FD-GATDR: A Federated-Decentralized-Learning Graph Attention Network for Doctor Recommendation Using EHR

LUNING BI, Iowa State University, USA
YUNLONG WANG, IQVIA, USA
FAN ZHANG, IQVIA, USA
ZHUQING LIU, The Ohio State University, USA
YONG CAI, IQVIA, USA
EMILY ZHAO, IQVIA, USA

In the past decade, with the development of big data technology, an increasing amount of patient information has been stored as electronic health records (EHRs). Leveraging these data, various doctor recommendation systems have been proposed. Typically, such studies process the EHR data in a flat-structured manner, where each encounter was treated as an unordered set of features. Nevertheless, the heterogeneous structured information such as service sequence stored in claims shall not be ignored. This paper presents a doctor recommendation system with time embedding to reconstruct the potential connections between patients and doctors using heterogeneous graph attention network. Besides, to address the privacy issue of patient data sharing crossing hospitals, a federated decentralized learning method based on a minimization optimization model is also proposed. The graph-based recommendation system has been validated on an EHR dataset. Compared to baseline models, the proposed method improves the AUC by up to 6.2%. And our proposed federated-based algorithm not only yields the fictitious fusion center’s performance but also enjoys a convergence rate of $O(1/T)$.

CCS Concepts: • Information systems → Collaborative filtering.

Additional Key Words and Phrases: graph learning, doctor decentralized recommendation, federated learning, heterogeneous data

1 INTRODUCTION

Electronic health record (EHR) system has been growing rapidly in the past decade. EHR contains patients’ medical information, history, diagnoses, medications, treatment plans, and laboratory and test results. It provides automate and streamline workflow to facilitate the decision-making of providers. Benefiting from the progress in the machine learning area, a variety of deep learning techniques and frameworks have been applied to clinical applications including such as information extraction, representation learning, outcome prediction, phenotyping, and a de-identification [18]. This study aims to build a doctor recommendation system using patients’ EHR history.

There are different types of EHR code representation applied in the existing studies. Word2vec methods have gained more popularity recently [2] [3]. In the word-level vector representation methods, all possible meanings of a word are transformed as a single vector representation, ignoring the fact that the word represent differently in different context. To solve this problem, Alsentzer et al. proposed the clinic BERT, a context-based model that was trained on all notes from MIMIC III, which have achieved the state-of-art performance [1]. By using these embedding methods, patients can be represented by the sequence of clinic service codes. Techniques are ranged from autoencoders for diabetic nephropathy [7], CNNs for unplanned readmission prediction [12], LSTMs for heart failure prediction [10], and GRU networks for predicting mortality [16][18]. However, in these studies, EHR data were treated as flat-structured information. The relationship among
services, patients and doctors were ignored. And in terms of prediction tasks, there is lack of studies explored the application of EHR data mining to doctor recommendation. Compared to treatment recommendation, doctor recommendation needs more structured information due to some factors such as speciality and the patient’s preference. Therefore, a method that can learn underlying information from the hidden structure in EHR data is in in immediately needs for the doctor recommendation.

Standard machine learning approaches require centralizing the training data on one machine or in a data center. However, collecting clinical datasets from isolated medical centers is unpractical since Clinic data such as disease symptoms and medical recordings are highly sensitive [9]. Under this context, federated Learning (FL), which enables local client model to collaboratively learn a shared prediction model while keeping all the training data private, eliminates the need to store the data in the cloud. However, most of the existing works in FL are limited to systems with i.i.d. datasets and centralized parameter servers. In the real world, the EHR data is fully decentralized among hospitals. The data is non i.i.d. due to the geographic influence on human health. Nevertheless, each hospital can share de-identified, non-sensitive, and intermediate statistics with its neighborhood hospitals. In this study, a federated decentralized learning is proposed to improve the client models’ performance while satisfying the privacy requirements.

Our contributions can be summarized as follows:

- To extract the structured information from the EHR data, a heterogeneous graph consisting different types of nodes and edges is built.
- We propose a heterogeneous graph attention network (HGAT) considering time sensitivity and node heterogeneity for the representation of the patient, doctor and service.
- A federated decentralized learning algorithm is proposed to address the issue of data sharing among hospitals.
- The case study shows that the proposed graph model can achieve better performance than other baseline models. The federated decentralized learning algorithm can realize comparable performance compared to the global training.

2 DOCTOR RECOMMENDATION USING EHR

In EHR, each interaction between patients and doctors is trying to answer two questions: “what is happening” and “what happens next” [13]. The first question is about diagnosis of patients’ health status. It also provides the answer to the second question by indicating the future disease risk and corresponding treatments. Traditionally, this step is completed by experienced doctors, which means expensive communication cost and time cost. Under this setting, EHR, which contains detailed information of patient medical history, has become an alternative. Therefore, this paper focuses on developing an end-to-end prototype for recommending a doctor based on EHR of the patients. However, there are two major challenges, i.e., data heterogeneity and data privacy.

2.1 Data Heterogeneity

Although in some way EHR is similar to human natural language, EHR contains a variety of information such as billing codes, medical service codes, laboratory measurements, patient demographic information, doctor’s speciality, etc. Utilizing part of the EHR as flat-structured data for the representation learning can lose a lot of valuable information. This study is aimed to extract heterogeneous structured from the EHR data. The concept is introduced as follows.

**Definition 1.** *Heterogeneous Graph.* A heterogeneous graph can be represented by $G = (V, E, A, R, \phi, \varphi)$, where $V$ denotes the set of nodes and $E$ represents the set of edges, $A$ and $R$ represent the node and edge
types respectively, and $\phi$ and $\varphi$ denote the mapping functions $\phi: V \rightarrow A$ and $\varphi: E \rightarrow R$. Moreover, in HIN, $|A| + |R| \geq 3$.

For better understanding, Fig. 1 is illustrated as an example of EHR graph. There are three types of nodes: patient, doctor and service. The patient node have the attributes such as age, sexuality and location. The doctor node have the attributes such as speciality and location. The service nodes can be divided into three sub-level node types: diagnosis, procedure and product. Each type of service node are stored as text in EHR, for example, "px-RADIOGRAPHY: SHOULDER AND UPPER ARM". The text or service codes can be used for diagnosis, risk prediction and decision-making of the following services/treatments. The edges of the graph represents the interactions between nodes.

Fig. 1. An example of EHR graph. The service nodes can be divided into diagnosis, procedure and product.

Four edge types are considered in this paper: patient-service, doctor-service, service-service and patient-doctor.

- **Patient-services** means that the patient receives the service.
- **Doctor-service** characterizes the speciality of the doctor.
- **Service-service** represents the chronological order of services. For example, some products are usually prescribed together for a specific disease. And a diagnosis service will be followed the corresponding products and procedure, or even another diagnosis requirement. These structured information provides the basic logic of the doctor recommendation because it not only connects the patient and doctor but also connect the present with the future by predicting the next services.
- **Patient-doctor** is the target of doctor recommendation. The prediction of patient-doctor edge is based on the learned pattern from the aforementioned node attributes and the other three types of edge. In other words, doctor recommendation is based on the current status and future prediction. Based on the heterogeneous graph, the doctor recommendation considered system can be defined as follows.

**Definition 2.** **Doctor Recommendation (DR).** Given a EHR set of $\langle V_P, V_D, V_S, W, B \rangle$ where $V_D$ denotes the set of doctors, $V_S$ denotes the set of services, $V_P$ denotes the set of patients, $W$ represents the attributes of the objects, and $B$ represents the interactions between two types of objects. The DR system is aimed to predict the next couple of services $\langle V_S \rangle$ needed in future as well as the suitable doctors $\langle V_P \rangle$. For the implementation, a simple scheme is to mask out a most recent part, e.g., 20%, of the EHR data and use the doctors appeared in the masked data as the ground-truth label.

**2.2 Data Privacy**

The performance of GL models are largely determined by the amount of the training data. However, there are several challenges to this exercise. First, data privacy and security are sensitive in the
medical area. It is difficult to integrate the data collected and aggregated across hospitals. Second, the EHR data collected across hospitals are often non i.i.d. due to the environmental factors. That means using only one hospital data may influence the generalization ability of the model. This paper will address these two concerns using a novel training method.

3 FEDERATED-DECENTRALIZED-LEARNING GRAPH ATTENTION NETWORK
To solve the problems aforementioned, FD-GATDR, a federated-decentralized-learning graph attention network for doctor recommendation is proposed.

3.1 Service Embedding
The first step is to convert the text of services into the representation in the computer language. Current research borrows the concepts from NLP, in which service codes are viewed as “words” and patients or encounters as the “sentences”. Sequences of patient encounters can also be seen as the “documents”. Deriving vector-based representation of clinical concepts is a common way to reduce the dimension of the code space and reveal complex relationships between different types of nodes, such as word2vec [6], GloVe [11] and BERT [4]. In the case study, all the three methods have been tested. It should be noted that the service embedding is just the first step to represent service. The service representation will be further updated by the graph learning model which is introduced in the following section.

3.2 Heterogeneous Graph Attention Network (HGAT)
The basic idea of the proposed model HGAT is to learn meaningful and robust representations for different types of nodes. The input to the model is a set of node features, \( h = \{\vec{h}_1, \vec{h}_2, ..., \vec{h}_N\} \), \( \vec{h}_i \in \mathbb{R}^{F_{ti}} \), where \( N \) denotes the number of nodes, \( t_i \) represents the node type of the \( \vec{h}_i \) and \( F_{ti} \) is the number of features for the node type of the \( \vec{h}_i \).

**Neighborhood nodes sampling.** Different from other recommendation system, doctor recommendation is more time-sensitive, meaning that it should recommend the most needed services or doctors. It is assumed that the closer interactions have more prevalence with recommendation. Thus, instead of using first-order neighbor nodes or selecting nodes randomly, the neighbor nodes is selected based on the weight of edges using the roulette method. Specifically, the weights can be calculated based on the timestamp of the interaction using a custom-built linear or log method.

**Heterogeneous node representation.** To address the data heterogeneity issue, Eq. 1 is proposed to obtain the attention coefficients.

\[
\alpha_{ij} = \frac{\exp \left( \text{LeakyReLU} \left( \vec{a}^T \left[ Q\vec{h}_i \| Q\vec{h}_j \| V_{t_i,t_j} \right] \right) \right)}{\sum_{k \in N_i} \exp \left( \text{LeakyReLU} \left( \vec{a}^T \left[ Q\vec{h}_i \| Q\vec{h}_k \| V_{t_i,t_j} \right] \right) \right)}
\]

Where \( Q \) denotes input linear transformation’s weight matrix and \( V_{t_i,t_j} \) is a shared one-dimension vector that represents the attention between two nodes in the node type level. Compared the original version in [19], the attention between node types are specified to improve the model performance on heterogeneous graph.

Specifically, we repeat the attention for \( K \) times and concatenate the learned embeddings as the embedding:

\[
\vec{h}'_i = \sigma \left( \frac{1}{K} \sum_{k=1}^{K} \sum_{j \in N_i} \alpha_{kj}^k Q^k \vec{h}_j \right)
\]
Where $k$ represents concatenation, $\alpha_{ij}^k$ are normalized attention coefficients computed by the $k$-th attention mechanism, and $Q^k$ is the corresponding input linear transformation’s weight matrix.

### 3.3 Reconstruction Loss

In this paper, it is assumed that there is an edge between every two successive services in the EHR of a patient. The robustness of a service-service edge is determined by the frequency of the edge in different patients’ EHR. Moreover, instead of being an attribute of the doctor node, the doctor’s speciality is masked out and used as a prediction task by using the doctor embedding obtained by HGAT.

Therefore, three types of reconstruction loss are considered: doctor-specialty, service-service and patient-doctor as in Eq. 3. The weights of each loss term can be manually set or using a Bayesian task weight learner introduced in [20].

$$\mathcal{L} = \mathcal{L}_{ds} + \mathcal{L}_{ss} + \mathcal{L}_{pd}$$ (3)

Where $\mathcal{L}_{ds}$ denotes the reconstruction loss of doctor-specialty, $\mathcal{L}_{ss}$ denotes the reconstruction loss of service-service and $\mathcal{L}_{pd}$ denotes the reconstruction loss of patient doctor.

$\mathcal{L}_{ds}$ and $\mathcal{L}_{ss}$ can be calculated the cross entropy loss function since they are multi-class prediction tasks. For the patient-doctor reconstruction, a pairwise personalized ranking loss, namely, Bayesian personalized ranking (BPR) [15], is used. The BPR loss is maximizing the match degree between the patient and the doctor recommended (positive sample) while minimizing the match degree between the patient and the doctor not recommended (negative sample). The model performance is evaluated using recall and AUC.

### 3.4 Federated Decentralized Learning

To address the data privacy issue discussed in 2.2, a federated decentralized learning method is proposed. In the federated decentralized scenario, the hospitals form a peer-to-peer network system, which can be represented by an undirected connected graph $\mathcal{G} = (\mathcal{N}, \mathcal{L})$. Here, $\mathcal{N}$ and $\mathcal{L}$ are the sets of workers for local hospitals and the edges between hospitals, respectively, with $|\mathcal{N}| = m$.

The workers are capable of local computation and communicating with their neighboring workers via the edges in $\mathcal{L}$. The goal of fully decentralized FL is to have the workers distributively and collaboratively solving the global optimization problem in the following form:

$$\min_{x \in \mathbb{R}^p} \frac{1}{m} \sum_{i=1}^{m} f(x_i) = \min_{x \in \mathbb{R}^p} \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} f(x_i; \xi_j),$$ (4)

where each local objective function $f_i(x) \triangleq \frac{1}{n} \sum_{j=1}^{n} f_i(x; \xi_j)$ is only observable to worker $i$ and not necessarily convex. Here, $\mathcal{D}_i$ represents the distribution of the dataset at node $i$, which is heterogeneous across workers.

First, we reformulate Problem (4) in the following equivalent form by introducing a local model copy at each worker:

Minimize $$\frac{1}{m} \sum_{i=1}^{m} f(x_i)$$ (5)

subject to $x_i = x_{i'}, \quad \forall (i, i') \in \mathcal{L}$. 


Algorithm 1 Federated Decentralized Learning

1: Set $x_{i,0} = x^0$ and $y_{i,0} = g_{i,0} = \nabla f_i(x_{i,0})$ at worker $i$, for all $i \in [m]$.
2: for $k = 0, \ldots, K-1$ do
3:     for worker $i, i \in [m]$ do
4:         Share $(x_{i,k}, y_{i,k})$ with neighboring nodes;
5:         Consensus Update $x_{i,k} = \sum_{l \in N_i} [W]_{i,l} x_{l,k} - \gamma y_{i,k}$;
6:    if mod(k,q)=0 then
7:        Calculate $g_{i,k+1} = \nabla f_i(x_{i,k})$;
8:     else
9:        Calculate $g_{i,k+1} = g_{i,k} + \frac{1}{|S_{i,k}|} \sum_{j \in S_{i,k}} (\nabla f_i(x_{i,k+1}; \zeta(j)) - \nabla f_i(x_{i,k}; \zeta(j)))$;
10:    end if
11:    Gradient Tracking $y_{i,k+1} = \sum_{j \in N_j} [W]_{i,j} y_{j,k} + g_{i,k+1} - g_{i,k}$;
12: end for
13: end for

where $x \doteq [x_1^T, \ldots, x_m^T]^T$, and $x_i$ is an introduced local copy at worker $i$. To solve Problem (5), we consider an $\epsilon^2$-stationary point $x$ defined as follows:

$$
\left\| \frac{1}{m} \sum_{i=1}^m \nabla f(x) \right\|^2 + \frac{1}{m} \sum_{i=1}^m \|x_i - \bar{x}\|^2 \leq \epsilon^2,
$$

where $\bar{x} \doteq \frac{1}{m} \sum_{i=1}^m x_i$ represents the global average across all workers. The first term is the gradient norm of the global loss function and the second term is the average consensus error across all local copies. In this work, we aim to develop an efficient algorithm to attain an $\epsilon^2$-stationary point for fully decentralized FL.

To solve Problem (4) in decentralized network systems where workers reach a consensus on a global optimal solution, a common approach in the literature is to let workers aggregate neighboring information through a consensus matrix $W \in \mathbb{R}^{m \times m}$. Let $[W]_{ij}$ represent the element in the $i$-th row and the $j$-th column in $W$. Then, a consensus matrix $W$ should satisfy the following properties:

(a) **Doubly Stochastic:** $\sum_{i=1}^m [W]_{ij} = \sum_{j=1}^m [W]_{ij} = 1$.
(b) **Symmetric:** $[W]_{ij} = [W]_{ji}, \forall i, j \in N$.
(c) **Network-Defined Sparsity Pattern:** $[W]_{ij} > 0$ if $(i, j) \in \mathcal{L}$; otherwise $[W]_{ij} = 0, \forall i, j \in N$.

The above properties imply that the eigenvalues of $W$ are real and can be sorted as $-1 < \lambda_m(W) \leq \cdots \leq \lambda_2(W) < \lambda_1(W) = 1$. We define the second-largest eigenvalue in magnitude of $W$ as $\lambda \doteq \max\{|\lambda_2(W)|, |\lambda_m(W)|\}$ for further notation convenience. It can be seen later that $\lambda$ plays an important role in the step-size selection and characterizing the algorithm’s convergence rate.

We start with stating the following assumptions:

**Assumption 1.** The objectives $f(\cdot)$ and $f_i(\cdot)$ satisfy:

1. $f(x)$ is bounded from below, i.e., there exists an $x^* \in \mathbb{R}^p$, such that $f(x) \geq f(x^*), \forall x \in \mathbb{R}^p$;
2. The function $f_i(x)$ is continuously differentiable and has $L$-Lipschitz continuous gradients, i.e., there exists a constant $L > 0$ such that $\|\nabla f_i(x_1) - \nabla f_i(x_2)\| \leq L\|x_1 - x_2\|_2, \forall x_1, x_2 \in \mathbb{R}^p$;

**Theorem 1 (Convergence of Defra).** Under Assumption 1, with a constant-level step-size $\gamma$, which satisfies
\[ \gamma \leq \min \left\{ \frac{1}{3}, \frac{1}{2} \sqrt{\frac{1-\lambda}{72mL}}, \frac{1}{2} \sqrt{\frac{1-\alpha}{40L}}, \frac{1-\alpha}{120L}, \frac{(1-\alpha)^2}{3}, \frac{1-\alpha}{6L}, \sqrt{\frac{1-\alpha}{12L^2}} \right\}. \]

in Algorithm 1 (See detailed requirements in our Appendix), then Algorithm 1 has the following convergence result:

\[ \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \left[ \left\| \nabla f(\bar{x}_k) \right\|^2 \right] + \frac{1}{m} \sum_{i=1}^{m} \left\| x_{i,k} - \bar{x}_k \right\|^2 = O\left( \frac{1}{T} \right) \]

4 CASE STUDY

In this section, the proposed method is compared with other baseline service representation methods, graph learning methods and training methods. All the models were trained using CPU AMD 5950X and GPU NVIDIA RTX 3090.

4.1 Dataset

In this study, a dataset of 238,846 claims of 1,005 patients was used to validate the proposed approach. In the graph model, there are 1,005 patient nodes, 2,233 service nodes, and 15,044 doctor nodes. The dataset was collected during 2010 to 2015. Each patient has a record range from 750 days to 1,750 days. The patient record length is distributed between 100 words and 1,500 words.

4.2 Comparisons with Other Service Representation Methods

To investigate which service embedding method is more suitable for this application, skip-gram, GloVe and BERT are implemented. Then the embedding are used as the input of a LSTM model for a binary classification of Alzheimer disease. The GloVe model is loaded with the pretrained parameters on Common Crawl which has 840B tokens and 2.2M vocab [11]. Each word is converted to a 300-dimension vector. Then the model is fine-tuned with LSTM for the classification. To boost the performance of BERT, Med-BERT [14], a contextualized embedding model pretrained on a structured EHR dataset of 28,490,650 patients, is used to obtain the service embedding. Then the BERT is fine-tuned with LSTM. For the train-test split, 65%, 15% and 20% of the 1005 patients are randomly selected as train, validation and test set.

Table 1 showed that BERT-LSTM substantially improves the prediction accuracy by up to 9%, boosting the AUC by up to 6%. It should be noted that only the service sequences of patients are used for this preliminary experiment. It proves the effectiveness of the context-based embedding method in the clinic application. In the following experiments, the graph information including patient attribute, doctor attribute and edge information is used. And BERT is used to obtain the service embedding.

Table 1. Comparisons among different service embedding methods for the binary classification of Alzheimer disease

| Methods          | Test recall | Test AUC |
|------------------|-------------|----------|
| Skip-gram-LSTM   | 0.53        | 0.54     |
| GloVe-LSTM       | 0.57        | 0.55     |
| BERT-LSTM        | **0.63**    | **0.61** |

4.3 Comparisons with Other Graph Learning Methods

To validate the effectiveness of our proposed model, we compare HGAT with three baseline models: graph convolutional network (GCN), graph neural network (GNN) and graph attention network (GAT). All the four models used 64 hidden neuron, 128 embedding dimension and 4 layers of the
modules. In GAT and HGAT, the number of head of the multi-head attention is set as 4. For the train-test split, 65%, 15% and 20% of the 1,005 patients are randomly selected as train, validation and test set. In the training of the model, only EHR in train set are used to generate the graph. In the validation and testing, the validation set and test set are incorporated into the train graph and generate the prediction. For each patient, the first 65% record are used to construct the graph and the remaining 35% is masked. In the testing, the model is aimed to predict the 5~10 ground-truth doctors from a set of 200~350 doctors.

The results are shown in Table 2. In terms of recall, HGAT can obtain 70%, meaning that the model can find 70% of the doctors appeared in the future record. We can see that although the HGAT is only 2% higher than GAT for test recall, the AUC of HGAT is 6% higher than that of GAT. And the performance of attention-based graph learning models, i.e., GAT and HGAT, is superior to the others.

Table 2. Comparisons among GCN, GNN, GAT and HGAT

| Methods   | Test recall | Test AUC |
|-----------|-------------|----------|
| GNN [17]  | 0.63        | 0.65     |
| GCN [8]   | 0.66        | 0.64     |
| GAT [19]  | 0.68        | 0.70     |
| HGAT      | 0.70        | 0.76     |

4.4 Comparisons with Global Training and Local Training

To simulate the situation that the EHR data is stored in different centers, we divided all the patients into six groups based on their location (state level). Each group have about 145, 158, 177, 207, 147 and 171 patients, respectively. It is assumed that the EHR data in each group is under an agreement and stored in one data center. A consensus matrix $W$ is generated to represent the connection between the six groups as in Eq. 7. The train-test split and labeling are the same as described in Sec. 4.3. The base model is HGAT used in 4.3.

$$ W = \begin{bmatrix} 0.64 & 0.18 & 0 & 0 & 0.18 & 0 \\ 0.18 & 0.64 & 0 & 0 & 0 & 0.18 \\ 0 & 0 & 0.82 & 0.18 & 0 & 0 \\ 0 & 0 & 0.18 & 0.64 & 0 & 0.18 \\ 0.18 & 0 & 0 & 0 & 0.82 & 0 \\ 0 & 0.18 & 0 & 0.18 & 0 & 0.64 \end{bmatrix} $$

The comparison results are shown in Table 3. Compared to the local training, the performance of FDL is superior. For example, for region 1, 5 and 6, the test recall and test AUC of FDL is 14%, 19% higher than that of the local training, respectively. The reason is that in the local training, the generalization ability of the model is limited due to the lack of train data. Compared to the global training, FDL can achieve a comparable performance. The difference in test recall and test AUC is between 2% and 5%. Our proposed FDL yields the fictitious fusion center’s performance. The proof of FDL convergence is given in the supplementary section.

5 CONCLUSION

Accurate doctor recommendation is the key to developing telehealth services and improving running efficiency of the healthcare system. This paper proposed a heterogeneous graph based
FD-GATDR: A Federated-Decentralized-Learning Graph Attention Network for Doctor Recommendation Using EHR

Table 3. Performance of federated decentralized learning on the six regions. Local: each local model is trained using the local graph only; Global: a fusion center is trained using the global graph. FDL: each local model is trained using proposed federated decentralized learning method, namely, Algorithm 1. The Recall and AUC are the model performance on the test dataset.

| Methods      | Recall | AUC  | Methods      | Recall | AUC  |
|--------------|--------|------|--------------|--------|------|
| Local-region1| 0.66   | 0.54 | Local-region4| 0.66   | 0.74 |
| Global-region1| 0.73  | 0.74 | Global-region4| 0.76  | 0.71 |
| FDL-region1  | 0.70   | 0.73 | FDL-region4  | 0.72   | 0.64 |
| Local-region2| 0.72   | 0.78 | Local-region5| 0.51   | 0.53 |
| Global-region2| 0.70  | 0.77 | Global-region5| 0.65  | 0.66 |
| FDL-region2  | 0.71   | 0.77 | FDL-region5  | 0.65   | 0.60 |
| Local-region3| 0.65   | 0.76 | Local-region6| 0.62   | 0.58 |
| Global-region3| 0.68  | 0.79 | Global-region6| 0.68  | 0.80 |
| FDL-region3  | 0.72   | 0.76 | FDL-region6  | 0.66   | 0.75 |

model for doctor recommendation using EHR. First, a latent vector representing the edge between two types of heterogeneous nodes is incorporated into the message passing. Second, to enrich the representation of the doctor, patient and service node, three relationships, i.e., service-service, doctor-speciality, and patient-doctor, are reconstructed simultaneously. Moreover, to solve the data privacy issue in the medical area, a federated decentralized learning is developed to help improve the local model performance. The results show that the proposed model FD-GATDR can achieve a high prediction accuracy. In future, we will use a bi-level optimization model to improve the model prediction accuracy as well as the communication efficiency for more complex heterogeneous graph.

REFERENCES

[1] Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. arXiv preprint arXiv:1904.03323 (2019).
[2] Tian Bai, Ashis Kumar Chanda, Brian L Egleston, and Slobodan Vucetic. 2017. Joint learning of representations of medical concepts and words from ehr data. In 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 764–769.
[3] Zhengping Che, Yu Cheng, Zhaonan Sun, and Yan Liu. 2017. Exploiting convolutional neural network for risk prediction with medical feature embedding. arXiv preprint arXiv:1701.07474 (2017).
[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018).
[5] Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. 2018. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. Advances in Neural Information Processing Systems 31 (2018).
[6] David Guthrie, Ben Allison, Wei Liu, Louise Guthrie, and Yorick Wilks. 2006. A closer look at skip-gram modelling.. In LREC, Vol. 6. Citeseer, 1222–1225.
[7] Takayuki Katsuki, Masaki Ono, Akira Koseki, Michiharu Kudo, Kyoichi Haida, Jun Kuroda, Masaki Makino, Ryosuke Yanagiya, and Atsushi Suzuki. 2018. Risk Prediction of Diabetic Nephropathy via Interpretable Feature Extraction from EHR Using Convolutional Autoencoder.. In MIE. 106–110.
[8] Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016).
[9] Songtao Lu, Yawen Zhang, Yunlong Wang, and Christina Mack. 2019. Learn electronic health records by fully decentralized federated learning. arXiv preprint arXiv:1912.01792 (2019).
[10] G Maragatham and Shobana Devi. 2019. LSTM model for prediction of heart failure in big data. Journal of medical systems 43, 5 (2019), 1–13.
[11] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 1532–1543.

[12] Trang Pham, Truyen Tran, Dinh Phung, and Svetha Venkatesh. 2016. Deepcare: A deep dynamic memory model for predictive medicine. In Pacific-Asia conference on knowledge discovery and data mining. Springer, 30–41.

[13] Trang Pham, Truyen Tran, Dinh Phung, and Svetha Venkatesh. 2017. Predicting healthcare trajectories from medical records: A deep learning approach. Journal of biomedical informatics 69 (2017), 218–229.

[14] Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. 2021. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. NPJ digital medicine 4, 1 (2021), 1–13.

[15] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2012. BPR: Bayesian personalized ranking from implicit feedback. arXiv preprint arXiv:1205.2618 (2012).

[16] Saranya Sankaranarayanan, Jagadeshwar Balan, Jesse R Walsh, Yanhong Wu, Sara Minnich, Amy Piazza, Collin Osborne, Gavin R Oliver, Jessica Lesko, Kathy L Bates, et al. 2021. Covid-19 mortality prediction from deep learning in a large multistate electronic health record and laboratory information system data set: Algorithm development and validation. Journal of medical Internet research 23, 9 (2021), e30157.

[17] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. 2008. The graph neural network model. IEEE transactions on neural networks 20, 1 (2008), 61–80.

[18] Benjamin Shickel, Patrick James Tighe, Azra Bihorac, and Parisa Rashidi. 2017. Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. IEEE journal of biomedical and health informatics 22, 5 (2017), 1589–1604.

[19] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. stat 1050 (2017), 20.

[20] Jiawei Zheng, Qianli Ma, Hao Gu, and Zhenjing Zheng. 2021. Multi-view Denoising Graph Auto-Encoders on Heterogeneous Information Networks for Cold-start Recommendation. In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. 2338–2348.
A SUPPORTING LEMMA

**Lemma 1 (Iterates Contraction).** The following contraction properties of the iterates hold:

\[
\| x_k - 1 \otimes x_k \|^2 \leq (1 + c_1) \lambda^2 \| x_{k-1} - 1 \otimes x_{k-1} \|^2 + (1 + \frac{1}{c_1}) \gamma^2 \| y_{k-1} - 1 \otimes y_{k-1} \|^2, \\
\| y_k - 1 \otimes y_k \|^2 \leq (1 + c_2) \lambda^2 \| y_{k-1} - 1 \otimes y_{k-1} \|^2 + (1 + \frac{1}{c_2}) \| g_k - g_{k-1} \|^2,
\]

where \( c_1 \) and \( c_2 \) are arbitrary positive constants. Additionally, we have

\[
\| x_k - x_{k-1} \|^2 \leq 8 \| x_{k-1} - 1 \otimes x_{k-1} \|^2 + 4 \gamma^2 \| y_{k-1} - 1 \otimes y_{k-1} \|^2 + 4 \gamma^2 m \| y_{k-1} \|^2.
\]

**Proof.** Define \( \widetilde{W} = W \otimes I_m \). First for the iterates \( x_k \), we have the following contraction:

\[
\| \widetilde{W} x_k - 1 \otimes x_k \|^2 = \| \widetilde{W} (x_k - 1 \otimes x_k) \|^2 \leq \lambda^2 \| x_k - 1 \otimes x_k \|^2.
\]

This is because \( x_k - 1 \otimes x_k \) is orthogonal to \( 1 \), which is the eigenvector corresponding to the largest eigenvalue of \( \widetilde{W} \), and \( \lambda = \max \{ |\lambda_2|, |\lambda_m| \} \). Recall that \( x_k = x_{k-1} - \gamma y_{k-1} \), hence,

\[
\| x_k - 1 \otimes x_k \|^2 = \| \widetilde{W} x_k - 1 \otimes x_k - \gamma y_{k-1} - 1 \otimes (x_{k-1} - \gamma y_{k-1}) \|^2
\]

\[
\leq (1 + c_1) \| \widetilde{W} x_k - 1 \otimes x_k \|^2 + (1 + \frac{1}{c_1}) \gamma^2 \| y_{k-1} - 1 \otimes y_{k-1} \|^2
\]

\[
\leq (1 + c_1) \lambda^2 \| x_k - 1 \otimes x_k \|^2 + (1 + \frac{1}{c_1}) \gamma^2 \| y_{k-1} - 1 \otimes y_{k-1} \|^2,
\]

where (a) is because of triangle inequality and (b) is from eqs.(11).

For \( y_k \), we have

\[
\| y_k - 1 \otimes y_k \|^2 = \| \widetilde{W} y_{k-1} + g_k - g_{k-1} - 1 \otimes (y_{k-1} + g_k - g_{k-1}) \|^2
\]

\[
\leq (1 + c_2) \lambda^2 \| y_{k-1} - 1 \otimes y_{k-1} \|^2 + (1 + \frac{1}{c_2}) \| g_k - g_{k-1} - 1 \otimes (g_k - g_{k-1}) \|^2
\]

\[
\leq (1 + c_2) \lambda^2 \| y_{k-1} - 1 \otimes y_{k-1} \|^2 + (1 + \frac{1}{c_2}) \| (I - \frac{1}{n} (11^\top) \otimes I) (g_k - g_{k-1}) \|^2
\]

\[
\leq (1 + c_2) \lambda^2 \| y_{k-1} - 1 \otimes y_{k-1} \|^2 + (1 + \frac{1}{c_2}) \| g_k - g_{k-1} \|^2,
\]

where (a) is due to \( \| I - \frac{1}{n} (11^\top) \otimes I \| \leq 1 \).

According to the updating

\[
\| x_k - x_{k-1} \|^2 = \| \widetilde{W} x_{k-1} - \gamma y_{k-1} - x_{k-1} \|^2
\]

\[
= \| (\widetilde{W} - I) x_{k-1} - \gamma y_{k-1} \|^2 \leq 2 \| (\widetilde{W} - I) x_{k-1} \|^2 + 2 \gamma^2 \| y_{k-1} \|^2
\]

\[
= 2 \| (\widetilde{W} - I)(x_{k-1} - 1 \otimes x_{k-1}) \|^2 + 2 \gamma^2 \| y_{k-1} \|^2
\]

\[
\leq 8 \| (x_{k-1} - 1 \otimes x_{k-1}) \|^2 + 4 \gamma^2 \| y_{k-1} - 1 \otimes y_{k-1} \|^2 + 4 \gamma^2 m \| y_{k-1} \|^2.
\]

\[\Box\]

B PROOF OF THEOREM

Next, we provide the proofs of Theorem 1.
where (a) follows from Lemma 1 in [5], (b) is telescoping the result of (a) over iterations.

Next, we bound the error of the gradient estimators as the follows:

The proof is as follows:

where (a) is because of Lipschitz continuous gradients of \( f \), (b) follows from the update rule of \( x \). (c), (d) and (f) are follow from the triangle inequality. (e) is because \( \tilde{y}_t = \tilde{g}_t \) and Lipschitz continuous gradients of \( f \).

Next, we bound the error of the gradient estimators as the follows:

Next, taking the expectation of both sides of eqs.(15) and plugging the result in eqs.(16), we have

(15)

(16)
we have

\[ \mathbb{E} f(\bar{x}_{k+1}) - \mathbb{E} f(\bar{x}_{(n_k-1)q}) \leq \frac{Y}{2} \mathbb{E} \sum_{t=(n_k-1)q}^{k} \| \nabla f(\bar{x}_t) \|^2 - (\frac{Y^2}{2}) \mathbb{E} \sum_{t=(n_k-1)q}^{k} \| \bar{y}_t \|^2 + \frac{Y L^2}{m} \sum_{t=(n_k-1)q}^{k} \sum_{i=1}^{m} \mathbb{E} \| x_{i,t} - \bar{x}_t \|^2 + \gamma \frac{Y}{2} \sum_{t=(n_k-1)q}^{k} \sum_{i=1}^{m} \sum_{j=(n_k-1)q}^{k} \frac{L^2}{|S_{i,t}|} \mathbb{E} \| x_{i,t+1} - x_{i,t} \|^2 \]

Next, telescoping eqs.17 over \( k \) from \( (n_k - 1) \) to \( k \) where \( k \leq n_k q - 1 \) and since \( q = |S_{i,t}| = \lfloor \sqrt{n} \rfloor \), we have

\[ \mathbb{E} f(\bar{x}_{k+1}) - \mathbb{E} f(\bar{x}_{(n_k-1)q}) \leq \frac{Y}{2} \mathbb{E} \sum_{t=(n_k-1)q}^{k} \| \nabla f(\bar{x}_t) \|^2 - (\frac{Y^2}{2}) \mathbb{E} \sum_{t=(n_k-1)q}^{k} \| \bar{y}_t \|^2 + \frac{Y L^2}{m} \sum_{t=(n_k-1)q}^{k} \sum_{i=1}^{m} \mathbb{E} \| x_{i,t} - \bar{x}_t \|^2 + \gamma \frac{Y}{2} \sum_{t=(n_k-1)q}^{k} \sum_{i=1}^{m} \sum_{j=(n_k-1)q}^{k} \frac{L^2}{|S_{i,t}|} \mathbb{E} \| x_{i,t+1} - x_{i,t} \|^2 \]

Next, combining the result of eqs.(18), eqs.(8)- (10) and telescoping over \( k \) from 0 to \( K \), we have

\[ \mathbb{E} f(\bar{x}_{K+1}) - \mathbb{E} f(\bar{x}_{0}) + \| x_{K} - 1 \otimes \bar{x}_0 \|^2 - \| x_{0} - 1 \otimes \bar{x}_0 \|^2 + \gamma \left[ \| y_{K} - 1 \otimes \bar{y}_0 \|^2 - \| y_{0} - 1 \otimes \bar{y}_0 \|^2 \right] \leq \frac{Y}{2} \sum_{k=0}^{K} \mathbb{E} \| \nabla f(\bar{x}_k) \|^2 - (\frac{Y^2}{2}) \sum_{k=0}^{K} \mathbb{E} \| \bar{y}_k \|^2 \]

\[ + \frac{Y L^2}{m} \sum_{k=0}^{K} \mathbb{E} \| x_k - \bar{x}_k \|^2 + \gamma \sum_{k=0}^{K} \frac{L^2}{|S_{i,t}|} \mathbb{E} \| x_{k+1} - x_k \|^2 \]

\[ + \frac{Y L^2}{m} \sum_{k=0}^{K} \sum_{i=1}^{m} \mathbb{E} \| x_{i,t} - \bar{x}_t \|^2 + \gamma \frac{Y}{2} \sum_{k=0}^{K} \sum_{i=1}^{m} \sum_{j=(n_k-1)q}^{k} \frac{L^2}{|S_{i,t}|} \mathbb{E} \| x_{i,t+1} - x_{i,t} \|^2 \]

\[ + \left( (1 + c_1) \lambda^2 - 1 \right) \sum_{k=0}^{K} \mathbb{E} \| x_{k+1} - 1 \otimes \bar{x}_k \|^2 + (1 + \frac{1}{c_1}) \gamma^2 \sum_{k=0}^{K} \mathbb{E} \| y_{k} - 1 \otimes \bar{y}_k \|^2 \]
\begin{align}
+ ((1 + c_2) \lambda^2 - 1) y \sum_{k=0}^{K} \mathbb{E}\|y_k - 1 \otimes \bar{y}_k\|^2 &+ (1 + \frac{1}{c_2}) y \sum_{k=0}^{K} \mathbb{E}\|g_{k+1} - g_k\|^2. \\
\end{align}

For \(\mathbb{E}\|g_k - g_{k-1}\|^2\) and let \(\nabla f_i = [\nabla f(x_{i,1})^\top, \cdots, \nabla f(x_{i,m})^\top]^\top\), we have

\begin{align}
\mathbb{E}\|g_k - g_{k-1}\|^2 &= \mathbb{E}\|g_k - \nabla_x f_t + \nabla_x f_t - \nabla_x f_{t-1} + \nabla_x f_{t-1} - v_{t-1}\|^2 \\
&\leq 3 \mathbb{E}\|g_k - \nabla_x f_t\|^2 + 3 \mathbb{E}\|\nabla_x f_t - \nabla_x f_{t-1}\|^2 + 3 \mathbb{E}\|\nabla_x f_{t-1} - g_{k-1}\|^2 \\
&\leq 0 + 3L^2 \mathbb{E}\|x_t - x_{t-1}\|^2 + 0.
\end{align}

Next, combing the result of eqs.(19), eqs.(20) and eqs.(14), we have

\begin{align}
\mathbb{E}(\bar{x}_{K+1}) - \mathbb{E}(\bar{x}_0) + \|x_K - 1 \otimes \bar{x}_K\|^2 - \|x_0 - 1 \otimes \bar{x}_0\|^2 \\
+ y\[\|y_K - 1 \otimes \bar{y}_K\|^2 - \|y_0 - 1 \otimes \bar{y}_0\|^2\] \\
\leq - \frac{Y}{2} \sum_{k=0}^{K} \mathbb{E}\|\nabla f(\bar{x}_k)\|^2 - \left(\frac{Y}{2} - \frac{L\gamma}{2}\right) \sum_{k=0}^{K} \mathbb{E}\|\bar{y}_k\|^2 \\
+ \frac{1}{2} \sum_{k=0}^{K} \mathbb{E}\|\nabla f(\bar{x}_k)\|^2 - \frac{2L\gamma}{2} \sum_{k=0}^{K} \mathbb{E}\|\bar{y}_k\|^2 \\
+ \sum_{k=0}^{K} \mathbb{E}\|x_k - \bar{x}_k\|^2 + ((1 + c_1) \lambda^2 - 1) \sum_{k=0}^{K} \mathbb{E}\|x_k - 1 \otimes \bar{x}_k\|^2 \\
+ (1 + \frac{1}{c_1}) \sum_{k=0}^{K} \mathbb{E}\|y_k - 1 \otimes \bar{y}_k\|^2 + ((1 + c_2) \lambda^2 - 1) \sum_{k=0}^{K} \mathbb{E}\|y_k - 1 \otimes \bar{y}_k\|^2 \\
&+ \left[([1 + \frac{1}{c_2}]y(3L^2 + 2\gamma L^2) \sum_{k=0}^{K} \mathbb{E}\|x_k - 1 \otimes \bar{x}_k\|^2 + 4\gamma^2 \|y_K - 1 \otimes \bar{y}_K\|^2 + 4\gamma^2 m \|\bar{y}_K\|^2\right] \\
&= - \frac{Y}{2} \sum_{k=0}^{K} \mathbb{E}\|\nabla f(\bar{x}_k)\|^2 - \left(\frac{Y}{2} - \frac{L\gamma}{2} - 4\gamma^2 m \left(1 + \frac{1}{c_2}\right)\gamma(3L^2 + 2\gamma L^2) \right) \sum_{k=0}^{K} \mathbb{E}\|\bar{y}_k\|^2 \\
+ \{(1 + c_1) \lambda^2 - 1 + \gamma L^2 + 8\left(1 + \frac{1}{c_2}\right)\gamma(3L^2 + 2\gamma L^2) \} \sum_{k=0}^{K} \mathbb{E}\|x_k - 1 \otimes \bar{x}_k\|^2 \\
+ \left[([1 + \frac{1}{c_2}]y(3L^2 + 2\gamma L^2) + 4\gamma^2 \left(1 + \frac{1}{c_2}\right)\gamma(3L^2 + 2\gamma L^2) \right) \sum_{k=0}^{K} \mathbb{E}\|y_k - 1 \otimes \bar{y}_k\|^2
\end{align}
where

\[
C_1 = -\left(\frac{Y}{2} - \frac{LY^2}{2} - 4Y^2m[(1 + \frac{1}{c_2})\gamma 3L^2 + \gamma L^2]\right)
\]

\[
C_2 = \{(1 + c_1)\lambda^2 - 1 + \gamma L^2 + 8[(1 + \frac{1}{c_2})\gamma 3L^2 + \gamma L^2]\}
\]

\[
C_3 = \{(1 + c_2)\lambda^2 - 1)\gamma + (1 + \frac{1}{c_1})\gamma^2 + 4\gamma^2[(1 + \frac{1}{c_2})\gamma 3L^2 + \gamma L^2]\}
\]

Next, let \(c_1 = c_2 = \frac{1}{\lambda} - 1\), we have

\[
C_1 = -(\frac{Y}{2} + \frac{LY^2}{2} + 4\gamma^2m[(1 + \frac{1}{c_2})\gamma 3L^2 + \gamma L^2]) \leq -\frac{Y}{2} + \frac{Y}{6} + \frac{Y}{6} = 0,
\]

where \(\gamma \leq \min\{\frac{1}{3}, \sqrt[3]{\frac{1}{24mL^2}}, \sqrt[6]{\frac{1}{120L^2}}\}.
\]

\[
C_2 = \{(1 + c_1)\lambda^2 - 1 + \gamma L^2 + 8[(1 + \frac{1}{c_2})\gamma 3L^2 + \gamma L^2]\}
\]

\[
= \lambda - 1 + \gamma L^2 + 8[\frac{1}{1 - \lambda}\gamma 3L^2 + \gamma L^2] \leq \frac{1}{5} - 1 + \frac{1}{5} + \frac{1}{5} = -\frac{1}{5},
\]

where \(\gamma \leq \min\{\frac{1}{5}, \frac{1}{40L^2}, \frac{1}{120L^2}\}.
\]

\[
C_3 = \{(1 + c_2)\lambda^2 - 1)\gamma + (1 + \frac{1}{c_1})\gamma^2 + 4\gamma^2[(1 + \frac{1}{c_2})\gamma 3L^2 + \gamma L^2]\}
\]

\[
= (\lambda - 1)\gamma + \frac{1}{1 - \lambda}\gamma^2 + 4\gamma^2[\frac{1}{1 - \lambda}\gamma 3L^2 + \gamma L^2]
\]

\[
\leq - (\lambda - 1)\gamma + \frac{(1 - \lambda)\gamma}{3} + \frac{(1 - \lambda)\gamma}{3} = 0,
\]

where \(\gamma \leq \min\{\frac{(1 - \lambda)^2}{3}, \frac{1 - \lambda}{6L}, \sqrt[3]{\frac{1}{12L^2}}\}.
\]

Thus, we can conclude that

\[
\mathbb{E}f(\hat{x}_{K+1}) - \mathbb{E}f(\hat{x}_0) + \mathbb{E}\|x_K - 1 \otimes \bar{x}_K\|^2 - \mathbb{E}\|x_0 - 1 \otimes \bar{x}_0\|^2
\]

\[
+ \gamma\mathbb{E}\|y_K - 1 \otimes \bar{y}_K\|^2 - \mathbb{E}\|y_0 - 1 \otimes \bar{y}_0\|^2 \leq -\frac{Y}{2} \sum_{k=0}^{K} \mathbb{E}\|\nabla f(\hat{x}_k)\|^2 - \frac{1}{5} \sum_{k=0}^{K} \mathbb{E}\|x_k - 1 \otimes \bar{x}_k\|^2.
\]

Define \(Q_T = \mathbb{E}f(\hat{x}_{T+1}) + \mathbb{E}\|x_T - 1 \otimes \bar{x}_T\|^2 + \gamma\mathbb{E}\|y_T - 1 \otimes \bar{y}_T\|^2\), then we have

\[
\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\|\nabla f(\hat{x}_t)\|^2 + \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\|x_t - 1 \otimes \bar{x}_t\|^2 \leq \frac{Q_T - Q_0}{T \min\{\frac{1}{3}, \frac{1}{2}\}} = O(\frac{1}{T}).
\]

Thus, we can conclude our result in Theorem. 1.

\[
\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\|\nabla f(\hat{x}_k)\|^2 + \frac{1}{m} \sum_{i=1}^{m} \mathbb{E}\|x_{i,k} - \bar{x}_k\|^2 = O(\frac{1}{T}),
\]
where $\gamma$ is a constant-level step-size, which satisfies
\[
\gamma \leq \min\left\{ \frac{1}{3L}, \sqrt{\frac{1-\lambda}{72mL^2}}, \sqrt{\frac{1-\lambda}{24mL^2}}, \frac{1}{5}, \frac{1-\lambda}{120L^2}, \frac{(1-\lambda)^2}{3}, \frac{1-\lambda}{6L}, \sqrt{\frac{1-\lambda}{12L^2}} \right\}.
\]