Biomedical Data Science
2025, VOL. 5, NO. 2, 35-40
DOI: 10.47297/wspbdsWSP2752-630506.20250502

Optimization of Health Recommendation Based on Graph Neural Network and Transformer

Xiaolong Luo

Business School, Southwest University, Chongqing 402460, China

ABSTRACT
With the swift progress of intelligent health management, health recommendation systems have become a crucial part of the healthcare area. Yet, traditional recommendation systems have many issues, such as data sparsity, inaccurate feature extraction, and insufficient personalization. This paper presents an optimized health recommendation algorithm based on the Transformer model, aiming to enhance the accuracy and user experience of health recommender systems. It combines the strengths of Graph Neural Networks (GNNs) and the Transformer model. By aggregating users' historical behavior data, it can capture potential features in health information and improve recommendations. Experimental results show that the proposed algorithm outperforms traditional models in recommendation accuracy and user satisfaction.

KEYWORDS
Transformer; Graph Neural Network; Health Recommendation; Attention Mechanism

1 Introduction

In the healthcare field, personalized health recommendation systems are vital for boosting users' health management efficiency. With digital health data expanding rapidly, how to extract valuable knowledge and recommend it precisely based on users' needs is a key research focus. Yet, traditional recommendation systems face challenges like data sparsity, complex user behaviors, and highly specialized health information[1].

Recent deep learning advances have offered new ideas for recommendation system optimization. Specifically, GNN - based methods can model the complex links between users and health information, and the Transformer model, with its strong feature extraction, has delivered good results in NLP and sequence modeling tasks[2]. Consequently, this paper proposes an optimized health recommendation algorithm integrating GNNs and Transformers to enhance recommendation accuracy and user experience.

The main contributions of this project are as follows:

(1) Combining GNNs' graph - structure learning and Transformers' sequence - modeling strengths, this study propose an optimized health recommendation method.

(2) Considering health information's unique nature, an improved feature aggregation mechanism is designed to capture user preferences more precisely.

(3) Experiments prove the method's effectiveness. Compared with mainstream recommendation algorithms, it shows superiority in recommendation accuracy and user satisfaction.

2 Literature review

2.1 Traditional recommendation algorithm

Traditional recommendation systems mainly include collaborative filtering (CF), content - based recommendation, and hybrid recommendation. CF recommends based on user or item similarities, with typical algorithms like k - nearest neighbors (KNN) and singular value decomposition (SVD) [3]. However, CF is vulnerable to data sparsity and struggles with cold - start problems. Content - based recommendation matches user interests by analyzing health - information features (e.g., text descriptions, tags), yet it relies on feature engineering and fails to capture deep semantics. Hybrid recommendation combines CF and content - based methods to enhance recommendation robustness and personalization[1]. Still, traditional recommendation algorithms, while effective in scenarios with sufficient data, single data sources, and simple feature extraction, have significant limitations in handling the temporality, cold - start, data sparsity, and dynamic demand changes in user health data. Consequently, they increasingly fall short of the actual application needs in complex health - care scenarios.

Correspondence: a18575146568@163.com

2.2 Recommendation Based on Graph Neural Networks

Graph Neural Networks (GNNs) have gained significant attention in recommendation systems recently. Since user - item relationships can be represented as graph structures, GNNs can effectively capture complex interaction patterns, improving recommendation quality. GNN - based recommendation methods, such as Graph Convolutional Networks (GCNs) and Graph Attention Networks (GATs) [4], can propagate information on the user - health information interaction graph, enabling finer - grained feature learning. In health recommendations, GNNs can mine the associations in health information, like users' medical records and health - knowledge graphs, making recommendations more precise. However, GNNs have limitations, such as difficulty in handling long - distance dependencies and limited sequence - data modeling capabilities, necessitating further optimization[5].

2.3 Recommendation Based on Transformer

The Transformer model initially achieved breakthroughs in natural language processing and has been increasingly applied to recommendation systems, particularly excelling in sequential recommendation tasks[6]. The Transformer, based on self - attention mechanisms, can model user - behavior sequences and capture long - range dependencies effectively. In health recommendation system, it can process time - series health data, such as user health - behavior records and symptom - trend data, making recommendations more personalized[7]. Despite its good performance, the Transformer has high computational complexity and room for improvement in structured - data modeling.

Thus, using GNNs to capture complex user - health information relationships and mine potential associations, and Transformers to process time - series data and identify health - data patterns, can enhance recommendation accuracy and better meet users' diverse health needs.

3 Health Recommendation Based on GNNs and Transformer

3.1 Problem Definition

Given a user list $U=\{u_1,u_2,\cdots,u_S\}$ and a candidate health - information list $I=\{d_1,d_2,\cdots,d_M\}$,where S is the total number of users and M is the total number of health information items. The historical interaction behaviors of users can be represented by a two - dimensional matrix:

$$X\in R^{S\times M}\# \tag{1}$$

In matrix X, the element x_{ij} indicates the interaction between user u_i and health information d_j. For example, if $x_{i_j}=1$, user u_i has viewed or interacted with health information d_j; if $x_{i_j}=0$, no interaction has occurred. Additionally, user ratings of health information can be represented by another two - dimensional matrix:

$$E\in R^{S\times M}\# \tag{2}$$

In matrix E, the element e_{i_j} represents user u_i's rating of health information d_j, typically ranging from 1 to 5. A rating of $e_{i_j}=0$ means user u_i has not rated health information d_j.

The model aims to generate a personalized health recommendation list for each user u_i. Specifically, the system predicts user interest in un interacted health information d_j based on historical behavior and rating data, then generates a personalized recommendation list from these predictions.

3.2 Overall Architecture

The overall architecture of the health - recommendation model based on GNNs and Transformer is shown in Figure 1. The model takes the user - health - information interaction matrix X and rating matrix E as input. First, the GNN module aggregates neighbor - node information to build a bipartite graph of users and health information, mapping input features to high - dimensional embedding vectors for the Transformer module. Then, the Transformer module uses self - attention to learn complex interactions and multi - head attention to focus on diverse user interests, alleviating data sparsity. Finally, the model fuses GNN and Transformer output features into a joint representation. This is mapped to a recommendation score space via a fully connected layer, forming a Top-N ranking optimized by a cross - entropy loss function.

3.3 Graph Neural Network Module

In health recommendation systems, users and health information are regarded as nodes in a graph, with interaction relationships serving as connecting edges. Suppose there are S users and M health information items; they form a bipartite graph where the adjacency matrix X is an $S \times M$ matrix. GNN updates node representations through multi - layer graph convolution operations. At the l - th layer, the representation of node u_i, $h_i^{(l)}$, is updated using:

$$h_i^{(l)} = ReLU\left(W^{(l)} \cdot \sum_{j \in \mathcal{N}(i)} \alpha_{ij} \cdot h_j^{(l-1)} + b^{(l)}\right) \# \tag{3}$$

ReLU() is the activation function, and $W^{(l)}$ and $b^{(l)}$ are the weight matrix and bias term of the l-th layer, respectively. $\mathcal{N}(i)$ denotes the set of health - information nodes connected to node u_i(a user). By stacking multiple GNN layers, nodes can aggregate information from neighbors, capturing complex user - preference patterns (see Figure 2).

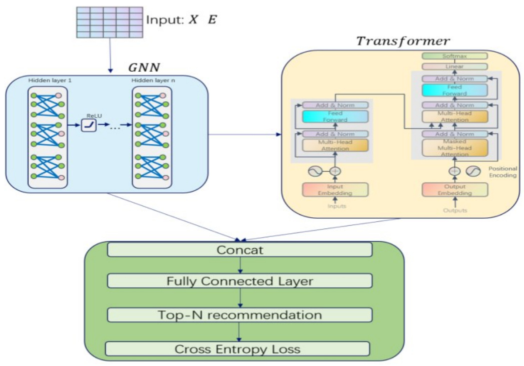

Figure 1 Overall Architecture

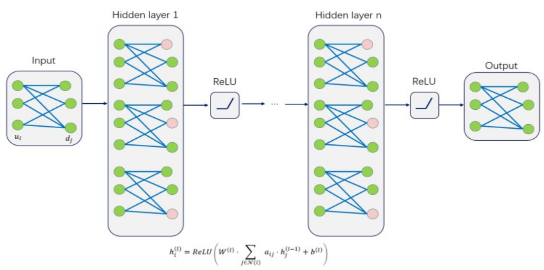

Figure 2 GNN Architecture Diagram

3.4 Transformer Module

The Transformer model processes sequential data through a self - attention mechanism, capturing long - range dependencies in users' historical behavior sequences. In health recommendation, user historical behavior data is typically time - series data, such as browsing history. Transformer can effectively model this sequential data, enhancing the model's ability to adapt to dynamic changes in user interests (see Figure 3).

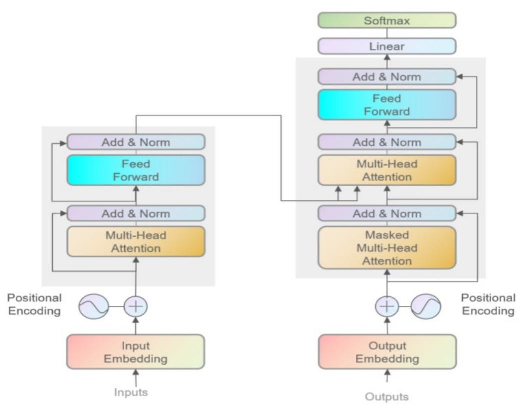

Figure 3 Transformer Architecture Diagram

The self - attention computation consists of:

(1)Calculating Query (Q), Key (K), and Value (V)

$$Q = r \cdot W^Q, K = r \cdot W^K, V = r \cdot W^V \# \tag{4}$$

(2)Computing attention weights

$$A = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) \# \tag{5}$$

(3)Weighted sum calculation

$$Attention\left(Q,K,V\right)=A\cdot V\#\tag{6}$$

r is the user's behavioral - representation matrix, containing recent health - information interaction data. W^Q,W^K,W^V are learnable weight matrices for Q, K, V calculation. d_k is the key - vector dimension for scaling. A is the attention - weight matrix indicating the correlation between each query - key pair.

(4)Multi - head attention mechanism

Transformer uses multi - head attention to boost expressiveness (see Figure 4). The formula is:

$$\text{MultiHead}\left(Q,K,V\right)=\text{Concat}\left(\text{ head }_1,\text{head }_2,\cdots,\text{ head }_h\right)W^O\#\tag{7}$$

$\text{head}_i=\text{Attention}\left(QW_i^Q,KW_i^K,VW_i^V\right)$ is the i-th attention head. W^O is the output weight matrix.。

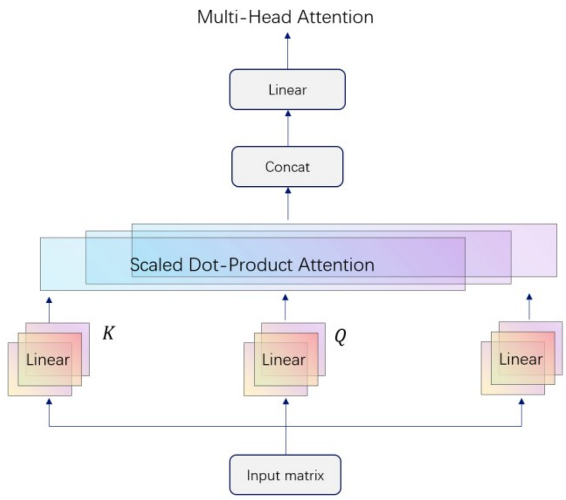

Figure 4 Multi - Head Attention Mechanism Diagram

(5)Positional Encoding

As Transformer cannot process sequence order, positional encoding is applied to introduce position - related information. The formula is:

$$PE_{(pos,2i)}=\sin\left(\frac{pos}{10000^{2i/d}}\right),PE_{(pos,2i+1)}=\cos\left(\frac{pos}{10000^{2i/d}}\right)\#\tag{8}$$

pos is the current position index, i is the dimension index, and d is the positional - encoding dimension.

3.5 Model Fusion and Optimization

In this study, the output features of GNN and Transformer are fused to optimize health - information recommendation. The process is as follows:

(1)Feature Fusion: Concatenate the output features of GNN and Transformer for a joint representation:

$$h_{\text{fused}}=\text{Concat}\left(h_{\text{GNN}},h_{\text{Transformer}}\right)\#\tag{9}$$

(2)Fully Connected Layer: Map the fused features to a recommendation - score space via a fully connected layer:

$$\hat{y}=W_{\text{final}}\cdot h_{\text{fused}}+b\#\tag{10}$$

(3)Rating Prediction: Generate personalized health - information recommendation lists based on the predicted scores.

3.6 Loss Function and Training

For model training, a cross - entropy loss function is adopted to optimize the prediction results:

$$L=-\left(\sum_{X\in\Delta^+}y\ln\hat{y}+\sum_{X\in\Delta^-}(1-y\ln(1-\hat{y})\right)\#\tag{11}$$

Δ^+ and Δ^- denote positive and negative sample sets, $\hat{y}$ is the predicted click probability, and y is the actual click label.

4 Experimental Results and Analysis

This study utilized a public dataset (HealthCare Dataset) encompassing users' health - information browsing history and associated metadata. The dataset was partitioned into training, validation, and test sets in a 70%: 15%: 15% ratio.

Evaluation metrics comprised Accuracy, Recall, F1 score, and AUC. Experiments on this dataset assessed the health - recommendation system integrating GNN and Transformer, contrasting it with traditional recommendation algorithms (see Figure 5).

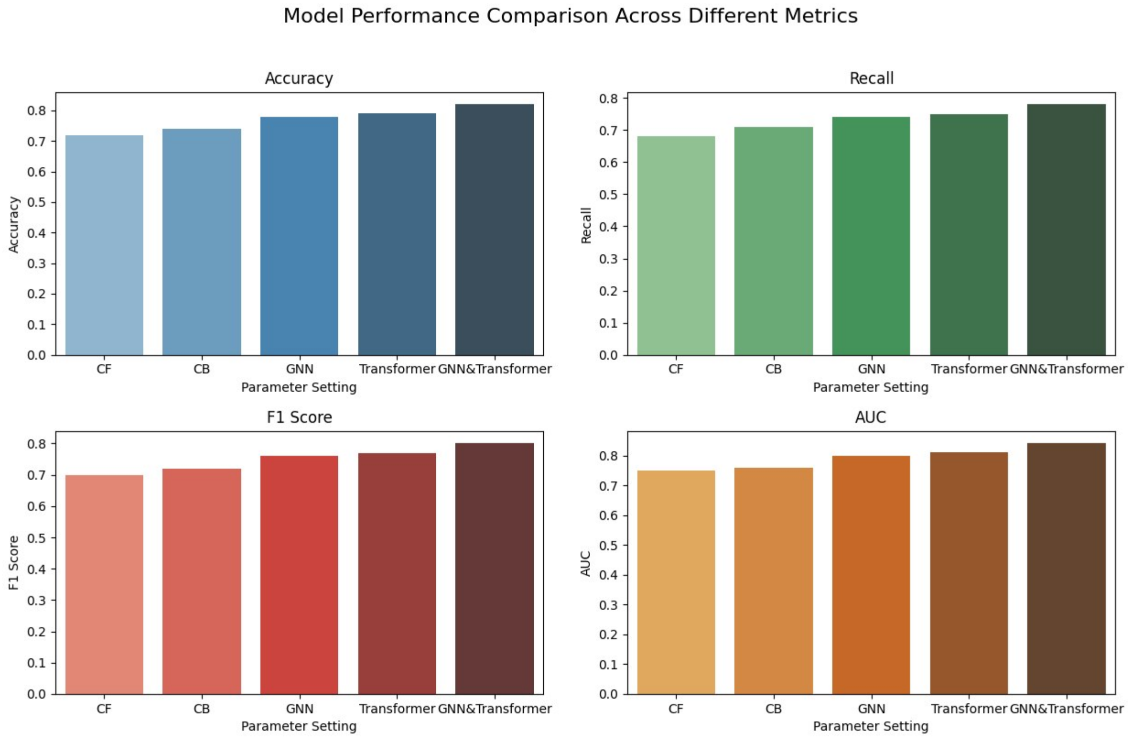

Figure 5 Comparison of Different Models' Evaluation Metrics

As shown in the results, the model combining GNN and Transformer outperforms others in all evaluation metrics, indicating its better performance. This suggests that using GNN to capture the complex interactions between users and health information, and leveraging Transformer to process user historical behavior sequences, can effectively enhance the accuracy and personalization of recommendation systems.

5 Discussion

Compared to traditional recommendation algorithms, the fusion model based on GNN and Transformer can better capture users' long - term and short - term interests. GNN effectively models the interaction graph between users and health information, capturing complex user - preference patterns; Transformer handles long - range dependencies in user historical - behavior sequences. This fusion approach leverages the strengths of both models, enhancing recommendation - system performance. To maintain recommendation accuracy while reducing computational complexity, techniques like model pruning (removing neurons with minimal performance impact) or knowledge distillation (transferring knowledge from a large model to a smaller one) can be introduced to cut down computational overhead[8]. In practical applications, it is recommended to employ differential privacy techniques to safeguard the privacy of user health data, ensuring the security of user information. During the personalized - recommendation process, encryption and privacy - protection methods can be utilized to guarantee personalization without leaking sensitive user data[9].

The health - information recommendation optimization model based on GNN and Transformer proposed in this research outperforms traditional recommendation algorithms on multiple evaluation metrics. Future research can further explore model - optimization strategies, such as incorporating more sophisticated attention mechanisms (e.g., combining multi - head self - attention with local attention), multimodal data fusion, reinforcement learning, adaptive - recommendation mechanisms, and enhancing explainability and transparency, to further boost recommendation - system performance and user experience.

References

[1] Yue W., Wang Z., Zhang J., & Liu X. (2021). An overview of recommendation techniques and their applications in healthcare. IEEE/CAA Journal of Automatica Sinica, 8(4), 701-717.

[2] Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

[3] Karypis G. (2001, October). Evaluation of item-based top-n recommendation algorithms. In Proceedings of the tenth international conference on Information and knowledge management (pp. 247-254).

[4] W. Ying et al., "Graph neural networks in recommender systems: A survey," ACM Computing Surveys, vol. 55, no. 9, pp. 1-36, 2023.

[5] Paul S. G., Saha A., Hasan M. Z., Noori S. R. H., & Moustafa A. (2024). A systematic review of graph neural network in healthcare-based applications: Recent advances, trends, and future directions. IEEE Access, 12, 15145-15170.

[6] Ma M., Ren P., Chen Z., Ren Z., Liang H., Ma J., & De Rijke M. (2023). Improving transformer-based sequential recommenders through preference editing. ACM Transactions on Information Systems, 41(3), 1-24.

[7] Li X., Yu X., Liu G., Fan X., Ge F., Zhao Y., & Zheng X. (2024). Transformer-based medication recommendation with a multiple graph augmentation strategy. Expert Systems with Applications, 257, 125091.

[8] Muralidharan S., Turuvekere Sreenivas S., Joshi R., Chochowski M., Patwary M., Shoeybi M., ... & Molchanov, P. (2024). Compact language models via pruning and knowledge distillation. Advances in Neural Information Processing Systems, 37, 41076-41102.

[9] Hu H., Dobbie G., Salcic Z., Liu M., Zhang J., Lyu L., & Zhang X. (2023). Differentially private locality sensitive hashing based federated recommender system. Concurrency Computation, 35(14), 1-16.