Electronics, Communications and Networks
A.J. Tallón-Ballesteros et al. (Eds.)

doi:10.3233/FAIA231258

Design and Implementation on Citation Network Link Prediction System Based on GAT

Yiqin BAO[a1], Wenbin XU[b] and Lei WANG[c]
[a] College of Information Engineering of Nanjing XiaoZhuang University, China
[b] Jiangsu United Vocational and Technical College Suzhou Branch, China
[c] Nanjing Yilaichuang Electronic Technology Co., Ltd, China

Abstract. Link prediction has important practical value in many fields, such as social networks, bioinformatics, citation networks, etc. However, existing link prediction methods mainly have two major problems: firstly, they neglect the attribute information of nodes or edges, which limits the accuracy and robustness of prediction; Secondly, when dealing with large-scale complex networks, especially those with rich attribute information, their performance still needs to be improved. To address these issues, the paper designs a citation network link prediction system based on Graph Attention Network (GAT). The system preprocesses the citation network data, constructs and trains a GAT model, and then uses heterogeneous graph convolution and temporal graph convolution to handle the heterogeneity and dynamism of the citation network. Then, a graph sampling strategy is used to handle large-scale citation networks, Finally, use the trained GAT model to predict possible links.

Keywords. Link prediction, graph attention network, graph convolutional network

1. Introduction

The research on link prediction problems began in the last century, with the initial methods mainly based on statistical methods and matrix decomposition methods. Statistical methods typically calculate the likelihood of connections between two nodes based on the topology of the network, such as the number of shared neighbors, Jaccard coefficients, etc. The method based on matrix decomposition predicts missing edges by decomposing the adjacency matrix or similarity matrix of the graph. These methods mainly have two major problems: firstly, they rely on the topology information of the network and ignore the attribute information of nodes or edges, resulting in low accuracy of prediction results; Secondly, when dealing with large-scale complex networks, performance is often limited, especially when dealing with complex

[1] Corresponding author: Yiqin BAO, email: 392335241@qq.com

networks with rich attribute information (such as academic citation networks), their performance needs to be improved [1].

In recent years, with the rapid development of deep learning, neural network-based link prediction methods have begun to receive attention. They usually use the embedded representation of learning nodes to predict links, such as graph convolutional networks (GCN), graph attention networks (GAT) [2-4]. These methods consider the attribute information of nodes and edges, improving prediction accuracy. However, they rely on a large amount of training data, and their performance will be limited when dealing with large-scale networks. These issues also propose new topics and directions for future research in this article.

Most current methods focus on solving the problem of link prediction for a single network, while multi network link prediction presents higher complexity, requiring simultaneous consideration of the topology and attribute information of multiple networks. Traditional link prediction methods face challenges such as data sparsity, feature diversity, and imbalanced samples, therefore a novel method is needed to improve the accuracy and efficiency of link prediction [5].

This paper designs a citation network link prediction system based on GAT. By incorporating attention mechanisms into the node feature aggregation process, GAT can adaptively adjust the influence weights of adjacent nodes on the current node. This not only mines the topology structure of the network, but also considers the attribute information of nodes and edges, thereby significantly improving the accuracy of prediction.

The contributions of this paper are as follows:

1) Summarized link prediction technology and GAT algorithm.

2) Design the architecture of a citation network link prediction system based on GAT, and briefly describe each module.

3) The system was tested and analyzed.

The remaining parts of the paper are organized as follows. The second section studies relevant technologies, the third section designs the citation network link prediction system for GAT, the fourth section tests the system, and the fifth section summarizes the full text and prospects.

2. Related technology

2.1 Link prediction technology

The field of link prediction technology refers to the research field that reveals potential correlations and patterns in network structure by analyzing and predicting the connection relationships between nodes in the network. One of its main applications is to predict citation relationships between academic papers, that is, to predict potential new citation links in existing citation networks, helping researchers understand important issues such as knowledge dissemination, academic collaboration, and academic influence in the academic community.

The link prediction method based on graph attention network (GAT) belongs to an innovative method in the field of this technology [6]. The GAT method constructs a graph structure of academic citation networks, and combines the content and structural information of nodes to adaptively adjust the connection weights between nodes through attention mechanisms, thereby achieving accurate link prediction. Compared

with traditional link prediction methods, GAT method has advantages in addressing challenges such as data sparsity, multimodal features, and imbalanced sample ratios in academic citation networks, and can provide more accurate and interpretable prediction results [7].

2.2 GAT algorithm

GAT adopts the Attention mechanism, which can assign different weights to different nodes. During training, it relies on paired adjacent nodes rather than specific network structures, and can be used for inductive tasks [8-9].

Assuming the Graph contains N nodes, each with a feature vector of hi, a dimension of F, and a node feature vector of h, as shown in formula (1).

$$h=\{h_1, h_2, \dots, h_N\} \quad h_i \in R^F \tag{1}$$

By performing a linear transformation on the node feature vector h, a new feature vector h_i' with dimension F' can be obtained. As shown in formula (2) and formula (3), W is the matrix of the linear transformation.

$$h_i'=Wh_i \quad W \in R^{F' \times F} \tag{2}$$

$$h_i'=\{h_1', h_2', \dots, h_N'\} \quad h_i' \in R^{F'} \tag{3}$$

The Attention mechanism can be used to calculate the importance of node j to node i. The attention score is shown in formula (4) and formula (5), where node j is a neighbor of node i.

$$e_{ij}= Attetion(Wh_i, Wh_j) \tag{4}$$

$$\alpha_{ij}=Softmax_j(e_{ij})=\frac{exp(e_{ij})}{\sum_{k \in N_i} exp(e_{ik})} \tag{5}$$

The specific attention method for calculating GAT is to concatenate the feature vectors of nodes i and j together, and then calculate the inner product with a 2F' dimensional vector a to represent the concatenation operation. The activation function uses LeakyReLU, and the GAT formula is shown in formula (6).

$$\alpha_{ij}=Softmax_j(e_{ij})=\frac{exp(LeakyRelU(a^T [Wh_i \| Wh_j]))}{\sum_{k \in N_i} exp(LeakyRelU(a^T [Wh_i \| Wh_k]))} \tag{6}$$

Schematic diagram of GAT Attention, as shown in Figure 1.

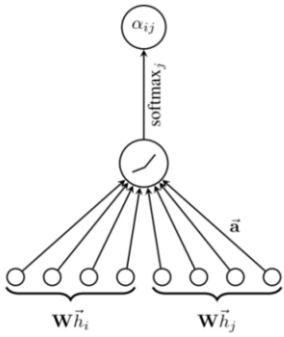

Figure 1. GAT attention schematic diagram

The feature vector of node i after Attention is shown in formula (7).

$$\boldsymbol{h_i'} = \alpha\left(\sum_{\mathbf{k}\in\mathbf{N_i}} {}_{\mathbf{ij}}\,\mathbf{W}\ \mathbf{h_j}\right) \tag{7}$$

GAT can also use Multi Head Attention, which means multiple Attention, as shown in Figure 2.

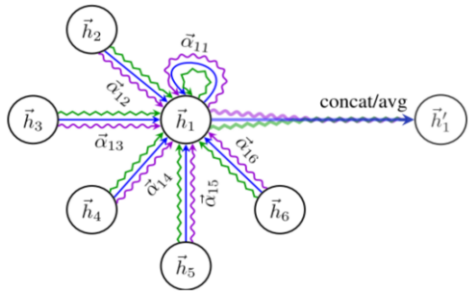

Figure 2. GAT Multi-Head Attention

If there are K Attention, the vectors generated by K Attention need to be concatenated together, as shown in formula (8).

$$h_i' = \text{concat}\left(\ \alpha\left(\sum_{\mathrm{j}\in\mathrm{N_i}} {}^{K}_{ij}\,W^K\mathrm{h_j}\right)\right) \tag{8}$$

If it is the last layer, the output of K Attention is not concatenated, but the average value is calculated, as shown in formula (9).

$$h_i' = \alpha\left(\frac{1}{K}\sum_{\mathrm{K=1}}^{K}\sum_{\mathrm{j}\in\mathrm{N_i}} {}^{K}_{ij}\,W^K\mathrm{h_j}\ \right) \tag{9}$$

3. System Design

3.1 System architecture

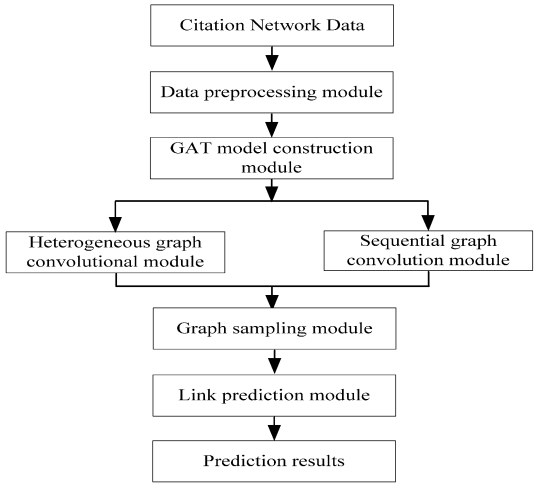

Figure 3. Link prediction system architecture

The GAT based citation network link prediction system designed in this paper consists of multiple modules, and the system framework is shown in Figure 3:

1） Data preprocessing module

This module is responsible for cleaning raw data, converting data formats, and extracting attribute information of nodes and links. It includes the following steps:

a. Data cleaning: Remove invalid or incorrect nodes and links, such as untitled papers, citation relationships without citation context, etc.

b. Data conversion: Transforming raw data into a format acceptable to the model, such as converting the title and abstract of a paper into word vectors, converting citation counts into numerical values, etc.

c. Feature extraction: Extracting useful features from the attribute information of nodes and links, such as extracting topics from the title and abstract of a paper, extracting emotions from citation contexts, etc.

2） GAT model construction module

This module is responsible for building and training the GAT model, which includes the following steps:

a. Model construction: Based on input data and predefined parameters (such as the size of the hidden layer, the number of attention heads, etc.), construct the structure of the GAT model.

b. Model training: Use optimization algorithms (such as gradient descent) to train the parameters of the GAT model based on training data and loss functions (such as cross entropy loss).

3） Heterogeneous Graph Convolutional Module

This module is mainly responsible for handling the heterogeneity of citation networks, designing different convolutional kernels for each type of node and link. Through convolutional kernels, complex relationships between different types of nodes and links can be captured [10-11]. The module includes the following steps:

a. Convolutional kernel design: Design different convolutional kernels based on the type of nodes and links.

b. Heterogeneous graph convolution: Use a designed convolution kernel to perform heterogeneous graph convolution operations on the citation network.

4）Sequential graph convolution module

This module is mainly responsible for handling the dynamism of the citation network, designing different convolutional kernels for each time step [12], including the following steps:

a. Convolutional kernel design: Design different convolutional kernels based on time steps.

b. Temporal graph convolution: Use a designed convolution kernel to perform temporal graph convolution operations on the citation network.

5） Graph sampling module

This module is mainly responsible for handling large-scale citation networks, including the following steps:

a. Sampling strategy design: Design a new graph sampling strategy so that the sampled graph can still maintain the main characteristics of the original graph.

b. Graph sampling: Use a designed sampling strategy to perform graph sampling operations on the citation network.

6) Link prediction module

This module is mainly responsible for using the trained GAT model to predict possible links in the citation network. It includes the following steps:

a. Feature extraction: Use the GAT model to extract features from nodes and links in the citation network.

b. Link prediction: Based on extracted features, use predefined prediction functions (such as sigmoid function) to predict possible links.

4. System testing

4.1 Experimental dataset and parameter settings

This paper uses the Cora dataset, a commonly used academic citation network dataset that includes 2708 scientific publications as nodes and 5429 citation relationships as edges. There are a total of 7 categories, namely neural networks, reinforcement learning, rule learning, probability methods, genetic algorithms, theoretical research, and case studies. The characteristics of each paper are obtained through a word bag model, with a dimension of 1433. Each dimension represents a word, with 1 indicating that the word has appeared in the paper and 0 indicating that it has not appeared. This section corresponds to the input of the "content" attention mechanism. The information of the adjacency matrix corresponds to the input of the structural attention mechanism. The goal of a dataset is to predict the domain category to which each node belongs based on its characteristics and citation relationships. In the experiment, the dataset was divided into training, validation, and testing sets, accounting for 80%, 10%, and 10%, respectively.

In terms of parameter settings, factors such as the number of layers of GAT, the number of hidden units per layer, the number of attention heads, learning rate, and training frequency were mainly considered. Specifically, set the number of layers for GAT to 2, the number of hidden layer nodes in each layer to 8, and the number of attention heads to 8. During the training process, use the Adam optimizer and set the learning rate to 0.005. In addition, L2 regularization has been added to prevent overfitting, with a regularization coefficient of 0.0005. The training frequency is 1000, and an early stop strategy has been implemented on the validation set. If the loss function value does not decrease, the training will be stopped to prevent overfitting.

4.2 Experimental analysis

In the experiment, the main focus is on the accuracy of link prediction. This paper uses Area Under Curve (AUC) under the ROC curve as an evaluation indicator. The experimental results are shown in Table 1.

Table 1. Comparison of GAT model

No.	GAT model	AUC
1	Using Content Attention	0.895
2	Using Structural Attention	0.910
3	Combining Content and Structure Attention	0.932

By comparing the experimental results, the dual attention mechanism can better utilize the information in the citation network and improve the accuracy of link prediction. The combination of these two attention mechanisms enables the model to better understand and utilize the information in the citation network, thereby achieving better results in link prediction tasks.

5. Conclusions

The paper designs a GAT based academic citation network link prediction system. The system comprehensively models node content and structure through a dual attention mechanism, achieving accurate link prediction and providing an effective solution for link prediction problems in academic citation networks.
The paper only implements an academic citation network link prediction system based on graph attention networks. However, further research is needed to achieve academic citation network link prediction through deep learning methods.

Acknowledgement

This work is supported by Natural Science Foundation Project of China (61976118), Key topics of the '13th five-year plan' for Education Science in Jiangsu Province (B-b /2020/01/18).

References

[1] Avros R, Keshet S, Kitai DT, Vexler E, Volkovich Z. Detecting Pseudo-Manipulated Citations in Scientific Literature through Perturbations of the Citation Graph. Mathematics. 2023; 11(18):3820.
[2] Zhu J, Li B, Zhang Z, Zhao L, Li H. High-Order Topology-Enhanced Graph Convolutional Networks for Dynamic Graphs. Symmetry. 2022; 14(10):2218.
[3] Wu J, Zhu Y, Wang C, Li J, Zhu X. A Prior Knowledge-Guided Graph Convolutional Neural Network for Human Action Recognition in Solar Panel Installation Process. Applied Sciences. 2023; 13(15):8608.
[4] Li L, Liu L, Du X, Wang X, Zhang Z, Zhang J, Zhang P, Liu J. CGUN-2A: Deep Graph Convolutional Network via Contrastive Learning for Large-Scale Zero-Shot Image Classification. Sensors. 2022; 22(24):9980.
[5] Nair LS, Jayaraman S, Krishna Nagam SP. An Improved Link Prediction Approach for Directed Complex Networks Using Stochastic Block Modeling. Big Data and Cognitive Computing. 2023; 7(1):31.
[6] Lin S, Hong J, Lang B, Huang L. DAG: Dual Attention Graph Representation Learning for Node Classification. Mathematics. 2023; 11(17):3691.
[7] Lu, J.; Shi, L.; Liu, G.; Zhan, X. Dual-Channel Edge-Featured Graph Attention Networks for Aspect-Based Sentiment Analysis. Electronics 2023, 12, 624.
[8] Zhang, W., Yin, Z., Sheng, Z., Li, Y., Ouyang, W., Li, X., Tao, Y., Yang, Z. & Cui, B. Graph attention multilayer perceptron in Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (2022), 4560-4570.
[9] Zhao, Y., Du, H., Liu, Y., Wei, S., Chen, X., Zhuang, F., Li, Q. & Kou, G. Stock Movement Prediction Based on Bi-Typed Hybrid-Relational Market Knowledge Graph Via Dual Attention Networks. IEEE Transactions on Knowledge and Data Engineering (2022).
[10] Wang Y, Xu X. ERGCN: Enhanced Relational Graph Convolution Network, an Optimization for Entity Prediction Tasks on Temporal Knowledge Graphs. Future Internet. 2022; 14(12):376.

[11] Ye Z, Zhao H, Zhang K, Zhu Y. Multi-View Network Representation Learning Algorithm Research. Algorithms. 2019; 12(3):62.
[12] Sighencea BI, Stanciu IR, Căleanu CD. D-STGCN: Dynamic Pedestrian Trajectory Prediction Using Spatio-Temporal Graph Convolutional Networks. Electronics. 2023; 12(3):611.