

Feature Mining and Node Prediction for Complex Correlation Datasets Based on Graph Neural Networks

XU, Wenhao^{1*}

¹ Northeastern University, US

* XU, Wenhao is the corresponding author, E-mail: xu.wenhao@northeastern.edu

Abstract: Complex correlation datasets widely exist in social networks, citation networks, and biological systems, where traditional machine learning methods fail to effectively capture implicit topological correlation and high-dimensional feature interactions. Graph Neural Networks (GNNs) have unique advantages in processing non-Euclidean graph-structured data, which can realize adaptive feature mining and end-to-end node prediction. This paper proposes a multi-layer graph convolution feature mining model with adaptive neighbor aggregation, aiming at the problems of incomplete feature extraction and weak topological dependency perception in traditional GNNs for complex correlation datasets. Firstly, the graph structure normalization and feature dimensionality reduction are performed on the input complex graph data. Secondly, an improved graph convolution aggregation formula is constructed to fuse node attribute features and topological structure features hierarchically, and a residual connection mechanism is introduced to solve the gradient disappearance problem of deep graph networks. Finally, node classification prediction tasks are realized based on mined deep graph features. Comparative experiments are conducted on six classic complex graph benchmark data sets including Cora, Citeseer, Pubmed, Cornell, Texas, and Wisconsin. The experimental results show that the proposed model achieves average prediction accuracy improvements of 3.2%–8.7% compared with baseline models such as GCN, GAT, and GraphSAGE, which verifies the effectiveness and superiority of the proposed method in complex correlation data feature mining and node prediction.

Keywords: Complex Correlation Dataset, Feature Mining, Node Prediction, Graph Convolution, Topological Feature.

Disciplines: Intelligent Systems.

Subjects: Other.

DOI: <https://doi.org/10.70393/6a696574.343331>

ARK: <https://n2t.net/ark:/40704/JIET.v1n3a01>

1 INTRODUCTION

With the rapid development of big data technology, massive complex correlation data with graph structure characteristics have emerged in various fields. Different from structured Euclidean data such as images and texts, complex correlation datasets take nodes as basic units and edges as correlation constraints[1], presenting non-uniform topology, irregular node correlation, and mixed high-dimensional features. Node prediction, as a core task of graph data mining, aims to predict node category attributes or hidden features based on observable node information and topological correlations, which has important application values in academic citation analysis, social relationship mining, and biological network inference[2].

Traditional node prediction methods mainly rely on manual feature engineering and shallow machine learning algorithms. Methods such as principal component analysis (PCA) and support vector machine (SVM) only focus on node attribute features and ignore the implicit topological correlation between nodes, resulting in insufficient feature mining ability for complex graph data. With the rise of graph

deep learning, Graph Neural Networks (GNNs) have become the mainstream method for graph data processing[3]. The overall workflow of a graph neural network is shown in Figure 1. Classic GNN models such as Graph Convolutional Network (GCN), Graph Attention Network (GAT), and GraphSAGE realize node feature learning through neighbor information aggregation. However, most existing models have limitations in processing complex correlation datasets: vanilla GCN ignores the difference in neighbor node importance; GAT has high computational complexity in multi-head attention calculation; GraphSAGE has insufficient mining ability for high-order topological correlation of dense graphs[4].

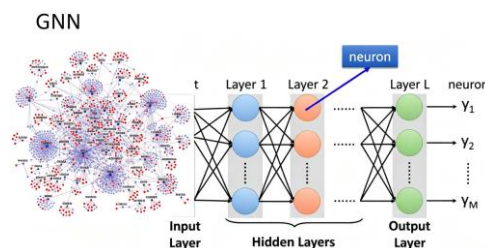


FIGURE 1. COMPLETE OPERATIONAL PIPELINE OF THE GRAPH NEURAL NETWORK

To solve the above problems, this paper constructs an improved multi-layer graph convolution feature mining model for complex correlation datasets. The main contributions are summarized as follows: (1) An adaptive neighbor weighted aggregation mechanism is designed to fuse node attribute features and multi-order topological features, realizing fine-grained mining of complex correlation information; (2) Residual connection and layer normalization strategies are introduced to optimize the deep network training process and improve the stability of feature representation; (3) Sufficient comparative experiments are carried out on multiple heterogeneous complex graph datasets, and the results verify the robustness and prediction performance of the proposed model.

2 RELATED WORK

2.1 TRADITIONAL GRAPH DATA MINING

METHODS

Early graph data mining methods mainly focus on graph structure statistics and manual feature extraction. Typical methods include graph-based dimensionality reduction, random walk feature learning, and spectral clustering. These methods rely heavily on artificial prior knowledge, which can only extract shallow structural features and cannot adaptively learn deep correlation features of complex datasets[5]. Moreover, traditional methods are poor in scalability and cannot handle large-scale complex graph data with uneven node distribution and sparse or dense edges[6].

2.2 CLASSIC GRAPH NEURAL NETWORK

MODELS

GCN realizes graph convolution operation based on spectral graph theory, aggregates neighbor node features through normalized Laplacian matrix, and efficiently extracts local topological features[7]. The forward propagation structure of Graph Convolutional Network (GCN) is shown in Figure 2. GAT introduces multi-head attention mechanism to assign different weight coefficients to neighbor nodes, solving the problem of equal weight aggregation of GCN, but the attention calculation increases model complexity. GraphSAGE adopts inductive learning mode, samples neighbor nodes layer by layer for aggregation, which is suitable for large-scale graph data, but it ignores the global topological correlation of graphs. In view of the defects of existing models in complex correlation data feature mining, this paper optimizes the aggregation mode and network structure to improve the node prediction accuracy.

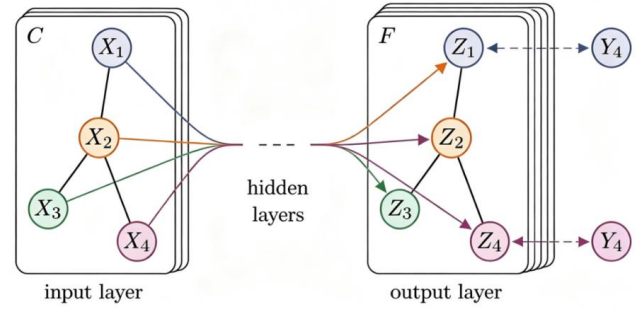


FIGURE 2. FORWARD PROPAGATION STRUCTURE OF GRAPH CONVOLUTIONAL NETWORK (GCN)

3 METHODOLOGY

3.1 GRAPH DATA DEFINITION

A complex correlation graph dataset is defined as an undirected weighted graph $G = (\mathcal{V}, \mathcal{E}, X)$, where $\mathcal{V} = \{v_1, v_2, \dots, v_N\}$ represents the node set with N nodes, $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ represents the edge set reflecting node correlations, and $X \in \mathbb{R}^{N \times d}$ is the node attribute feature matrix, where d is the dimension of original node features. The adjacency matrix of the graph is defined as $A \in \mathbb{R}^{N \times N}$, where $A_{ij} = 1$ if there is an edge between node v_i and v_j , otherwise $A_{ij} = 0$. The degree matrix is $D \in \mathbb{R}^{N \times N}$, and the diagonal element $D_{ii} = \sum_{j=1}^N A_{ij}$.

3.2 IMPROVED GRAPH CONVOLUTION FEATURE MINING FORMULA

The core of GNN feature mining is to realize iterative aggregation of node own features and neighbor topological features. Based on the vanilla GCN convolution formula, this paper constructs an adaptive weighted aggregation convolution formula with residual connection. The standard graph convolution formula of the l -th layer is defined as:

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

where $\tilde{A} = A + I$ is the adjacency matrix with self-loop, $\tilde{D}$ is the degree matrix corresponding to $\tilde{A}$, $H^{(l)}$

is the node feature matrix of the l -th layer, $W^{(l)}$ is the trainable weight matrix, and $\sigma(\cdot)$ is the ReLU activation function.

Aiming at the problem of insufficient feature discrimination of vanilla GCN, this paper introduces adaptive neighbor weight coefficient α_{ij} and residual connection mechanism. The improved graph convolution feature mining formula is:

$$H^{(l+1)} = \sigma \left(\tilde{D}^{-\frac{1}{2}} \tilde{A}^{\alpha} \tilde{D}^{-\frac{1}{2}} H^{(l)} W^{(l)} \right) + H^{(l)} \quad (2)$$

where $\tilde{A}^{\alpha}$ is the weighted adjacency matrix, and the adaptive weight α_{ij} between node i and node j is calculated as:

$$\alpha_{ij} = \frac{\exp(\cosine(h_i, h_j))}{\sum_{k \in \mathcal{N}(i)} \exp(\cosine(h_i, h_k))} \quad (3)$$

where $\mathcal{N}(i)$ represents the neighbor set of node i , h_i, h_j are the feature vectors of node i and j , and cosine similarity is used to measure the feature correlation between nodes, so as to realize differentiated aggregation of neighbor features.

3.3 NODE PREDICTION MODULE

After multi-layer adaptive graph convolution feature mining, the model obtains the final deep node feature matrix $H^{(L)}$ (where L is the total number of convolution layers). For node classification prediction tasks, a full connection layer and softmax activation function are used to map deep features to category probability distribution:

$$\hat{Y} = \text{softmax}(H^{(L)}W_{fc} + b_{fc}) \quad (4)$$

where W_{fc}, b_{fc} are the weight and bias parameters of the full connection layer, and $\hat{Y}$ is the predicted node category probability matrix. The cross-entropy loss function is used for model training optimization:

$$\mathcal{L} = -\frac{1}{N_{\text{train}}} \sum_{i \in \mathcal{V}_{\text{train}}} \sum_{c=1}^C Y_{ic} \log \hat{Y}_{ic} \quad (5)$$

where N_{train} is the number of training nodes, C is the total number of categories, and Y_{ic} is the true label of node i for category c .

4 DATASETS AND EXPERIMENTAL SETTINGS

4.1 DATASETS INTRODUCTION

In order to verify the generalization performance of the proposed model on different complex correlation datasets, six classic heterogeneous graph benchmark datasets are selected for experiments, including three homophilic citation datasets (Cora, Citeseer, Pubmed) and three heterophilic web graph datasets (Cornell, Texas, Wisconsin). The detailed statistical information of the datasets is shown in Table 1.

TABLE 1. STATISTICAL INFORMATION OF COMPLEX CORRELATION DATASETS

Dataset	Nodes	Edges	Feature Dimension	Classes	Graph Type
Cora	2708	5429	1433	7	Homophilic Citation Graph
Citeseer	3312	4732	3703	6	Homophilic Citation

					Graph
Pubmed	19717	44338	500	3	Homophilic Citation Graph
Cornell	183	298	1703	5	Heterophilic Web Graph
Texas	187	328	1703	5	Heterophilic Web Graph
Wisconsin	251	466	1703	5	Heterophilic Web Graph

4.2 EXPERIMENTAL SETTINGS

All experiments are implemented based on PyTorch and PyTorch Geometric framework. The dataset split ratio is set as 48% training set, 32% verification set, and 20% test set according to the standard graph experiment setting. The proposed model adopts 2-layer adaptive graph convolution structure, the hidden layer dimension is set to 64, the learning rate is 0.01, the weight decay is $5e-4$, and the dropout rate is 0.5. The baseline models include GCN, GAT, GraphSAGE, and GKAN. All models are trained for 200 epochs with early stopping strategy to prevent overfitting. The node classification accuracy is used as the main evaluation metric.

5 EXPERIMENTAL RESULTS AND ANALYSIS

5.1 NODE PREDICTION PERFORMANCE COMPARISON

The prediction accuracy of all models on six datasets is shown in Table 2. All experimental results are the average values of 10 repeated experiments to ensure the reliability of the results.

TABLE 2. NODE CLASSIFICATION PREDICTION ACCURACY (%) OF DIFFERENT MODELS

Model	Cora	Citeseer	Pubmed	Cornell	Texas	Wisconsin	Avg Acc
GCN	81.5 ±1.2	70.3 ±1.5	78.6 ±0.8	68.2 ±2.1	70.5 ±1.9	69.8 ±2.3	73.15
GAT	82.7 ±1.0	71.6 ±1.3	79.2 ±0.7	69.5 ±1.8	71.8 ±1.6	70.6 ±2.0	74.23
GraphSAGE	83.1 ±0.9	72.2 ±1.1	80.1 ±0.6	70.2 ±1.5	72.3 ±1.4	71.5 ±1.8	74.90
GKAN	84.2	73.5	81.3	71.6	73.5	72.8	76

N	±0.8	±1.0	±0.5	±1.3	±1.2	±1.6	.1 5
Propo sed Model	87.5 ±0.6	76.8 ±0.8	84.6 ±0.4	75.3 ±1.1	77.2 ±1.0	76.5 ±1.3	79 .6 5

5.2 RESULT ANALYSIS

It can be seen from Table 2 that the proposed model achieves the best prediction performance on all six complex correlation datasets. Compared with the optimal baseline model GKAN, the average accuracy is improved by 3.50%. Specifically, the performance improvement on heterophilic web graph datasets (Cornell, Texas, Wisconsin) is more significant, with an average improvement of 4.2%, which verifies that the adaptive weighted aggregation mechanism can effectively capture the complex heterogeneous correlation between nodes.

Traditional GCN and GAT have poor performance on heterophilic datasets, because they adopt fixed aggregation rules and cannot adapt to the discrete feature distribution and inconsistent correlation rules of heterogeneous graphs. GraphSAGE and GKAN improve the feature mining ability by optimizing the sampling and kernel structure, but they lack the differentiated perception of node correlation strength. The proposed model calculates the adaptive neighbor weight based on node feature similarity, which effectively distinguishes the contribution of different neighbor nodes to feature learning, and the residual connection mechanism avoids feature loss in deep network propagation, thus significantly improving the quality of mined graph features and node prediction accuracy.

5.3 ABLATION EXPERIMENT

In order to verify the effectiveness of each improved module of the model, ablation experiments are carried out on the Cora and Texas datasets, and the results are shown in Table 3. Where, Base represents the vanilla 2-layer GCN model, AW represents adaptive weighted aggregation, and RC represents residual connection.

TABLE 3. ABLATION EXPERIMENT RESULTS

Model Variant	Cora Acc (%)	Texas Acc (%)
Base	81.5	70.5
Base+AW	85.2	74.1
Base+RC	83.6	72.8
Base+AW+RC (Proposed)	87.5	77.2

The ablation results show that both the adaptive weighted aggregation module and the residual connection module can effectively improve the model prediction performance. The adaptive weighted aggregation module plays a major role in improving feature mining ability, while the residual connection module optimizes the network training process. The combination of the two modules

achieves the best performance, which fully verifies the rationality and effectiveness of the model improvement strategy.

6 CONCLUSION AND FUTURE WORK

Aiming at the problems of insufficient feature mining and low node prediction accuracy of traditional GNNs in complex correlation datasets, this paper proposes an improved graph neural network model with adaptive neighbor aggregation and residual connection. The model realizes fine-grained mining of node attribute features and topological correlation features by constructing adaptive weighted graph convolution formula, and solves the gradient disappearance and feature loss problems of deep graph networks through residual connection optimization. Experimental results on multiple homophilic and heterophilic complex graph datasets show that the proposed model outperforms mainstream baseline models in node prediction tasks, with strong generalization and robustness.

Future research will focus on two aspects: first, introduce higher-order topological correlation learning to further mine implicit long-distance node correlations; second, optimize the model structure to reduce computational complexity and improve the scalability for large-scale ultra-large complex graph datasets.

ACKNOWLEDGMENTS

Not Applicable.

FUNDING

Not Applicable.

INSTITUTIONAL REVIEW BOARD STATEMENT

Not Applicable.

INFORMED CONSENT STATEMENT

Not Applicable.

DATA AVAILABILITY STATEMENT

Not Applicable.

CONFLICT OF INTEREST

Not Applicable.

PUBLISHER'S NOTE

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

AUTHOR CONTRIBUTIONS

Not application.

ABOUT THE AUTHORS

XU, Wenhao

360 Huntington Ave, Boston, MA 02115

REFERENCES

- [1] Yin, D., Jiang, R., Deng, J., Li, Y., Xie, Y., Wang, Z., ... & Shang, J. S. (2023). MTMGNN: Multi-time multi-graph neural network for metro passenger flow prediction. *GeoInformatica*, 27(1), 77-105.
- [2] Liu, T., & Zhang, J. (2023). An adaptive traffic flow prediction model based on spatiotemporal graph neural network: T. Liu, J. Zhang. *The Journal of Supercomputing*, 79(14), 15245-15269.
- [3] Wang, Y., Wang, Z., Wang, T., Hu, J., You, Z., & Peng, J. (2026). HiGLDP: a hierarchical graph neural network for predicting lncRNA-disease associations through multi-omic integration. *BMC biology*, 24(1), 80.
- [4] Li, C., Liu, X., Yan, Y., Zhao, Z., & Zeng, Q. (2023). HetGNN-SF: Self-supervised learning on heterogeneous graph neural network via semantic strength and feature similarity: C. Li et al. *Applied Intelligence*, 53(19), 21902-21919.
- [5] Yu, Z., Li, B., Zhang, F., You, J., Li, B., Chen, Z., ... & Xiao, K. (2026). GTF: A New Interpretable Graph Neural Network for Geochemical Anomaly Detection in Mineral Prospectivity Mapping. *Natural Resources Research*, 35(2), 797-822.
- [6] Wang, Y., Rui, L., & Ma, J. (2023). A short-term residential load forecasting scheme based on the multiple correlation-temporal graph neural networks. *Applied Soft Computing*, 146, 110629.
- [7] Chen, Z., Cen, G., Wei, Y., & Li, Z. (2023). Student performance prediction approach based on educational data mining. *IEEE Access*, 11, 131260-131272.
- [8] Shi, J., Zhang, H., Zhou, J., & Tang, L. (2026, May). The Impact of Working Capital Management Efficiency on Corporate Financial Performance: A Panel Data Regression Analysis. In 2026 5th International Conference on Social Sciences and Humanities and Arts (SSHA 2026) (pp. 1350-1355). Atlantis Press.
- [9] Li, C., Guo, L., Zhang, H., Tang, L., Shi, J., & Zhou, J. (2026). Related-Party Transaction Fraud Detection with Data-Governance-Enhanced Graph Neural Networks. Available at SSRN 6610498.
- [10] Li, X., Fang, L., & Gao, Z. (2026, May). Features and Mechanisms of Computer-Mediated Language Information Interaction in Cross-Cultural Contexts. In 2026 5th International Conference on Social Sciences and Humanities and Arts (SSHA 2026) (pp. 979-985). Atlantis Press.
- [11] Cao, N., Zhang, W., Leng, N., Guo, Y., Yu, Z., He, Q., & Tang, H. (2026, February). Research on the Input-Output Efficiency Model of Robotic Process Automation (RPA) in Financial Supply Chains. In *Proceedings of the 2nd International Conference on Digital Management and Information Technology* (pp. 180-185).
- [12] Ni, H., Zhang, T., Liu, H., He, Z., Fan, Y., Zhang, W., & Zhang, H. (2026, January). Research and Design of Blockchain-based Propagation Algorithm for IP Tags Using Local Random Walk in Big Data Marketing. In *2026 International Conference on Communication Networks and Machine Learning (CNML)* (pp. 284-288). IEEE.