Counterfactual Graph Learning for Anomaly Detection with Feature Disentanglement and Generation (Student Abstract)

Yutao Wei[1], Wenzheng Shu[1], Zhangtao Cheng[1,3]*, Wenxin Tai[1,3], Chunjing Xiao[2], Ting Zhong[1,3]

[1]University of Electronic Science and Technology of China, Chengdu, Sichuan 610054, China
[2]Henan University, Kaifeng, Henan 475000, China
[3]Kash Institute of Electronics and Information Industry, Kashgar 844000, China
yutaowei@std.uestc.edu.cn, shuwenzheng926@gmail.com, wxtai@std.uestc.edu.cn, chunjingxiao@gmail.com, zhongting@uestc.edu.cn, zhangtao.cheng@outlook.com

Abstract

Graph anomaly detection has received remarkable research interests, and various techniques have been employed for enhancing detection performance. However, existing models tend to learn dataset-specific spurious correlations based on statistical associations. A well-trained model might suffer from performance degradation when applied to newly observed nodes with different environments. To handle this situation, we propose **C**ounter**F**actual **G**raph **A**nomaly **D**etection model, CFGAD. In this model, we design a gradient-based separator to disentangle node features into class features and environment features. Then, we present a weight-varying diffusion model to combine class features and environment features from different nodes to generate counterfactual samples. These counterfactual samples will be adopted to enhance model robustness. Comprehensive experiments demonstrate the effectiveness of our CFGAD.

Introduction

Anomalous nodes within graphs are considered as data objects that significantly deviate from the majority in terms of structures and/or properties. Detecting anomalous nodes is a crucial task in a great many industrial applications, as a few anomalies may cause tremendous loss. Correspondingly, significant progress has been achieved by leveraging various techniques on *graph anomaly detection* (GAD) (Ma et al. 2021; Xiao et al. 2023b).

Despite the encouraging results made by existing models, they still struggle in poor generalization beyond training data distribution. The decline in performance can be attributed to distribution shift and shortcut learning, which means that deep learning models tend to learn dataset-specific spurious correlations based on statistical associations. Problems with these characteristics arise when the test data has a different distribution from the training data. Consequently, even a well-trained model might suffer from performance degradation when applied to newly observed nodes with different environments. Existing work (Yang et al. 2021) has shown that learning causal relations can efficiently alleviate this issue in the field of natural language processing.

*corresponding author

In this paper, we try to learning causal relations in the field of graph learning and propose a **C**ounter**F**actual **G**raph **A**nomaly **D**etection model, CFGAD. In this model, we first design a gradient-based selector to disentangle node features of the graph into the class feature and environment feature. Second, we present a weight-varying diffusion model to generate counterfactual samples based on disentangled features. Thus CFGAD provides a direct way for generating counterfactual samples based on decomposed features in the task of GAD. Extensive experiments conducted on two real-world datasets show that our model achieves state-of-the-art results.

Methodology

Problem Definition. Given a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \boldsymbol{X})$, $\mathcal{V}$ donates the node set, $\mathcal{E}$ represents the edge set, and $\boldsymbol{X}$ refers to the set of node features. The task of GAD aims to learn a function f to measure the anomaly score $s_i = f(v_i)$ for each node v_i in $\mathcal{V}$. Herein, the anomaly score s_i indicates the abnormality degree of node v_i, i.e., the higher the anomaly scores, the more likely the nodes are anomalous.

Framework. Our model comprises two main modules. The first module focuses on performing feature separation under a constraint condition, with the goal of separating node features into class features C and environment features E. The second module is responsible for generating counterfactual samples. It combines the class feature from one node with the environment feature from another node to create a counterfactual sample. This counterfactual sample is then utilized to enhance the model's generalization capability and improve the detection performance

Gradient-based Feature Separation. Inspired by (Chen et al. 2020; Pope et al. 2019), we design a gradient-based feature separator to decompose node features $\boldsymbol{X}$ into class features C and environment features E. We aim for C to inherit most of the informative characteristics of the nodes. Meanwhile, We expect E can capture local structure near nodes. To measure the importance of features in the node, we transform the node feature $\boldsymbol{X}$ into a learnable model parameter F and adopt the gradients of F to evaluate the feature importance. Here, F is initialized with the original feature data. The contribution of the k-th feature in F to the anomaly detection at layer l as:

$$\alpha_k^{(l)} = \frac{1}{N}\left|\sum_{n=1}^{N}\frac{\partial y}{\partial F_{k,n}^{(l)}}\right|, \quad (1)$$

Where y is the predicted probability of ground truth, F is the feature representation in hidden layer, and N is the number of samples. Based on this gradient score α, the feature selector uses top-K sampling to adaptively select class-specific features, denoted as C, while treating the remaining features as environment features, denoted as E.

Counterfactual Sample Generation. Due to the superiority of diffusion models (Xiao et al. 2023a), we present a weight-varying diffusion model to combine the class features from one node, c^a, and the environment feature from another node, e^b, to generate counterfactual samples. To this end, we resize the environment feature e^b as the prior, which is fed into the reverse diffusion process to generate a clean embedding through gradual denoising, i.e., $\mathbf{x}^T \rightarrow \mathbf{x}^t \rightarrow \hat{\mathbf{x}}^{t-1} \rightarrow \mathbf{x}^{t-1} \rightarrow \mathbf{x}^0$. During this process, we regard the class feature c^a as the condition, which is exerted into the generative process with varying weights. Specifically, the prior is computed as:

$$x^T = m \odot e^b + (1-m) \odot \left(g\left(e^b\right)\beta + \hat{x}^T(1-\beta)\right), \quad (2)$$

where m is a binary sequence, $g(\cdot)$ is the bicubic interpolation function and β is the weight parameter to adjust the importance of two terms. We impose the condition on the reverse diffusion iteration. According to the Markov chain, the conditional reverse diffusion aims to predict $\hat{x}^{t-1}$ based on $\hat{x}^t$ and c^a. After adding the condition c^a, every step of the reverse process becomes:

$$\begin{aligned} x^{t-1} = m \odot \left((1-h(t-1))\hat{x}^{t-1} + h(t-1)c^a\right) \\ +(1-m) \odot \hat{x}^{t-1}. \end{aligned} \quad (3)$$

We use the real observed values c^a to replace the generated data at each time step to avoid data deviation problem. $m \odot (\cdot)$ and $(1-m) \odot \hat{x}^{t-1}$ refer to the generated values of condition data and environment data respectively. $h(\cdot)$ is a monotonic function for adjusting the weights of the conditions. This generative process iteratively refines the distribution until reaching a clean sample x^0, donated as $\boldsymbol{X}_{cf}$.

Anomaly Detection Model. Having generated counterfactual feature $\boldsymbol{X}_{cf}$ and original feature $\boldsymbol{X}$, we first compute the corresponding node representations:

$$\mathbf{z}_i = \mathrm{AGG}(\mathrm{MLP}(\boldsymbol{X} + \boldsymbol{X}_{cf})), \quad (4)$$

where AGG is a aggregation function that aggregate the neighbor information to form the node representation $\mathbf{z}_i$. Further, we detect anomalies based on these representations:

$$s_i = f_\theta(v_i) = \mathrm{Softmax}(\mathbf{u}^T \cdot (\mathbf{W} \cdot \mathbf{z}_i + \mathbf{b})), \quad (5)$$

where $\mathbf{u}$ and $\mathbf{W}$ are learnable wight vector and weight matrix, respectively. $\mathbf{b}$ is bias term, and $f_\theta(\cdot)$ is a binary classifier that maps the node representations into probability values through $\mathrm{Softmax}$. s_i is the probability value of node v_i being anomalous. Because of generated counterfactual samples, the model can learn a more robust causal representation of the nodes on the graph during training, allowing us to better distinguish abnormal and normal nodes, improving the performance of anomaly detection.

Method	YelpChi		Amazon	
	macro-F_1	AUC	macro-F_1	AUC
GCN	0.5171	0.5689	0.6054	0.8667
GAT	0.5164	0.7403	0.6426	0.8499
GraphConsis	0.6577	0.7853	0.7894	0.9516
CARE-GNN	0.6433	0.7925	0.8988	0.9491
CFGAD	**0.6832**	**0.8129**	**0.9054**	**0.9629**

Table 1: Overall performance on two datasets.

Experiments

Datasets & Baselines. We conduct experiments on two datasets, **YelpChi** is a review network collected from yelp.com and **Amazon** includes product reviews under the Musical Instruments category. Both of the datasets are attributed multi-relation graph. We compare our model with the following baselines: GCN, GAT, CARE-GNN (Dou et al. 2020).

Performance Comparison. The overall comparison results are reported in Table 1. We can observe that CFGAD outperforms all baselines. This result confirms the effectiveness of the gradient-based selector in disentangling node features and learning fine-grained class and environment features. Additionally, the diffusion model can generate high-quality counterfactual samples by leveraging the disentangled features and improve model's generalization and robustness.

Acknowledgments

This work was supported in part by National Natural Science Foundation of China (Grant No.62176043 and No.62072077).

References

Chen, L.; Yan, X.; Xiao, J.; Zhang, H.; Pu, S.; and Zhuang, Y. 2020. Counterfactual samples synthesizing for robust visual question answering. In *CVPR*, 10800–10809.

Dou, Y.; Liu, Z.; Sun, L.; Deng, Y.; Peng, H.; and Yu, P. S. 2020. Enhancing graph neural network-based fraud detectors against camouflaged fraudsters. In *CIKM*, 315–324.

Ma, X.; Wu, J.; Xue, S.; Yang, J.; Zhou, C.; Sheng, Q. Z.; Xiong, H.; and Akoglu, L. 2021. A comprehensive survey on graph anomaly detection with deep learning. *TKDE*.

Pope, P. E.; Kolouri, S.; Rostami, M.; Martin, C. E.; and Hoffmann, H. 2019. Explainability methods for graph convolutional neural networks. In *CVPR*, 10772–10781.

Xiao, C.; Gou, Z.; Tai, W.; Zhang, K.; and Zhou, F. 2023a. Imputation-based Time-Series Anomaly Detection with Conditional Weight-Incremental Diffusion Models. In *KDD*, 2742–2751.

Xiao, C.; Xu, X.; Lei, Y.; Zhang, K.; Liu, S.; and Zhou, F. 2023b. Counterfactual Graph Learning for Anomaly Detection on Attributed Networks. *TKDE*.

Yang, S.; Yu, K.; Cao, F.; Liu, L.; Wang, H.; and Li, J. 2021. Learning causal representations for robust domain adaptation. *TKDE*.