文章编号:1674-2974(2025)04-0124-11 DOI:10.16339/j.cnki.hdxbzkb.2025272

基于多域图神经网络的疾病预测模型

罗熹[1],刘洋[1,2],安莹[2†]
(1. 湖南警察学院 网络犯罪侦查湖南省普通高校重点实验室,湖南 长沙 410138;
2. 中南大学 大数据研究院,湖南 长沙,410083)

摘 要:电子病历数据类型多样以及时序不规则,现有的基于深度学习的方法在特征学习的过程中大多无法同时兼顾对不同类型临床数据间静态关联和就诊记录间动态时序依赖的有效捕获.针对该问题,本文提出了一种基于多域图神经网络的疾病预测模型.该方法首先利用一个结合编码级注意力和时间感知LSTM的时序特征学习模块获得患者每次就诊的初始特征表示.然后,根据就诊序列中不同就诊间的相关性和时间间隔信息分别构建了一个就诊亲和图和一个就诊时序图,并通过图卷积神经网络从图中挖掘就诊记录间的静态语义关联和动态时序依赖.最后,利用一个基于自注意力机制的多域特征融合模块将时序特征和语义关联特征结合起来得到最终的患者融合特征表示,用于患者未来的疾病预测.在两个真实临床数据集上的实验结果表明,本文方法超过其他现有的方法获得了更高的预测准确性.

关键词:电子病历;疾病预测;图神经网络;注意力机制
中图分类号:TP391 **文献标志码**:A

Disease Prediction Model Based on Multi-domain Graph Neural Network

LUO Xi[1], LIU Yang[1,2], AN Ying[2†]
(1. Key Laboratory of Network Crime Investigation of Hunan Provincial Colleges, Hunan Police Academy, Changsha 410138, China;
2. Big Data Institute, Central South University, Changsha 410083, China)

Abstract: Due to the characteristics of electronic medical records (EMRs), such as the diversity of data types and temporal irregularity inherent, most existing deep learning-based methods cannot simultaneously capture static correlations between different types of clinical data and dynamic temporal dependencies between visits during the feature learning process. To address this issue, this paper proposes a disease prediction model based on multi-domain graph neural network. In this model, a temporal feature learning module that combines code level attention and time aware LSTM is first utilized to obtain the initial feature representation of patient visits. Then, based on the correlation and time interval information between different visits, a visit affinity graph and a visit sequence graph are constructed, and a graph convolutional neural network is used to mine the static and dynamic semantic associations between visit records from these graphs. Finally, a multi-domain feature fusion module based on self-attention mechanism is utilized to combine temporal features and semantic association features to obtain the final patient

* **收稿日期**:2024-04-16
基金项目:湖南省教育厅科学研究重点项目(23A0702),Key Scientific Research Project of Hunan Provincial Department of Education (23A0702)
作者简介:罗熹(1980—),女,湖南长沙人,湖南警察学院副教授,硕士生导师
† 通信联系人,E-mail: anying@csu.edu.cn

fusion representation for future disease prediction. The experimental results on two real clinical datasets show that our method outperforms other existing methods and achieves higher prediction accuracy.

Key words: electronic medical records; disease prediction; graph neural network; attention mechanism

疾病风险预测因其在疾病防控和治疗等方面突出的临床意义一直是医疗卫生领域关注的重要研究课题.传统的基于病例队列研究的疾病预测方法耗时冗长,人力、物力投入巨大,难以满足服务临床应用的需求.随着大数据、人工智能技术的发展以及医院信息化程度的提高,各个医疗机构积累了丰富的临床电子病历(electronic medical records, EMR)数据,为研究人员提供了大量类型多样、易于获取的患者诊疗全流程临床信息.因此,如何利用深度学习技术对电子病历数据进行分析,挖掘患者疾病发展的潜在模式以支持疾病风险预测等相关临床应用的辅助决策,一直是该领域的研究热点之一.

为了从电子病历的历史诊疗记录中捕获相关临床事件之间的时序关联以获得患者健康状态的动态特征,卷积神经网络(CNN)、循环神经网络(RNN)及其变体(LSTM 和 GRU)等深度模型已经得到了广泛的应用[1-4].然而,这些模型通常难以有效地捕获就诊记录间的长期依赖关系,在信息传播过程中容易忽略或丢失较早发生的历史临床信息,无法形成患者临床记录的完整的全局表示.尽管在后续的研究中,涌现了一系列基于 Transformer 和注意力机制的改进方法[5-8],通过结合患者临床记录中的时间间隔信息并利用自注意力机制来有效学习不同临床事件对于患者健康状态发展的动态时序依赖,一定程度上提升了这些模型的时序建模能力.但是,由于疾病的发展是一个受到多方面因素共同影响的复杂过程,电子病历中记录的不同类型的临床信息(如,诊断信息、用药信息、治疗信息等)彼此间存在着不同的相关性,且它们对于患者未来疾病发展的影响也不尽相同.这些复杂的关联关系并不能简单地通过现有的时序模型得以完整的表达.

图是一种描述复杂关系的常用数据结构,近年来,许多学者尝试将电子病历数据中不同临床实体间的关系表示为图结构,并利用图神经网络(graph neural network, GNN)来提取和挖掘其中蕴含的语义特征和交互模式[9-12].这些方法通过对图结构信息的学习以及节点的聚合虽然增强了医疗实体间语义关系的表征能力,但是,由于它们所使用的图大多基于医学本体的层次结构或医学文献中的先验知识和规则而构建,只能部分地反映临床记录中相关实体或事件的静态语义关联,而无法准确地体现其中隐含的动态时序关系.因此,大大限制了这一类方法对患者健康状态变化模式的捕获能力.

为了解决上述问题,本文提出一种基于多域图神经网络(multi-domain graph neural network, MD-GNN)的疾病预测模型,在得到患者每次就诊的初始特征表示的基础上,根据就诊序列中不同就诊间的相关性和时间间隔信息分别构建了一个就诊亲和图和一个就诊时序图.然后,采用图卷积神经网络来学习患者就诊时间序列中各次就诊记录间的静态语义关联以及动态时序依赖.最后,通过一个多域特征融合模块将时序特征和两种语义关联特征表示结合起来得到最终的患者融合特征表示,用于患者未来的疾病预测.

1　相关工作

1.1　基于 EMR 的疾病预测研究

电子病历是医疗机构对患者临床诊疗和指导干预的、数字化的医疗服务工作记录,详细全面地反映了患者的病程发展以及相关诊疗过程,成了医疗大数据分析应用的关注热点.大量研究人员利用深度学习等技术挖掘 EMR 中的重要临床信息来实现患者未来疾病发展趋势的预测[13-15].Nguyen 等人[16]提出一种基于 CNN 的端到端学习系统,通过自动捕获临床记录历史序列中的相关特征来预测患者的患病风险.Lee 等人[17]提出了一种基于医疗上下文注意力的 RNN 模型,利用一个条件变分自编码器(CVAEs)学习患者的临床表型差异,从而实现有效的诊断预测.Ye 等人[18]提出了 LSAN 模型,基于患者就诊记录中的诊断编码,利用卷积神经网络和 Transformer 两种不同的神经网络分别学习样本就诊记录的短期和长期依赖关系来实现疾病风险预测.为了更好地捕获 EMR 数据中隐含的患者健康状态变化模式,部分

研究考虑了时间序列采样间隔的不规则性来增强模型的时序建模能力.Tan等人[19]提出了一个基于双注意力机制的时间感知GRU模型(DATA-GRU),引入时间衰减函数来学习不规则临床时序数据中的动态特性以提高临床风险预测的性能.An 等人[20]提出了一个时间感知的多类型数据融合表示学习框架,利用一个 BiLSTM 与 CNN 相结合的多分支并行网络,综合诊断、药物、检验和检查等多方面的临床特征实现了患者健康状态变化模式的表示与评估.文献[21]提出了一种基于异常偏移量分析的个性化临床时间序列表示学习模型,利用时序监测数据中各生理指标观测值与正常参考值间的绝对偏移量以及相邻时间步上观测值间的相对偏移量来更准确地生成患者的临床特征表示.由于EMR中包含的异类异构临床数据之间存在着密切的关联,其对患者病程发展的作用也各不相同.上述方法虽然在捕获临床事件随时间变化的动态性方面有了明显的提升,但是对于不同类型数据间相互关系的学习方面却存在着一定的不足.尽管不少学者设计了一系列基于注意力机制的特征融合方法[22-24],但是依然无法完整地挖掘和表达这些复杂的语义关联和交互作用关系.

1.2 基于GNN的疾病预测研究

图是描述现实世界中对象之间复杂关系的常用结构.为了有效地对图结构数据中的复杂模式进行建模,Scarselli 等人[25]提出了图神经网络,通过图节点间的消息传递来捕获图结构数据中的依赖关系.近年来,大量的研究试图将生物医学领域中相关实体间的复杂关系表示为图,并利用图神经网络来实现实体及其关联关系的特征学习[26-28].例如,Ye 等人[29]提出了 MedPath 模型,利用从大量在线医学知识中提取的疾病-疾病以及疾病-症状的相互关系来构建患者个性化知识图谱,再通过图神经网络编码器从知识图谱中学习得到患者病程发展相关的特征表示.文献[30]将患者的就诊信息表示为一张图,然后提出了一种图卷积 Transformer 网络模型(GCT)来学习电子病历数据中隐藏的结构信息,从而得到语义更丰富的患者就诊表示.文献[31]提出了一种知识增强的电子病历表示学习方法,利用分层的图卷积神经网络结合对比学习从医学本体图和本体共现关系图中学习医学编码的特征表示并用于患者的诊断预测.Wu 等人[32]提出了一个基于图的分层医疗实体表示学习框架 ME2Vec,该方法首先根据患者临床记录中诊断、用药、治疗等医疗服务的共现关系构建医疗实体关系图,然后分别使用图注意力网络和一个改进的图节点嵌入算法来学习不同医疗实体的关联以得到医生和患者的特征表示.An 等人[33]提出了一种多图注意力表示学习框架,利用患者电子病历中的结构化编码、实验室检验和生命体征等临床监测数据以及人口学信息,在捕获患者个体临床特征的同时,通过多个并行的图注意力网络从多个方面提取相似患者的群体特征作为补充,以提升患者临床特征表示的完整性和有效性.然而,这些方法大多仅考虑了来自医学先验知识或临床记录中相关医学概念或临床事件间的静态语义关联,而没有充分挖掘隐藏在就诊时间序列数据中的患者健康状态动态变化模式.

2 基于多域图神经网络的疾病预测

2.1 问题描述

仅通过样本筛选的数据无法直接用于训练疾病预测模型,还需要将诊断编码序列转换成向量表示.为了便于对后续基于多域图神经网络的疾病预测模型的描述,将实验数据集中诊断编码的集合表示成 $C=\{c_1,c_2,\cdots,c_{|C|}\}$,$|C|$ 表示诊断编码的数量,任意元素 c_i 代表一个诊断编码.数据集中的样本集合被表示成 $S=\{s_1,s_2,\cdots,s_{|S|}\}$,$|S|$ 表示样本总数.对于任意样本 s_i 的电子病历数据可以由一个医疗就诊序列 $\left\langle V_1^{s_i},V_2^{s_i},\cdots,V_{T(i)}^{s_i}\right\rangle$ 表示,其中 $T(i)$ 是该样本的住院次数,$V_i^{s_i}$ 则表示样本 s_i 的第 i 个就诊记录,$V_j^{s_i}$ 是由一个或多个诊断编码组成的无序集合.此外,任意样本 s_i 每次就诊的时间被记录为 $\left\langle t_1^{s_i},t_2^{s_i},\cdots,t_{T(i)}^{s_i}\right\rangle$,相邻就诊之间的时间间隔序列是 $\delta^{s_i}=\left\langle d_1^{s_i},d_2^{s_i},\cdots,d_{T(i)}^{s_i}\right\rangle$,其中 $d_1^{s_i}=0$,$d_j^{s_i}=t_j^{s_i}-t_{j-1}^{s_i}$.因此,本文涉及的疾病预测问题可以定义为通过样本历史诊断信息 $\left\langle V_1^{s_i},V_2^{s_i},\cdots,V_{T(i)-1}^{s_i}\right\rangle$ 与时间间隔信息 δ^{s_i} 预测未来的诊断结果 $V_{T(i)-1}^{s_i}$.

2.2 模型架构

基于多域图神经网络的疾病预测模型 MD-GNN 的总体架构如图1所示.模型主要可以分为时序特征学习模块、语义特征学习模块和自适应的多域特征融合与预测模块.其中,时序特征学习模块通过编

码级的注意力机制获得每次就诊记录的嵌入向量用来提取患者就诊序列的初始时序特征.语义特征挖掘与表示模块根据就诊序列的相关性和时间间隔分别构建了就诊亲和图和就诊时序图以表示就诊序列之间的语义关系,然后利用加权的图神经网络分别从时序的角度和相关性的角度学习多重语义特征的表示.最后,自适应的多域特征融合与预测模块采用了注意力机制融合时序特征和两种语义关联特征,使模型自主选择地关注更重要的部分,从而得到全面且准确的患者表示.

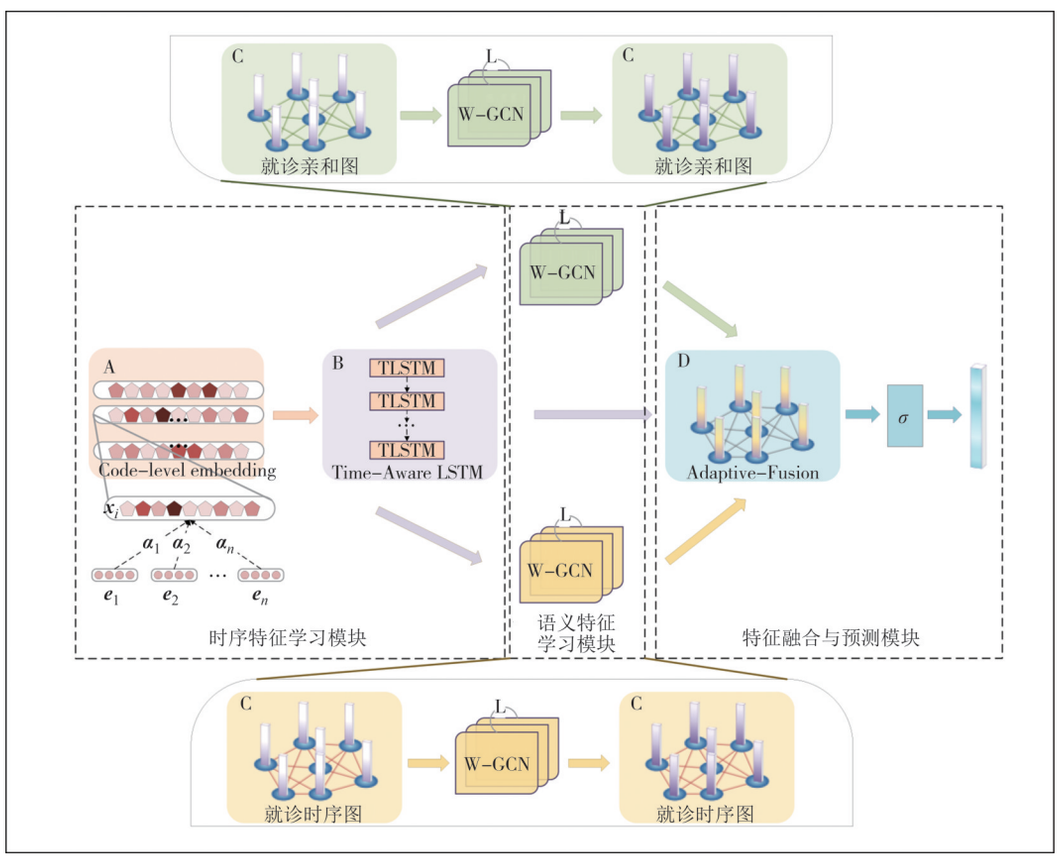


图 1　MD-GNN 模型架构图

Fig.1　Architecture of MD-GNN

2.2.1　基于时序关系的特征学习

首先将诊断编码集合 C 中的任意一个编码 c_i 表示成一个维度为 m 的向量 $\boldsymbol{e}_{c_i} \in \mathbb{R}^m$.那么,对于任意就诊记录 V_t,可以聚合其中的诊断编码信息来获得该次就诊的特征表示.考虑到不同诊断之间存在相关性,同时各诊断对于患者健康状态具有不同的影响,采用了自注意力机制,将任意编码的向量表示 $\boldsymbol{e}_{c_i}$ 通过权重矩阵分别映射为查询向量 $\boldsymbol{q}_{c_i}$、键向量 $\boldsymbol{k}_{c_i}$ 和值向量 $\boldsymbol{v}_{c_i}$.然后,通过注意力函数计算得到对应的注意力分数 α_{c_i},以此来表示诊断序列中各诊断编码对当前就诊的影响程度.具体计算过程如式(1)~式(4)所示.

$$\boldsymbol{q}_{c_i} = \boldsymbol{W}_q \boldsymbol{e}_{c_i} \tag{1}$$

$$\boldsymbol{k}_{c_i} = \boldsymbol{W}_k \boldsymbol{e}_{c_i} \tag{2}$$

$$\boldsymbol{v}_{c_i} = \boldsymbol{W}_v \boldsymbol{e}_{c_i} \tag{3}$$

$$\alpha_{c_i} = \mathrm{softmax}\left(\left[\frac{\boldsymbol{q}_{c_i}\boldsymbol{k}_1}{\sqrt{a}}, \frac{\boldsymbol{q}_{c_i}\boldsymbol{k}_2}{\sqrt{a}}, \cdots, \frac{\boldsymbol{q}_{c_i}\boldsymbol{k}_N}{\sqrt{a}}\right]\right) \tag{4}$$

式中:$\boldsymbol{W}_q \in \mathbb{R}^{a\times m}$,$\boldsymbol{W}_k \in \mathbb{R}^{a\times m}$,$\boldsymbol{W}_v \in \mathbb{R}^{a\times m}$,$a$ 是缩放因子.

于是,第 t 次就诊记录 V_t 的特征向量 $\boldsymbol{x}_t$ 可表示为该次就诊中所含诊断编码向量的加权和.

$$\boldsymbol{x}_t = \sum_{i=1}^{n} \alpha_{c_i} \boldsymbol{v}_{c_i} \tag{5}$$

考虑到大部分 ICU 患者的就诊序列较短,且在我们的模型中仅使用了诊断编码一种类型的数据,而未使用受时间影响较大的实验室检验等数值型数据.因此,我们采用了一个简单的时间感知长短期记忆网络(Time-aware LSTM),在 LSTM 网络的基础上引入了一个信息衰减函数 $\mathrm{decay}(\delta_t) = 1 - \dfrac{1}{1+\mathrm{e}^{-\delta_t}}$ 来

模拟临床事件对患者健康状态的影响随时间衰减的过程.具体来说,在第t个时间步,LSTM以隐藏状态$\boldsymbol{h}_{t-1}$、记忆单元状态$\boldsymbol{c}_{t-1}$和当前的输入$\boldsymbol{x}_t$为输入,输出当前的隐藏状态$\boldsymbol{h}_t$和记忆单元状态$\boldsymbol{c}_t$.Time-aware LSTM通过信息衰减函数decay(δ)将时间间隔转换为权重,用以控制历史记忆单元状态$\boldsymbol{c}_{t-1}$中的短期记忆$\boldsymbol{c}^s_{t-1}$的比重,加权后的短期记忆$\hat{\boldsymbol{c}}^s_{t-1}$再与长期记忆$\boldsymbol{c}^l_{t-1}$结合组成新的历史记忆单元状态$\hat{\boldsymbol{c}}_{t-1}$.后续的隐藏状态的更新与基础LSTM的步骤相同.主要计算过程如式(6)~式(11)所示.

$$\hat{\boldsymbol{c}}^s_{t-1}=\boldsymbol{c}^s_{t-1}\times\mathrm{decay}(\delta_t) \tag{6}$$

$$\boldsymbol{c}^l_{t-1}=\boldsymbol{c}_{t-1}-\boldsymbol{c}^s_{t-1} \tag{7}$$

$$\hat{\boldsymbol{c}}_{t-1}=\hat{\boldsymbol{c}}^s_{t-1}+\boldsymbol{c}^l_{t-1} \tag{8}$$

$$\tilde{\boldsymbol{c}}_t=\tanh(\boldsymbol{W}_c\boldsymbol{x}_t+\boldsymbol{U}_c\boldsymbol{h}_{t-1}+\boldsymbol{b}_c) \tag{9}$$

$$\boldsymbol{c}_t=\boldsymbol{f}_t\otimes\hat{\boldsymbol{c}}_{t-1}+\boldsymbol{i}_t\otimes\tilde{\boldsymbol{c}}_t \tag{10}$$

$$\boldsymbol{h}_t=\boldsymbol{o}_t\odot\tanh(\boldsymbol{c}_t) \tag{11}$$

这里,$\boldsymbol{o}_t$表示第t个时间步的输出门状态.这样,我们可以得到患者历史就诊记录的时序特征表示$\boldsymbol{H}=[\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_T]$.

2.2.2 基于语义关系的特征学习

为充分挖掘就诊记录之间的语义信息,我们基于患者的临床时序数据分别构建了一个就诊亲和图和一个就诊时序图,然后设计了一个基于图神经网络的语义关系特征提取模块.

1) 关系图的构建

我们根据就诊序列间的相关性和时间间隔构建就诊亲和图和就诊时序图,构图规则如下所示:

i)两个关系图中的每个节点都表示一次就诊,图节点的初始表示是Time-aware LSTM模块对应时间步的输出$\boldsymbol{H}=[\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_T]$.

ii)两个关系图均为全连接且带有边权重的无向图.

iii)就诊亲和图中边的权重的定义如下:采用Pearson相关系数p衡量两个就诊序列的相关性,计算过程如式(12)所示.

$$p(\boldsymbol{h}_{ri},\boldsymbol{h}_{rj})=\frac{\boldsymbol{E}\left[(\boldsymbol{h}_{ri}-\boldsymbol{\mu}_{h_{ri}})(\boldsymbol{h}_{rj}-\boldsymbol{\mu}_{h_{rj}})\right]}{\sigma_{h_{ri}}\sigma_{h_{rj}}} \tag{12}$$

式中:$\boldsymbol{E}\left[\left(\boldsymbol{h}_{ri}-\boldsymbol{\mu}_{h_{ri}}\right)\left(\boldsymbol{h}_{rj}-\boldsymbol{\mu}_{h_{rj}}\right)\right]$表示向量$\boldsymbol{h}_{ri}$和向量$\boldsymbol{h}_{rj}$之间的协方差;$\sigma_{h_{ri}}\sigma_{h_{rj}}$表示二者的标准差的乘积.Pearson相关系数$p$介于-1和1之间,$p$值越大表示相关性越高,$p<0$表示负相关,$p=0$表示不相关,$p>0$表示正相关.因此,当$p>0$时,两个就诊序列呈现正相关的关系,且$p$值越大这两个就诊序列之间的相互作用关系越强、影响越大.

iv)就诊时序图中边的权重定义如下:假设两次就诊之间相隔的时间越久,那么这两次就诊信息的相互影响就越小.因此,这里采用与2.2.1中相同的衰减函数计算连接两个就诊节点的边权重.

根据上述规则可以构建得到就诊亲和图$G_R=(\boldsymbol{H}_R,\boldsymbol{E},\Lambda_R)$和就诊时序图$G_T=(\boldsymbol{H}_\Delta,\boldsymbol{E},\Lambda_\Delta)$,其中$\boldsymbol{H}_R=\boldsymbol{H}_\Delta=\boldsymbol{H}$是关系图节点的表示,$\boldsymbol{E}$为元素全为1的矩阵,表示图节点的连接关系,$\Lambda_R$和$\Lambda_\Delta$则分别对应就诊亲和图和就诊时序图中的边权重.

2) 关系特征提取

采用图卷积神经网络(GCN)来更新每个节点的表示,它能够有效地聚合图节点及其邻域的局部信息.第l层的图卷积以前一层的图节点表示作为输入,然后通过卷积操作为每个节点生成一个新的表示.在模型MD-GNN中我们对传统的图卷积神经网络进行了改进.传统的图卷积神经网络应用于无向无权图,在进行卷积操作时对所有的邻接节点一视同仁,而不考虑各个邻接节点对当前节点的重要程度的不同.然而,本文中设计的两个语义关系图都是无向带权图,因此在进行卷积时需要考虑边的权重,即邻接节点在重要程度上的差异,从而有所侧重地聚合邻域信息.此外,为聚合到非直接邻接节点的信息,我们采用了一个两层的GCN聚合更新关系图中各节点的信息,基于就诊亲和图G_R和就诊时序图G_T的图卷积过程分别如式(13)和式(14)所示.

$$\boldsymbol{h}^{(l)}_{ri}=\sigma\left(\boldsymbol{W}^{(l)}\sum\nolimits_{j\in\aleph(i)\cup\{i\}}\frac{p_{j,i}}{\sqrt{\deg(i)}\sqrt{\deg(j)}}\boldsymbol{h}^{(l-1)}_{ri}\right) \tag{13}$$

$$\boldsymbol{h}^{(l)}_{\delta i}=\sigma\left(\boldsymbol{W}^{(l)}\sum\nolimits_{j\in\aleph(i)\cup\{i\}}\frac{\beta_{j,i}}{\sqrt{\deg(i)}\sqrt{\deg(j)}}\boldsymbol{h}^{(l-1)}_{\delta i}\right) \tag{14}$$

式中:$\sigma()$表示激活函数ReLU();$\boldsymbol{W}^{(l)}$是权重矩阵;$\aleph(i)$表示图节点i的邻接节点的集合;$\deg(i)$表示图节点i的度;$\boldsymbol{h}^{(0)}_{ri}=\boldsymbol{h}^{(0)}_{\delta i}=\boldsymbol{h}_i$.

2.2.3 多域特征融合与预测模块

通过基于时序关系的特征学习和基于语义关系的特征学习,可以得到时序特征表示$\boldsymbol{h}$以及两种类型的语义关系表示$\boldsymbol{h}_r$和$\boldsymbol{h}_\delta$.对于每一位患者,来自不

同邻域的特征对未来的疾病预测的影响程度不同.在此背景下,采用self-attention的融合机制使得模型可以自适应地为各部分特征分配权重,并通过加权求和进行特征融合.通过上述过程学习得到最终的患者表示$\boldsymbol{h}'$.

$$\boldsymbol{h}' = \text{Attention}\left(\left[\boldsymbol{h} \middle|\middle| \boldsymbol{h}_r \middle|\middle| \boldsymbol{h}_\delta\right]\right) \tag{15}$$

最后,采用一个带有激活函数Sigmoid()的线性层计算得到每种疾病的概率,计算如式(16)所示.

$$y' = \text{Sigmoid}(\boldsymbol{W}_u \boldsymbol{h}' + \boldsymbol{b}_u) \tag{16}$$

由于是多标签分类任务,我们采用二进制交叉熵损失函数BCELoss()计算真实标签y和预测概率y'之间的损失.假设Θ表示所有的参数,那么损失计算公式如(17)所示:

$$L(\Theta) = -\frac{1}{N}\sum_{n=1}^{N}\left(y_n \log\left(y_n'\right) + \left(1 - y_n\right)\log\left(1 - y_n'\right)\right) \tag{17}$$

3　实验结果与分析

3.1　数据描述与预处理

本文研究中使用了MIMIC-Ⅲ[34]和MIMIC-Ⅳ[35]两个数据集来进行模型的性能评估.其中,MIMIC-Ⅲ是一个大型的免费医疗数据集,该数据集收集了2001年至2012年之间波士顿贝斯以色列医院(Beth Israel Deaconess Medical Center)重症监护病房收治的4万多患者的诊疗记录.MIMIC-Ⅳ是对MIMIC-Ⅲ数据集的改进和扩展,记录了共256 878名患者的就诊信息,其中包括超过50 000名有重症监护单位经历的患者.

根据对两个数据集中患者的就诊次数的统计,我们将本文的研究目标设定为利用患者前3次历史就诊记录中的诊断编码来预测第4次就诊的患病风险.因此,对于MIMIC-Ⅳ数据集,我们滤除了就诊次数少于4次的患者.对于就诊次数大于4次的患者我们使用一个长度为4的滑动窗口以4为步长进行采样,以此来扩充样本集.而对于MIMIC-Ⅲ数据集,由于其中大多数患者的就诊次数均不足3次,所以,我们选择保留了就诊次数不小于2次的患者,并对就诊序列长度不足4次的采用零向量填充的方式予以补齐.最终数据集的统计信息如表1所示.

3.2　实验设置和评估指标

我们将经过筛选的数据按照8∶1∶1的比例划分训练集、验证集和测试集.为了保证比较的公平性,所有的实验均采用5折交叉验证来评估模型的性能,并设置了95%置信区间.为了评价模型的性能,我们根据患者每次就诊平均的诊断数量选择了Recall@10,Recall@20和AUC作为评价指标.为优化模型参数,采用BECLoss()作为损失函数,优化器为AdamW,以上设置也适用于所有的基准模型和消融模型.本文中涉及的所有实验均基于PyTorch 1.10.2深度学习框架实现并采用批量训练,batch size设置为128, epochs设置为150.我们提出的模型MD-GNN的基本参数设置如下,诊断编码嵌入维度为64,Time-aware LSTM的隐藏向量维度为128,两个图卷积维度均为128.为防止过拟合,我们采用了dropout机制,在MIMIC-Ⅲ和MIMIC-Ⅳ数据集上的实验中,dropout_rate分别设为0.2和0.3.

表1　最终数据集统计信息

Tab.1　The final dataset statistics

参数	MIMIC-Ⅲ	MIMIC-Ⅳ
患者数	6 609	11 733
样本数	6 797	17 124
ICD编码总数	931	1 036
平均就诊次数	2.74	6.82
单次就诊平均ICD编码数	12.76	10.58

3.3　Baseline方法

为验证我们提出的基于多域图神经网络的疾病预测模型MD-GNN的预测效果,我们选择了六种相关方法进行了性能对比,具体描述如下.

Dipole[1]:一个端到端的预测模型,它采样基于注意力机制的双向循环神经网络来学习患者每次就诊的表示,并将学习到的表示用于未来的诊断预测.同时,注意力机制也被用来学习每次就诊对未来诊断预测的重要性以此解释预测结果.

TLSTM[4]:一种改进的基于LSTM的方法,通过引入时间衰减函数来处理不规则时间间隔,再由时间衰减函数得到的权重来更改LSTM中之前的信息.

LSAN[18]:由层次注意力模块和时间聚合模块组成.首先通过层次注意力模块学习每次就诊的嵌入表示作为时间聚合模块的输入,使模型关注重要的诊断编码.时间聚合模块有两个通道,通道一由卷积神经网络构成用于提取短期时间依赖关系,通道二由Transformer组成用于学习长期依赖关系,然后拼接两个通道的输出.最后再次使用层次注意力模块为每个就诊赋予权重,使模型关注更重要的就诊

数据.

GCN-TLSTM:一个基于图卷积神经网络和TLSTM的端到端的预测模型.首先利用GCN学习诊断编码之间的依赖关系得到每次就诊的特征表示,然后再通过TLSTM学习电子病历数据的时序特征.

IICL[26]:一种基于图对比学习的疾病预测方法.该方法分别构建了一个全局静态疾病关系图和一个局部动态疾病关系图,并通过GNN来学习疾病间的相互关系,然后再利用多个对比学习任务来进一步提升疾病特征表示的准确性.

MulT-EHR[28]:一种基于异构图的EHR表示学习框架,其首先利用基于图对比学习的预训练模块增强图中节点的关系特征,并采用一个基于Transformer的图神经网络来学习节点级的特征表示,最后通过一个多任务学习模块,充分利用任务间的相关知识来优化训练过程.在本文实验中,MulT-EHR构建的异构图中我们只使用了诊断信息而未使用处方、实验室检验以及手术三类信息,以保持所有方法输入数据的一致性.同时,为了重点比较图表示学习方法的作用,我们删除了MulT-EHR中的多任务学习模块.

3.4 结果分析

3.4.1 整体性能分析

如表2所示,在疾病预测任务上我们提出的模型MD-GNN在两个公共数据集上均表现出了最佳性能,在MIMIC-Ⅲ数据集上Recall@10为0.347 0,在MIMIC-Ⅳ数据集上Recall@10达到了0.423 2.在EMR时序数据建模方面,与直接使用简单的序列模型相比,我们提出的模型对时间信息更加精细化的处理可以更准确地捕获患者健康状态随时间动态变化的特征,因此引入了时间衰减函数的序列模型TLSTM比直接使用双向循环神经网络的Dipole表现出了更高的Recall和AUC.LSAN不仅考虑到了每次就诊中各个诊断编码的重要程度的不同,还考虑了每次的就诊记录对于最终的疾病预测的贡献的差异,减少了干扰信息对模型预测的影响.同时,它使用了卷积神经网络和Transformer两种不同的神经网络分别学习患者EMR数据中的短期和长期依赖关系.然而,在本文实验中使用的就诊序列大部分都较短,一定程度上可能限制了Transformer优势的发挥.因此,LSAN在三种非图神经网络的方法中表现最佳.

表2 MD-GNN与其他baseline模型在两个数据集上的对比结果

Tab.2 Comparative results between baseline models and MD-GNN on two datasets

模型	MIMIC-Ⅲ			MIMIC-Ⅳ		
	Recall@10	Recall@20	AUC	Recall@10	Recall@20	AUC
Dipole	0.296 3 (±0.003 5)	0.422 2 (±0.002 3)	0.614 4 (±0.008 6)	0.3460 (±0.0067)	0.4833 (±0.0048)	0.676 2 (±0.006 7)
TLSTM	0.301 6 (±0.006 1)	0.434 7 (±0.007 4)	0.616 1 (±0.013 9)	0.364 9 (±0.002 4)	0.511 4 (±0.004 5)	0.690 9 (±0.002 1)
LSAN	0.337 6 (±0.005 0)	0.470 2 (±0.003 7)	0.637 8 (±0.008 9)	0.404 9 (±0.009 8)	0.535 6 (±0.011 4)	0.706 7 (±0.009 5)
GCN-TLSTM	0.308 6 (±0.007 8)	0.441 8 (±0.005 1)	0.617 5 (±0.009 3)	0.391 4 (±0.003 7)	0.534 8 (±0.002 4)	0.713 5 (±0.006 8)
IICL	0.338 7 (±0.006 7)	0.473 9 (±0.008 0)	0.641 6 (±0.010 3)	0.410 2 (±0.005 9)	0.552 0 (±0.007 3)	0.718 6 (±0.007 7)
MulT-EHR	0.314 4 (±0.005 7)	0.454 7 (±0.008 6)	0.628 1 (±0.007 9)	0.398 9 (±0.004 5)	0.539 1 (±0.006 5)	0.720 0 (±0.008 4)
MD-GNN	0.347 0 (±0.010 5)	0.482 1 (±0.011 5)	0.657 3 (±0.014 9)	0.423 2 (±0.006 4)	0.567 1 (±0.007 5)	0.734 4 (±0.006 4)

从表中结果还可以发现,基于图神经网络的几种方法的性能整体要优于其他对比方法.其中,GCN-TLSTM通过在TLSTM的基础上增加一个图卷积神经网络来增强就诊级的特征表示能力,其性能相比TLSTM得到了一定的提升.IICL利用图对比学习有效地捕获了疾病间的静态和动态关系,提高了特征表示的准确性,因此,其获得了仅次于MD-GNN的性能表现.MD-GNN之所以能取得较好的预测性能,我们认为主要得益于以下几方面:首先采用编码级的注意力机制学习每次就诊的嵌入表示,可以使模型更加关注与预测相关的诊断编码而减少一些无关记录对预测的影响.其次,时间感知的LSTM可以

根据时间间隔决定保留历史信息的数量,从而学习到具有不规则时序性的隐藏特征表示.更重要的是,我们引入了多域图神经网络从就诊序列间的相关性和就诊序列间的时间间隔两个角度构建就诊亲和图和就诊时序图,并利用图卷积神经网络学习就诊序列间的语义关系.就诊亲和图的语义特征学习根据就诊序列之间的相关性衡量就诊之间的影响,就诊时序图的语义特征学习不用考虑顺序而直接根据时间间隔计算就诊之间的影响程度,这使得信息的传递更加简单直接,一定程度上也减少了信息损失和噪声,因此,显著地提升了最终患者表示的有效性.

3.4.2　消融实验

为了检验MD-GNN不同组成部分的有效性,我们对MD-GNN在两个数据集上进行了消融实验.我们逐一删除MD-GNN的某一部分,同样通过Recall@10、Recall@20和AUC来验证它们的性能.在我们的消融实验中使用的模型如下:

MD-GNN(A):删除MD-GNN中的编码级注意力机制(code-level embedding),直接使用multi-hot向量作为Time-aware LSTM模块的输入.

MD-GNN(B):删除MD-GNN中的就诊亲和图特征提取部分.

MD-GNN(C):删除基于注意力机制的多域特征融合模块(adaptive-fusion),将融合方式改为向量的拼接.

MD-GNN(D):删除MD-GNN中的就诊时序图特征提取部分.

MD-GNN(E):完全移除语义特征学习模块,仅保留时序特征学习模块.

如表3所示,删除MD-GNN任意一部分都对预测性能有所影响.删除就诊亲和图的特征提取部分和就诊时序图的特征提取部分对模型的预测性能的影响比较显著.当我们移除了就诊亲和图的特征提取模块后,MD-GNN(B)与MD-GNN相比,它的Recall@10和AUC在数据集MIMIC-Ⅳ分别下降到0.390 7和0.718 6.在移除就诊时序图的特征提取模块后,MD-GNN(D)与MD-GNN相比,它的Recall@10和AUC在数据集MIMIC-Ⅲ分别下降了6.75个百分点和10.07个百分点.这说明通过图神经网络学习就诊序列之间的语义关系可以挖掘到患者EMR数据中更多的潜在信息,从而使最终的患者表示更全面和准确.与此同时,通过比较消融模型MD-GNN(A)和模型MD-GNN的预测性能,我们也可以发现,为每次就诊中的诊断编码赋予不同的权重使模型关注更重要的编码,有助于学习更加精确的就诊记录表示从而提高疾病预测的综合性能.表3中的比较结果也验证了基于注意力机制的多域特征融合方式在提高预测性能方面的有效性.此外,当完全移除语义特征学习模块后,模型MD-GNN(E)与完整模型MD-GNN相比在两个数据集上的性能均有所下降,这证明无论是从就诊亲和图还是就诊时序图中提取的语义特征都对患者未来的患病风险预测有积极作用.从表3可以发现,完全移除语义特征学习模块的模型MD-GNN(E),与仅移除部分语义信息的模型MD-GNN(B)和MD-GNN(D)相比,性能反而有所提升,这可能是因为筛选来自MIMIC-Ⅲ的数据大部分都没有完整的就诊序列,而使用了零向量进行填充.经过填充后的数据无法准确地反映出真实的就诊序

表3　MD-GNN与其变种模型的性能比较

Tab.3　Performance comparison for MD-GNN's variants

模型	MIMIC-Ⅲ			MIMIC-Ⅳ		
	Recall@10	Recall@20	AUC	Recall@10	Recall@20	AUC
MD-GNN(A)	0.315 9 (±0.005 3)	0.450 7 (±0.003 4)	0.641 6 (±0.001 8)	0.404 8 (±0.003 5)	0.553 2 (±0.002 9)	0.737 1 (±0.001 4)
MD-GNN(B)	0.301 1 (±0.020 4)	0.431 3 (±0.020)	0.606 9 (±0.041 6)	0.390 7 (±0.005 3)	0.535 6 (±0.006 6)	0.718 6 (±0.004 1)
MD-GNN(C)	0.280 4 (±0.018 0)	0.415 5 (±0.018 0)	0.562 1 (±0.044 5)	0.417 0 (±0.006 0)	0.561 6 (±0.006 0)	0.745 5 (±0.005 8)
MD-GNN(D)	0.279 5 (±0.015 1)	0.415 2 (±0.014 4)	0.556 6 (±0.038 0)	0.400 0 (±0.003 8)	0.544 1 (±0.001 6)	0.721 6 (±0.006 5)
MD-GNN(E)	0.301 8 (±0.022 7)	0.432 9 (±0.020 3)	0.601 9 (±0.036 5)	0.384 9 (±0.004 9)	0.532 6 (±0.004 7)	0.723 2 (±0.003 4)
MD-GNN	0.347 0 (±0.010 5)	0.482 1 (±0.011 5)	0.657 3 (±0.014 9)	0.423 2 (±0.006 4)	0.567 1 (±0.007 5)	0.734 4 (±0.006 4)

列的关系,甚至可能会产生部分噪声,从而导致上述情况的发生.除了各个消融模型的性能具有一定的差异外,我们可以发现所有的模型,无论是预测性能还是稳定性,在数据集MIMIC-Ⅳ上的表现均优于在MIMIC-Ⅲ上的表现,导致此现象的原因可能是MIMIC-Ⅲ的样本量远小于MIMIC-Ⅳ的样本量.而且最终的MIMIC-Ⅲ数据集中患者平均就诊次数仅为2.74,这意味着大量的样本均被使用了零向量进行填充而无法十分准确地表达患者的疾病进展情况.

3.4.3 案例分析

为进一步分析时序特征与语义特征对模型MD-GNN预测的影响,我们以患者B和患者C的就诊记录为例,通过自适应特征融合部分的权重观察不同特征对疾病风险预测的贡献.如图2所示,MD-GNN更加关注患者B就诊记录之间的相关性,通过W-GCN在就诊亲和图中学习得到的语义特征被赋予了0.82的权重,远高于时序特征和基于就诊时序图学习到的语义特征的权重.根据该患者的多次诊断记录,可以发现该患者第一次就诊中就包含药物滥用(酒精、致幻剂等)、肠胃炎和阿尔兹海默症等,其中药物滥用并不会在短期内消失甚至极可能反复与加重.此外,阿尔兹海默症此类神经退行性疾病对未来患病的影响也不会随时间的推移而减弱,因此相比于就诊时间对诊断的影响,模型更加关注每次诊断之间的亲和性.如图中患者C的诊断记录所示,与着重关注患者B的就诊间的相关性不同,MD-GNN对从患者C的电子病历中提取的三种类型的特征的关注度差异较小.一方面,该患者有擦伤、酒精中毒等,这类疾病通常是短暂出现且不会对未来患病产生较大的影响,可能会随时间推移而逐渐消失,因此模型会比较关注时序特征和时间间隔的影响;另一方面,该患者还患有风湿这类慢性疾病,且存在酒精依赖的情况,此类疾病一般会长期存在,因此模型需要关注就诊序列间的相关性,即从就诊亲和图中提取的语义特征.

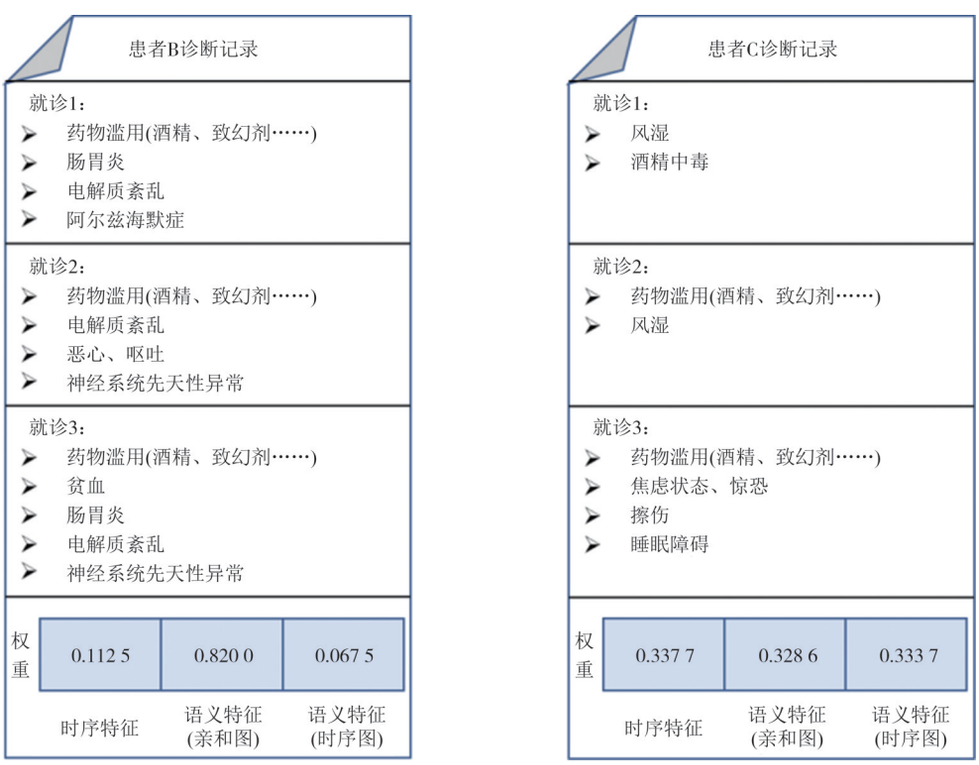


图2 患者就诊记录及MD-GNN捕获的特征权重示例

Fig.2 Example of patient visit records and feature weights captured by MD-GNN

4 结 论

本文提出了一个基于多域图神经网络的疾病预测模型MD-GNN,在学习就诊序列的时序特征的同时,提取就诊序列间的多重语义特征.在时序关系的特征学习模块中,采用了编码级注意力机制学习就诊嵌入并通过时间感知LSTM捕获不规则时序特征.在语义特征学习模块中,基于序列间的相关性和时间间隔分别构建两种语义关系图(就诊亲和图和就

诊时序图),加权的图卷积神经网络被用于语义特征的提取.最后引入自注意力机制使模型自适应地融合各类特征表示.该模型的疾病预测性能在数据集 MIMIC-Ⅲ 和 MIMIC-Ⅳ 上得到了验证,并开展了消融研究,证明了 MD-GNN 各核心模块的有效性.

参考文献

[1] MA F L, CHITTA R, ZHOU J, et al. Dipole: diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks[C]//Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Halifax NS, Canada. ACM, 2017: 13-17.

[2] KIM Y J, LEE Y G, KIM J W, et al. Highrisk prediction from electronic medical records via deep attention networks [C]// Proceedings of the 31st Conference on Neural Information Processing Systems. Long Beach, CA, USA. ACM, 2017: 1-10.

[3] AN Y, HUANG N J, CHEN X L, et al. High-risk prediction of cardiovascular diseases via attention-based deep neural networks[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2021, 18(3): 1093-1105.

[4] BAYTAS I M, XIAO C, ZHANG X, et al. Patient subtyping via time-aware LSTM networks [C]//Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Halifax, NS, Canada. ACM, 2017: 65-74.

[5] LUO J Y, YE M C, XIAO C, et al. HiTANet: hierarchical time-aware attention networks for risk prediction on electronic health records[C]//Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. Virtual Event, CA, USA. ACM, 2020: 647-656.

[6] LI R, GAO J. Multi-modal contrastive learning for healthcare data analytics[C]//2022 IEEE 10th International Conference on Healthcare Informatics (ICHI). Rochester, MN, USA. IEEE, 2022: 120-127.

[7] NIU S, YIN Q, SONG Y Y, et al. Label dependent attention model for disease risk prediction using multimodal electronic health records[C]//2021 IEEE International Conference on Data Mining (ICDM). Auckland, New Zealand. IEEE, 2021: 449-458.

[8] MENG Y W, SPEIER W, ONG M K, et al. Bidirectional representation learning from transformers using multimodal electronic health record data to predict depression [J]. IEEE Journal of Biomedical and Health Informatics, 2021, 25(8): 3121-3129.

[9] CHOI E, BAHADORI M T, SONG L, et al. GRAM [C]// Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Halifax, NS, Canada. ACM, 2017: 787-795.

[10] MA F L, YOU Q Z, XIAO H P, et al. KAME[C]//Proceedings of the 27th ACM International Conference on Information and Knowledge Management. Torino, Italy. ACM, 2018: 743-752.

[11] LI Y, QIAN B Y, ZHANG X L, et al. Knowledge guided diagnosis prediction via graph spatial-temporal network[C]//Proceedings of the 2020 SIAM International Conference on Data Mining. Philadelphia, PA: Society for Industrial and Applied Mathematics, 2020: 19-27.

[12] CHOI E, XIAO C, SUN J, et al. MiME: multilevel medical embedding of electronic health records for predictive healthcare [C] //Proceedings of the 32st Conference on Neural Information Processing Systems. Montreal, QC, Canada. ACM, 2018: 4547-4557.

[13] MENG Q W, QIAN H W, LIU Y, et al. MHCCL: Masked hierarchical cluster-wise contrastive learning for multivariate time series [C]//Proceedings of the AAAI Conference on Artificial Intelligence. Washington, DC, USA. ACM, 2023: 9153-9161.

[14] ANSARI A F, HENG A, LIM A, et al. Neural continuous-discrete state space models for irregularly-sampled time series [C]//Proceedings of the 40th International Conference on Machine Learning. Honolulu, Hawaii, USA. ACM, 2023: 926-951.

[15] RAGHU A, CHANDAK P, ALAM R, et al. Sequential multi-dimensional self-supervised learning for clinical time series[C]// Proceedings of the 40th International Conference on Machine Learning. Honolulu, Hawaii, USA. ACM, 2023: 28531-28548.

[16] NGUYEN P, TRAN T, WICKRAMASINGHE N, et al. Deepr: a convolutional net for medical records [J]. IEEE Journal of Biomedical and Health Informatics, 2017, 21(1): 22-30.

[17] LEE W, PARK S, JOO W, et al. Diagnosis prediction via medical context attention networks using deep generative modeling[C]// 2018 IEEE International Conference on Data Mining (ICDM). Singapore. IEEE, 2018: 1104-1109.

[18] YE M C, LUO J Y, XIAO C, et al. LSAN: modeling long-term dependencies and short-term correlations with hierarchical attention for risk prediction[C]//Proceedings of the 29th ACM International Conference on Information & Knowledge Management. Virtual Event, Ireland. ACM, 2020: 1753-1762.

[19] TAN Q X, YE M, YANG B Y, et al. DATA-GRU: dual-attention time-aware gated recurrent unit for irregular multivariate time series [C]//Proceedings of the AAAI Conference on Artificial Intelligence. New York, USA. ACM, 2020: 930-937.

[20] AN Y, TANG K, WANG J X. Time-aware multi-type data fusion representation learning framework for risk prediction of cardio vascular diseases [J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2022, 19(6): 3725-3734.

[21] AN Y, CAI G L, CHEN X L, et al. PARSE: a personalized clinical time-series representation learning framework via abnormal offsets analysis[J]. Computer Methods and Programs in Biomedicine, 2023, 242: 107838.

[22] AN Y, ZHANG H J, SHENG Y, et al. MAIN: multimodal attention-based fusion networks for diagnosis prediction[C]//2021

IEEE International Conference on Bioinformatics and Biomedicine (BIBM). Houston, TX, USA. IEEE, 2021: 809-816.

[23] ZHANG C H, CHU X, MA L T, et al. M3Care: learning with missing modalities in multimodal healthcare data[C]//Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. Washington DC, USA. ACM, 2022: 2418-2428.

[24] LIU S C, WANG X L, XIANG Y, et al. Multi-channel fusion LSTM for medical event prediction using EHRs[J]. Journal of Biomedical Informatics, 2022, 127: 104011.

[25] SCARSELLI F, GORI M, TSOI A C, et al. The graph neural network model[J]. IEEE Transactions on Neural Networks, 2009, 20(1): 61-80.

[26] KANG Y, ZHENG J Y, YANG M J, et al. Inter-structure and intra-semantics graph contrastive learning for disease prediction[J]. Knowledge-Based Systems, 2024, 300: 112059.

[27] WANG Y C, XU Y C, YANG J F, et al. Fully-connected spatial-temporal graph for multivariate time-series data[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, 38(14): 15715-15724.

[28] CHAN T H, YIN G S, BAE K, et al. Multi-task heterogeneous graph learning on electronic health records[J]. Neural Networks, 2024, 180: 106644.

[29] YE M C, CUI S H, WANG Y Q, et al. MedPath: augmenting health risk prediction via medical knowledge paths [C]// Proceedings of the Web Conference 2021. Ljubljana, Slovenia. ACM, 2021: 1397-1409.

[30] CHOI E, XU Z, LI Y J, et al. Learning the graphical structure of electronic health records with graph convolutional transformer[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(1): 606-613.

[31] AN Y, SHI Y H, GUO L, et al. Knowledge-enhanced difference-aware clinical time series representation learning for diagnosis prediction [C]//2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). Istanbul, Turkiye. IEEE, 2023: 1014-1021.

[32] WU T, WANG Y L, WANG Y, et al. Leveraging graph-based hierarchical medical entity embedding for healthcare applications[J]. Scientific Reports, 2021, 11(1): 5858.

[33] AN Y, LI R Z, CHEN X L. MERGE: a multi-graph attentive representation learning framework integrating group information from similar patients[J]. Computers in Biology and Medicine, 2022, 151: 106245.

[34] JOHNSON A E W, POLLARD T J, SHEN L, et al. MIMIC-Ⅲ, a freely accessible critical care database[J]. Scientific Data, 2016, 3: 160035.

[35] JOHNSON A E W, BULGARELLI L, SHEN L, et al. MIMIC-Ⅳ, a freely accessible electronic health record dataset[J]. Scientific Data, 2023, 10(1): 1.