Journal of Computer Science and Artificial Intelligence
ISSN: 3078-8242 (Print), ISSN: 3078-2309 (Online) | Vol. 2, No. 2, 2025

Research on Link Prediction Based on Improved Graph Convolutional Network

Ruijie Huang*

Faculty of Science and Technology, Beijing Normal University-Hong Kong Baptist University United International College, Zhuhai, Guangdong, China

* Corresponding author: s230034077@mail.uic.edu.cn

Abstract: Link prediction is a fundamental task in network analysis with extensive applications in social networks, recommendation systems, and biological networks. In this work, we propose an improved Graph Convolutional Network (IGCN) model that leverages an encoder–decoder architecture, integrating multi-layer GCN-based aggregation of node features and local structure with adaptive hard negative sampling and contrastive learning. Our framework is evaluated on three benchmark datasets—Cora, Citeseer, and Pubmed—and achieves state-of-the-art performance when compared with traditional methods and other graph neural network models. This study demonstrates that incorporating adaptive negative sampling and contrastive loss effectively enhances the discriminative power of node representations for link prediction. Future research will focus on model scalability and the integration of domain-specific knowledge to further broaden its application scope.

Keywords: Link Prediction, Graph Convolutional Network (GCN), Node Embedding, Negative Sampling, AUC Evaluation.

1. Introduction

With the wide application of complex networks in social networks, bioinformatics, recommendation systems and other fields, link prediction, as one of the core problems in network analysis, has become increasingly important. The goal of link prediction is to predict the possibility of future connections between unconnected nodes in the network, and its research results not only help to understand the structural evolution of the network, but also to discover potential relationships and optimize network performance. However, link prediction faces many challenges, such as network sparsity, high dimensional node features and complexity of network structure. Traditional link prediction methods are mainly based on heuristic rules or matrix decomposition. Although these methods have been successful to some extent, they are often difficult to capture the nonlinear structure and the complex relationship between nodes in the network. In recent years, Graph Convolutional Network (GCN), as a powerful graph neural network model, provides a new solution for link prediction tasks by efficiently aggregating node features and local structure information. This paper builds a link prediction model based on GCN, aiming to realize efficient link prediction by encoding node features and decoding potential connections. The main contributions of this paper include: 1) a link prediction framework based on GCN is designed, and the model training is optimized with negative sampling technology; 2) Detailed experimental verification was carried out on the Cora dataset, and the performance of the model was analyzed; 3) The effect of node embedding and prediction link is demonstrated by visual means, which provides support for the interpretability of the model. The structure of this paper is as follows: the second chapter introduces the relevant research work, the third chapter describes the problem definition and model design in detail, the fourth chapter describes the experimental design and implementation process, the fifth chapter analyzes the experimental results, the sixth chapter discusses the advantages and disadvantages of the model and the future improvement direction, and the seventh chapter summarizes the whole paper and looks forward to the practical application prospect. [1].

2. Related work

As an important research direction of network analysis, link prediction is mainly divided into traditional link prediction method and graph neural network based link prediction method. Traditional link prediction methods are usually based on heuristic rules or matrix decomposition techniques. Heuristic rule methods predict potential connections by defining similarity indexes between nodes (such as common neighbors, Jaccard coefficient, Adamic-Adar index, etc.). Although the computational efficiency is high, it is difficult to capture complex structures and nonlinear relationships in the network[2]. Matrix decomposition methods predict links by dissolving the adjacency matrix of the network into a low-dimensional representation. Such methods can reduce data sparsity to a certain extent, but they have limited processing capacity for large-scale networks and cannot make full use of node feature information. In recent years, with the rapid development of Graph Neural Networks (GNNs), link prediction methods based on graph neural networks have gradually become the mainstream. By aggregating local structure information and feature information of nodes, such methods can capture complex patterns of networks more effectively. Graph Convolutional Network (GCN), as an important branch of GNNs, iteratively aggregates node features through convolution operations. The performance of link prediction is significantly improved. GCN has been widely used in social networks, knowledge graphs, biological networks and other fields, but it still has some limitations, such as high computational complexity for large-scale networks, limited ability to represent heterogeneous graphs, and insufficient modeling of long-term dependence relationships[3]. In addition, the existing methods still have room for improvement in negative sampling strategy, loss function design and model interpretability. Based on the existing research and the advantages of GCN, this paper proposes an improved link

prediction framework to address the above limitations and further improve the prediction performance[4].

3. Methodology

In this section, we present our proposed method for link prediction that integrates self-supervised contrastive learning with an adaptive hard negative sampling strategy. Our approach builds upon a standard Graph Convolutional Network (GCN) architecture and refines the learned node representations through advanced data augmentation and a Focal Loss modulated NT-Xent contrastive loss. In addition, an adaptive hard negative sampling scheme is employed to focus the model on challenging negative samples, thereby enhancing its discriminative capability. The following subsections describe each module of our method in detail.

3.1. Graph Convolutional Network-based Link Prediction Model

Given an input node feature matrix $\mathbf{X} \in \mathbb{R}^{N \times d}$ and an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with $N = |\mathcal{V}|$ nodes, we employ a two-layer GCN to capture both feature and structural information. The propagation rule for the GCN is given by:

$$\mathbf{H}^{(l+1)} = \sigma(\tilde{\mathbf{D}}^{-\frac{1}{2}}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-\frac{1}{2}}\mathbf{H}^{(l)}\mathbf{W}^{(l)}),$$

Where $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}$ is the adjacency matrix with added self-loops, $\tilde{\mathbf{D}}$ is its corresponding degree matrix, $\mathbf{W}^{(l)}$ is the learnable weight matrix at layer l, and $\sigma(\cdot)$ is a non-linear activation function (e.g., ReLU). The input layer is initialized as $\mathbf{H}^{(0)} = \mathbf{X}$.

For the link prediction task, we employ an inner product decoder where the probability of an edge between node i and node j is estimated as:

Enhanced Graph Contrastive Module}

3.2. Enhanced Graph Contrastive Module

To improve the robustness and discriminative power of the node representations, we introduce an enhanced graph contrastive module that consists of two primary components: tailored data augmentation and a contrastive loss with focal modulation.

3.3. Data Augmentation Strategies

We generate two augmented views of the graph using the following perturbations:

Random Edge Dropping: For the set of positive edges $\mathcal{E}_{\text{pos}}$, each edge is randomly dropped with probability α. The perturbed positive edge set is defined as:

$$\tilde{\mathcal{E}}_{\text{pos}} = \{e \in \mathcal{E}_{\text{pos}} \mid r_e > \alpha\}, r_e \sim \mathcal{U}(0,1).$$

Adaptive Feature Dropping: For each node i, we compute the feature dropping probability using its degree d_i as follows:

$$p_i = \beta \cdot \frac{d_i - \min_j d_j}{\max_j d_j - \min_j d_j + \epsilon},$$

Where β is a baseline dropping ratio and ϵ is a small constant for numerical stability. A binary mask is applied to drop element(s) in the node feature vector with probability p_i.

These augmentations yield two distinct node embedding views, denoted as $\mathbf{Z}^{(1)}$ and $\mathbf{Z}^{(2)}$, which serve as inputs for the contrastive learning process.

3.4. Contrastive Loss with Focal Modulation

Let $\mathbf{z}_i^{(1)}$ and $\mathbf{z}_i^{(2)}$ be the representations of node i in the two augmented views. We compute the cosine similarity between any two representations $\mathbf{z}$ and $\mathbf{z}'$ as:

$$\text{sim}(\mathbf{z}, \mathbf{z}') = \frac{\mathbf{z}^\top \mathbf{z}'}{\| \mathbf{z} \| \| \mathbf{z}' \|}.$$

For a positive pair $\mathbf{z}_i^{(1)}, \mathbf{z}_i^{(2)}$, the contrastive probability is defined as:

$$p_i = \frac{\exp(\text{sim}(\mathbf{z}_i^{(1)}, \mathbf{z}_i^{(2)})/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(\mathbf{z}_i^{(1)}, \mathbf{z}_k)/\tau)},$$

Where τ is a temperature hyperparameter and $2N$ denotes the total number of samples obtained by concatenating the two views. To focus more on hard-to-discriminate pairs, we introduce a Focal Loss modulation in the contrastive loss:

$$\mathcal{L}_{\text{contrast}} = -\frac{1}{2N}\sum_{i=1}^{2N}(1 - p_i)^\gamma \log(p_i + \epsilon),$$

With γ being the focusing parameter and ϵ a small constant to ensure numerical stability.

3.5. Adaptive Hard Negative Sampling Strategy

Instead of relying on random negative sampling—which often generates easily distinguishable negatives—we introduce an adaptive hard negative sampling procedure within the training process. This strategy comprises the following steps:

Candidate Negative Sample Generation: For every positive edge in $\mathcal{E}_{\text{pos}}$, negative samples are generated through random negative sampling.

Negative Sample Ranking: For a candidate negative edge (i, j), a similarity score is computed as:

$$s_{ij} = \mathbf{z}_i^\top \mathbf{z}_j.$$

A higher score indicates that the negative sample is more challenging to separate from positive examples.

Hard Negative Selection: The top M negative samples with the highest similarity scores are selected to form the final set $\mathcal{E}_{\text{hard}}$ for training.

This adaptive sampling mechanism directs the model's attention towards difficult negatives, thereby sharpening its discriminative performance.

3.6. Loss Function

The overall training objective of our model is formulated as the sum of the link prediction loss and the contrastive loss:

$$\mathcal{L} = \mathcal{L}_{\text{link}} + \lambda \mathcal{L}_{\text{contrast}},$$

Where λ is a balancing parameter. The link prediction loss is defined using the binary cross-entropy loss:

$$\mathcal{L}_{\text{link}} = -\frac{1}{|\mathcal{E}_{\text{pos}}| + |\mathcal{E}_{\text{neg}}|}\sum_{(i,j) \in \mathcal{E}_{\text{pos}} \cup \mathcal{E}_{\text{neg}}} [y_{ij} \log(\hat{y}_{ij}) + (1 - y_{ij}) \log(1 - \hat{y}_{ij})],$$

With $y_{ij} = 1$ for positive samples and $y_{ij} = 0$ for negative samples.

4. Experiments

4.1. Datasets

The Cora dataset is a classic benchmark in graph representation learning, containing 2708 academic papers

represented by 1433-dimensional bag-of-words features across 7 categories; its edges are segmented into training, validation, and test sets via random sampling (ensuring no overlap), node features are normalized and stored in a sparse format, and model training employs the Adam optimizer with binary cross-entropy loss while performance is chiefly evaluated by the AUC and ROC metrics [7,8,9,10]. In contrast, the Citeseer dataset consists of 3312 publications in 6 classes using sparse bag-of-words vectors under similar preprocessing protocols, and the Pubmed dataset offers a larger and more challenging benchmark with 19,717 publications, TF-IDF weighted features, and 44,338 citation links across 3 classes.

4.2. Result

In this section, we first present a comparative study among different graph neural network models on three datasets, followed by an ablation analysis to examine the contributions of individual components in our GCN-based approach.

We compare our proposed method with several widely used GNN models, including GAT (Graph Attention Network), GIN (Graph Isomorphism Network), and GraphSAGE. In addition, we include our GCN-based model enhanced with both self-supervised contrastive learning and adaptive negative sampling. Table 1 summarizes the AUC performance of these models.

Table 1. Results on Cora, Citeseer, and PubMed (%)

Model	Cora	Citeseer	PubMed
GCN	89.12	91.46	97.21
GIN	90.26	92.34	97.69
GAT	91.22	93.55	98.34
GraphSAGE	90.85	94.27	98.26
IGCN	92.28	97.24	98.84

Our comparative experiments evaluate several cutting-edge GNN models on three benchmark datasets: Cora, CiteSeer, and PubMed. The results clearly indicate that the proposed IGCN model consistently outperforms traditional architectures such as GCN, GIN, GAT, and GraphSAGE across all datasets. Specifically, the IGCN model achieves state-of-the-art AUC scores on each dataset, demonstrating its superior ability to capture both local structural information and discriminative node features. This improvement can be attributed to the integration of adaptive hard negative sampling and self-supervised contrastive learning within the IGCN framework. The adaptive negative sampling efficiently selects challenging negative samples, enhancing the model's ability to distinguish between subtle differences in node relationships. Meanwhile, the contrastive learning strategy provides robust feature representations by aligning multiple augmented views of each node.

Overall, the enhanced performance of the IGCN model not only validates the effectiveness of these additional components but also underscores its potential as a robust solution for link prediction tasks in complex networks.

4.3. Ablation Study

To further understand the impact of each module in our GCN-based approach, we perform an ablation study with three variants: no_hard: GCN with self-supervised contrastive learning added; no_contrastive: GCN with adaptive negative sampling incorporated; IGCN: GCN that integrates both self-supervised contrastive learning and adaptive negative sampling. As shown in Table 2:

Table 2. Ablation study on Cora, Citeseer, and PubMed (%)

Model	Cora	Citeseer	PubMed
no_hard	91.52	94.32	97.33
no_contrastive	89.48	93.27	97.16
IGCN	92.28	97.24	98.84

In our ablation study, we evaluated the individual contributions of adaptive hard negative sampling and self-supervised contrastive learning by comparing two variant models against our full IGCN model. One variant removed the adaptive hard negative sampling component (no_hard), while the other omitted the contrastive learning module (no_contrastive). The results indicate that both components are crucial: each ablated model exhibits a noticeable reduction in performance compared to the full model.

Specifically, the variant lacking adaptive hard negative sampling (no_hard) shows inferior link prediction performance, suggesting that selecting hard negative samples is essential for challenging the model during training and enhancing robustness. Similarly, the variant without contrastive learning (no_contrastive) also underperforms, highlighting that contrastive loss plays an important role in refining node representations by enforcing consistency across different augmented views. Importantly, our IGCN model, which integrates both adaptive hard negative sampling and self-supervised contrastive learning, consistently achieves the best results across all datasets. This synergistic effect underscores that the combination of these two techniques contributes significantly to the improved discriminative power and overall performance of the model for link prediction tasks.

5. Conclusion

In this study, we introduced the Improved Graph Convolutional Network (IGCN) for link prediction. The IGCN model leverages an encoder–decoder architecture, integrating multi-layer GCN-based feature extraction with advanced techniques such as hard negative sampling and self-supervised contrastive learning. Our comprehensive experiments, including both comparative and ablation studies, consistently demonstrate the superiority of the full IGCN model over its variants. The ablation experiments clearly highlight that both hard negative sampling and contrastive learning are essential components. Without either module, the model's ability to capture complex node relationships and structural nuances is significantly diminished. The combined incorporation of these techniques not only enhances the quality of the generated node embeddings but also leads to marked improvements in link prediction performance, as evidenced by consistently high AUC values and stable training convergence.

Overall, our results validate that IGCN is an efficient, robust, and interpretable solution for link prediction in complex networks. Its ability to effectively integrate node features and structural information makes it a promising candidate for a wide range of applications—from social network analysis and recommendation systems to biological network modeling. Future work will focus on scaling the model for large-scale and heterogeneous graphs and incorporating additional domain knowledge to further enhance its performance and applicability.

References

[1] Zhang M, Chen Y. Link prediction based on graph neural networks[J]. Advances in neural information processing systems, 2018, 31.

[2] Cai L, Ji S. A multi-scale approach for graph link prediction[C]//Proceedings of the AAAI conference on artificial intelligence. 2020, 34(04): 3308-3315.

[3] Cai L, Li J, Wang J, et al. Line graph neural networks for link prediction[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 44(9): 5103-5113.

[4] Cukierski W, Hamner B, Yang B. Graph-based features for supervised link prediction[C]//The 2011 International joint conference on neural networks. IEEE, 2011: 1237-1244.

[5] Arrar D, Kamel N, Lakhfif A. A comprehensive survey of link prediction methods[J]. The journal of supercomputing, 2024, 80(3): 3902-3942.

[6] Nguyen T K, Fang Y. Diffusion-based negative sampling on graphs for link prediction[C]//Proceedings of the ACM Web Conference 2024. 2024: 948-958.

[7] Deng W, Zhang Y, Yu H, et al. Knowledge graph embedding based on dynamic adaptive atrous convolution and attention mechanism for link prediction[J]. Information Processing & Management, 2024, 61(3): 103642.

[8] Dileo M, Zignani M, Gaito S. Temporal graph learning for dynamic link prediction with text in online social networks[J]. Machine Learning, 2024, 113(4): 2207-2226.

[9] Li M, Wang Z, Liu L, et al. Subgraph-aware graph kernel neural network for link prediction in biological networks[J]. IEEE Journal of Biomedical and Health Informatics, 2024.

[10] Zhang Y, Chen J, Cheng Z, et al. Edge propagation for link prediction in requirement-cyber threat intelligence knowledge graph[J]. Information Sciences, 2024, 653: 119770.