] and Amazon [5] datasets, as well as the test accuracy of a recent GNN for these two classes. It is evident that anomalies constitute only a small portion of the total nodes, and the prediction accuracy for anomalies is significantly lower than that for normals.

In this poster, we propose a method MG-GNN to solve this problem, which generates minority class samples for GNN in the hidden space, mitigating the negative impact of class imbalances. Specifically, we first use a GNN to map the node feature and structural information into hidden

Posters, Demos, and Industry Tracks at ISWC 2024, November 13–15, 2024, Baltimore, USA

*Corresponding author.

ronghui_guo@tju.edu.cn (R. Guo); minghuizou@tju.edu.cn (M. Zou); zhang_sai@tju.edu.cn (S. Zhang); xiaowangzhang@tju.edu.cn (X. Zhang); zyfeng@tju.edu.cn (Z. Feng)

Table 1
The distribution of the number of nodes in the YelpChi and Amazon datasets, and the test accuracy of BWGNN[6] for anomalies and normals.

Dataset	# Nodes (Anomaly%)	Anomaly Acc(%)	Normal Acc(%)
YelpChi	45,954 (14.53%)	61.03	91.76
Amazon	11,944 (6.87%)	82.42	97.54

space. Then, based on the hidden representations of the minority class, we generate a large number of minority nodes to achieve a relatively balanced class distribution before performing classification. The experiments show that our method can handle the class imbalance issues.

2. Methodology

2.1. Problem Definition

Given an anomaly graph $\mathcal{G}$ containing both normal and anomalous nodes, the objective is to learn a classifier $f(\cdot)$ based on the graph $\mathcal{G}$ and a set of partially labeled nodes Y_{Train}. The classifier aims to predict the labels of the unlabeled nodes $\hat{Y}$Test, where 1 represents anomalies and 0 represents normal nodes. The task can be formalised as:

$$f(\mathcal{G}, Y_{Train}) \rightarrow \hat{Y}_{Test} \tag{1}$$

2.2. Model Overview

Our model, MG-GNN, consists of three main components. First, a GNN encoder transforms the node feature and structural information into hidden space. Next, based on the representation of the minority class in the hidden space, a large number of nodes representing the minority class are generated, ensuring that the numbers of normal and anomalous classes are relatively balanced. Finally, a classifier is used to perform classification under these balanced conditions.

2.3. GNN Decoder

To encode the anomaly graph, we utilize BWGNN as the backbone network due to its low- and band-pass characteristics. It is noteworthy that we do not use GCN [7] as the encoder here because GCN is based on the homophily assumption and cannot adequately handle the heterophily of anomaly graphs. The decoder is defined as

$$H = \text{BWGNN}(A, X) \tag{2}$$

where $X \in \mathbb{R}^{N \times d}$ represents the raw node features, A is the adjacency matrix of the graph, and H is the node representation in hidden space.

2.4. Synthetic Node Generator

After obtaining the node representations H, we use SMOTE [8] to generate synthetic anomalies. The basic idea is to interpolate between samples of the target minority class and their nearest neighbors in the hidden space. Specifically, let $h_v \in H$ represent the representation of an anomalous node v. First, the nearest node $h_u \in H$ of node v is found based on Euclidean distance:

$$u = \underset{m}{\operatorname{argmin}} \|h_m - h_v\|, h_m \in H \tag{3}$$

where, unlike [9, 10], we do not require node u to necessarily be an anomaly class, as recent research [11] has shown that this strategy can better expand the decision space of anomalies. Then, we generate a new minority class node h_u through linear interpolation:

$$h_k = \delta h_v + (1 - \delta) h_u \tag{4}$$

where $\delta \in [0, 1]$ is sampled from the Beta distribution. The synthesized node k is labeled as an anomalous node. Therefore, we can obtain a large number of synthesized anomalous nodes. Additionally, we can select more nearest neighbors to get more synthetic anomalous nodes.

2.5. Classifier

After synthesizing a large number of nodes, we stack the representations of the original nodes with the synthesized abnormal nodes to obtain a more balanced class, denoted as H'. Finally, we use another MLP as a classifier for final prediction.

$$\hat{Y} = softmax(MLP(H')) \tag{5}$$

Finally, the loss is calculated using cross-entropy. It is important to note that the loss is computed not only for the original nodes but also for the synthesized anomalous nodes.

3. Experiments

Table 2
Performance Results.

Dataset	YelpChi			Amazon		
Metric	F1-macro	AUC	GMean	F1-macro	AUC	GMean
BWGNN	76.92	90.47	75.68	91.45	96.61	90.12
MG-GNN	78.85	92.57	78.09	92.83	98.14	91.72

We employ three widely used class equalization metrics for fair comparisons, namely F1-macro, AUC and GMean. The experimental results from Table 2 show that after generating a large number of anomaly class nodes using our method, the class imbalance is better handled and the overall performance is improved. Additionally, from Table 3 and Table 1 we observe

Table 3
The test accuracy of MG-GNN for anomalies and normals.

Dataset	Anomaly Acc(%)	Normal Acc(%)
YelpChi	71.72	90.70
Amazon	85.42	97.21

a significant improvement in the accuracy of anomalies, while the accuracy of correct nodes remains largely unaffected. This demonstrates that our method effectively mitigates the impact of class imbalance.

4. Conclusion

In this poster, we propose a method MG-GNN, which generates minority class samples for GNN in the hidden space. Experimental results demonstrate that our method can enhance the classification performance of anomalous nodes while having minimal impact on normal nodes. In future work, we are interested in addressing the issue of class imbalance by leveraging the original distribution of the graph.

Acknowledgments

This work was supported by the Project of Science and Technology Research and Development Plan of China Railway Corporation (N2023J044).

References

[1] M. Zhou, Z. Gong, Graphsr: A data augmentation algorithm for imbalanced node classification, in: B. Williams, Y. Chen, J. Neville (Eds.), Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023, AAAI Press, 2023, pp. 4954–4962. URL: https://doi.org/10.1609/aaai.v37i4.25622. doi:10.1609/AAAI.V37I4.25622.

[2] Y. Liu, X. Ao, Z. Qin, J. Chi, J. Feng, H. Yang, Q. He, Pick and choose: A gnn-based imbalanced learning approach for fraud detection, in: J. Leskovec, M. Grobelnik, M. Najork, J. Tang, L. Zia (Eds.), WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021, ACM / IW3C2, 2021, pp. 3168–3177. URL: https://doi.org/10.1145/3442381.3449989. doi:10.1145/3442381.3449989.

[3] F. Liu, X. Ma, J. Wu, J. Yang, S. Xue, A. Beheshti, C. Zhou, H. Peng, Q. Z. Sheng, C. C. Aggarwal, DAGAD: data augmentation for graph anomaly detection, in: X. Zhu, S. Ranka, M. T. Thai, T. Washio, X. Wu (Eds.), IEEE International Conference on Data Mining,

ICDM 2022, Orlando, FL, USA, November 28 - Dec. 1, 2022, IEEE, 2022, pp. 259–268. URL: https://doi.org/10.1109/ICDM54844.2022.00036. doi:10.1109/ICDM54844.2022.00036.

[4] S. Rayana, L. Akoglu, Collective opinion spam detection: Bridging review networks and metadata, in: L. Cao, C. Zhang, T. Joachims, G. I. Webb, D. D. Margineantu, G. Williams (Eds.), Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, August 10-13, 2015, ACM, 2015, pp. 985–994. URL: https://doi.org/10.1145/2783258.2783370. doi:10.1145/2783258.2783370.

[5] J. J. McAuley, J. Leskovec, From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews, in: D. Schwabe, V. A. F. Almeida, H. Glaser, R. Baeza-Yates, S. B. Moon (Eds.), 22nd International World Wide Web Conference, WWW '13, Rio de Janeiro, Brazil, May 13-17, 2013, International World Wide Web Conferences Steering Committee / ACM, 2013, pp. 897–908. URL: https://doi.org/10.1145/2488388.2488466. doi:10.1145/2488388.2488466.

[6] J. Tang, J. Li, Z. Gao, J. Li, Rethinking graph neural networks for anomaly detection, in: K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvári, G. Niu, S. Sabato (Eds.), International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA, volume 162 of *Proceedings of Machine Learning Research*, PMLR, 2022, pp. 21076–21089. URL: https://proceedings.mlr.press/v162/tang22b.html.

[7] T. N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, in: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings, OpenReview.net, 2017. URL: https://openreview.net/forum?id=SJU4ayYgl.

[8] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, J. Artif. Intell. Res. 16 (2002) 321–357. URL: https://doi.org/10.1613/jair.953. doi:10.1613/JAIR.953.

[9] S. Shi, K. Qiao, C. Chen, J. Yang, J. Chen, B. Yan, Over-sampling strategy in feature space for graphs based class-imbalanced bot detection, in: T. Chua, C. Ngo, R. K. Lee, R. Kumar, H. W. Lauw (Eds.), Companion Proceedings of the ACM on Web Conference 2024, WWW 2024, Singapore, Singapore, May 13-17, 2024, ACM, 2024, pp. 738–741. URL: https://doi.org/10.1145/3589335.3651544. doi:10.1145/3589335.3651544.

[10] T. Zhao, X. Zhang, S. Wang, Graphsmote: Imbalanced node classification on graphs with graph neural networks, in: L. Lewin-Eytan, D. Carmel, E. Yom-Tov, E. Agichtein, E. Gabrilovich (Eds.), WSDM '21, The Fourteenth ACM International Conference on Web Search and Data Mining, Virtual Event, Israel, March 8-12, 2021, ACM, 2021, pp. 833–841. URL: https://doi.org/10.1145/3437963.3441720. doi:10.1145/3437963.3441720.

[11] W. Li, C. Wang, H. Xiong, J. Lai, Graphsha: Synthesizing harder samples for class-imbalanced node classification, in: A. K. Singh, Y. Sun, L. Akoglu, D. Gunopulos, X. Yan, R. Kumar, F. Ozcan, J. Ye (Eds.), Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2023, Long Beach, CA, USA, August 6-10, 2023, ACM, 2023, pp. 1328–1340. URL: https://doi.org/10.1145/3580305.3599374. doi:10.1145/3580305.3599374.