



# 计算机科学

COMPUTER SCIENCE

## 基于图卷积网络和注意力机制的诊断预测

杨仙明, 詹贤春, 程恒亮, 丁海燕

引用本文

杨仙明, 詹贤春, 程恒亮, 丁海燕. [基于图卷积网络和注意力机制的诊断预测](#)[J]. 计算机科学, 2023, 50(11A): 221100232-6.

YANG Xianming, ZHAN Xianchun, CHEN Hengliang, DING Haiyan. [Diagnosis Prediction Based on Graph Convolutional Network and Attention Mechanism](#) [J]. Computer Science, 2023, 50(11A): 221100232-6.

---

## 相似文章推荐 (请使用火狐或 IE 浏览器查看文章)

Similar articles recommended (Please use Firefox or IE to view the article)

### [一种融合CNN和Swin Transformer的医学显微图像分割模型](#)

Medical Microscopic Image Segmentation Model Based on CNN Structure and Swin Transformer  
计算机科学, 2023, 50(11A): 230200119-8. <https://doi.org/10.11896/jsjcx.230200119>

### [一种面向多模态医疗数据的联邦学习隐私保护方法](#)

Federated Learning Privacy-preserving Approach for Multimodal Medical Data  
计算机科学, 2023, 50(11A): 230800021-8. <https://doi.org/10.11896/jsjcx.230800021>

### [基于注意力机制和ConvLSTM的船舶交通流量预测算法](#)

Ship Traffic Flow Prediction Algorithm Based on Attention Mechanism and ConvLSTM  
计算机科学, 2023, 50(11A): 230800067-7. <https://doi.org/10.11896/jsjcx.230800067>

### [基于知识蒸馏和高效通道注意力的异常检测](#)

Novelty Detection Method Based on Knowledge Distillation and Efficient Channel Attention  
计算机科学, 2023, 50(11A): 220900034-10. <https://doi.org/10.11896/jsjcx.220900034>

### [融合物品关系的图神经网络推荐算法](#)

Graph Neural Network Recommendation Algorithm Based on Item Relations  
计算机科学, 2023, 50(11A): 230100019-9. <https://doi.org/10.11896/jsjcx.230100019>

# 基于图卷积网络和注意力机制的诊断预测

杨仙明 詹贤春 程恒亮 丁海燕

云南大学信息学院 昆明 650504

(yxm3053@163.com)

**摘要** 诊断预测是医疗保健中一项重要的预测任务,其目的是根据患者的历史健康记录预测其未来的诊断。基于注意力机制和循环神经网络的预测模型被广泛应用于解决这一任务,但容易受到数据不足的影响。此外,医学领域知识在改进诊断预测模型性能上已经显示出重要作用,但现有方法还不能充分利用这些领域知识。因此,设计了一种基于图卷积网络和注意力机制的诊断预测模型。具体而言,首先利用医学本体对医学概念之间的相关性进行建模,并将患者就诊信息构建为一个图;其次,通过图卷积模块在图结构上获取患者每次就诊中各医学代码之间的空间特征;最后利用多头注意力机制来对就诊特征和多级医学知识之间的相互关系进行建模,从而对患者的未来健康状况进行预测。在两个公开的医疗数据集上的实验结果表明,该模型的诊断预测性能优于已有诊断预测模型,可以更有效地利用医学知识图中的潜在信息。

**关键词:** 诊断预测;电子健康记录;医学领域知识;图卷积网络;注意力机制

**中图法分类号** TP391

## Diagnosis Prediction Based on Graph Convolutional Network and Attention Mechanism

YANG Xianming, ZHAN Xianchun, CHEN Hengliang and DING Haiyan

School of Information Science and Engineering, Yunnan University, Kunming 650504, China

**Abstract** Diagnosis prediction is an important prediction task in healthcare, which aims to predict the future diagnosis of patients based on their historical health records. Predictive models based on attention mechanisms and recurrent neural network are widely used to solve this task, but they are easy to be affected by insufficient data. In addition, medical domain knowledge plays an important role in improving the performance of diagnosis prediction, but existing methods still cannot make full use of those knowledge. Therefore, a diagnostic prediction model based on graph convolutional network and attention mechanism is designed. Firstly, the medical ontology is used to model the correlation between medical concepts, then the patient visit information is modeled as a graph. Secondly, the graph convolution module is used to obtain the spatial features between the medical codes in each visit of the patient. Finally, a multi-head attention mechanism is used to model the interrelationship between visit features and multi-level medical knowledge to predict the future health status of patients. Experimental results on two publicly available medical datasets show that the diagnosis prediction performance of the model is better than that of the existing diagnostic prediction models, and the potential information in the medical knowledge graph can be used more effectively.

**Keywords** Diagnostic prediction, Electronic health record, Medical domain knowledge, Graph convolutional network, Attention mechanism

## 1 引言

电子健康记录(Electronic Health Record, EHR)记录了与患者健康相关的入院情况、症状、检查、诊断和治疗等临床信息<sup>[1]</sup>,这些信息原是为提高医疗保健效率而存储的,但随着医疗行业的发展和EHR数据的爆炸式增长,越来越多的研究者发现EHR数据具有较高的研究价值并且可以被二次有效利用<sup>[2]</sup>。目前,EHR数据已被成功应用于多项医疗研究任务中,在描述患者病情和预测患者再住院、死亡率、未来发病率和疾病预测等方面起着重要作用,充分利用和挖掘EHR数据中的有效信息对医疗卫生事业的发展起着重要的推动作用<sup>[3-4]</sup>。

诊断预测旨在根据患者的历史就诊信息预测患者未来可能发生的疾病,是临床事件预测任务中最重要的任务之一,已成为个性化医疗保健的基石,大量可用的EHR数据和不断

发展的深度学习技术为诊断预测研究提供了机会。深度学习模型可以有效地从原始数据中学习到相关的特征,在各个任务中实现了比传统机器学习模型更优越的性能。循环神经网络(Recurrent Neural Network, RNN)在处理时间序列的预测任务上已经有了很不错的性能,基于RNN的相关模型在疾病诊断预测中也取得了不错的成果。例如,Doctor AI<sup>[5]</sup>将RNN与注意力机制相结合,根据患者已有的就诊记录预测下次就诊中的患病情况。RETAIN<sup>[6]</sup>受临床实践的启发,认为越接近当前预测的就诊可能贡献越大,利用RNN从逆向时间上对EHR数据进行建模。Dipole<sup>[7]</sup>使用双向RNN和注意力机制来预测患者未来的健康状况。尽管上述方法不断改进诊断预测,但在可解释性上存在不足。

为了解决上述问题,以医学领域知识为导向的诊断预测方法成为一个有效的解决方案,该诊断预测方法主要利用

基金项目:国家自然科学基金(32060151)

This work was supported by the National Natural Science Foundation of China(32060151).

通信作者:丁海燕(teidhy@163.com)

额外的资源来更好地学习医学概念的有效表示。GRAM<sup>[8]</sup>利用医学本体和基于图的注意力机制来学习健壮的医学代码表示,在数据不足的情况下能够获得较好的预测性能。在GRAM基础上,KAME<sup>[9]</sup>进一步探索了医学本体中不同层级的医学概念之间的关系,学习医学本体中的疾病与上层医学概念的嵌入表示,以实现更加准确的预测效果。CAMP<sup>[10]</sup>使用记忆网络增强了门控循环单元,更进一步提高了诊断预测性能。上述方法均证明了附加的医学领域知识可以有效地弥补数据不足,提高模型的可解释性,但它们都只利用了医学领域知识中的部分信息,忽略了可能同等重要的图结构特征。此外,这些研究利用医学本体中蕴含的医学知识作为从EHR数据中分离出来的外部信息,在训练模型时可能会引入额外的噪声。

为解决上述问题,NKAM<sup>[11]</sup>使用Node2vec学习医学概念的低维向量表示,并设计一种知识注意力机制实现了更优越的预测性能。GNDP<sup>[12]</sup>首先提出以时空图的形式重建患者的历史健康数据,将医学领域知识转化为内部特征,同时利用EHR的顺序信息和医学领域知识来学习更健壮和更准确的患者表示,并在诊断预测任务中得到了更好的预测。受GNDP的启发,本文设计了一种基于图卷积网络和注意力机制的诊断预测模型(Diagnosis Prediction Based on Graph Convolutional Network and Attention Mechanism,GCAP),该模型利用医学领域知识中的分层关系,使用图结构对EHR数据进行建模,利用GCN学习富含医学领域知识的分类代码嵌入向量,采用图卷积操作来提取患者历史就诊信息中的空间特征,通过多头注意力机制来对就诊特征和多级医学知识之间的相互关系进行建模,从而对患者的未来健康状况进行预测。本文的贡献如下:

(1)使用图结构对EHR数据进行建模,不仅包含了来自EHR的序列信息,还体现了医学本体中蕴含的知识,也为进一步学习医学概念的低维向量表示提供了支持。

(2)基于GCN的嵌入方法提供了一种定量方法,在特定医学领域知识的指导下区分相似和不相似的医学概念,可以将医学领域知识中所蕴含的多级语义信息整合到患者就诊的EHR序列中,进一步提高所代表的特征的丰富程度。

(3)设计了一种多头注意力机制并与门控循环单元结合,实现患者就诊特征与医学领域知识的有效融合。

(4)在两个EHR数据集上的实验结果表明,所提出的GCAP的诊断预测性能优于现有的诊断预测方法。

## 2 诊断预测模型

### 2.1 基本符号

我们将来自EHR数据集中所有诊断代码的集合表示为 $D = \{d_1, d_2, \dots, d_{|D|}\}$ ,其中 $d_i$ 对应一个唯一的诊断代码, $|D|$ 是EHR数据集中诊断代码的数量。EHR数据集中的每位患者的健康记录由一系列就诊信息组成,每一次就诊包括诊断代码集中的一个子集。EHR数据集中的患者数可以用 $P$ 来表示,一个就诊次数为 $n$ 的第 $P$ 位患者,其就诊序列可以表示为 $V_1^P, V_2^P, \dots, V_n^P$ ,其中 $V_i^P \subseteq D$ 对应患者 $P$ 的第 $i$ 次就诊所对应的诊断代码。

医学本体是生物医学领域长期的知识积累,在国际疾病分类(International Classification of Diseases, ICD)<sup>[13]</sup>和临床分类软件(Clinical Classifications Software, CCS)<sup>[14]</sup>中,均以亲子关系表达医学概念的层次结构,可将疾病分类看作医学

本体。我们使用CCS作为医学本体 $G$ , $G$ 由分层结构中具有父子关系的不同医学概念组成。 $G$ 中的每个叶子节点对应诊断代码集合 $D$ 中的一个的诊断代码,每个祖先节点则代表与其叶子节点相关的更一般的医学概念,我们将所有祖先节点的集合表示为 $C = \{c_1, c_2, \dots, c_{|C|}\}$ ,其中 $c_i$ 对应一个唯一的祖先代码, $|C|$ 是EHR数据集中祖先代码的数量。一个就诊次数为 $n$ 的第 $P$ 位患者,其就诊的祖先代码序列可以表示为 $S_1^P, S_2^P, \dots, S_n^P$ ,其中 $S_i^P \subseteq \{0, 1\}^{|C|}$ ,如果患者 $P$ 的第 $i$ 次就诊中出现祖先代码 $c_j$ ,那么 $S_{ij}^P = 1$ ,否则 $S_{ij}^P = 0$ 。

目前,预训练分布式表示法广泛应用于各项主流任务中<sup>[15]</sup>。图神经网络不断发展,在处理图数据方面显示出其优越性,图卷积网络<sup>[16]</sup>可以通过提取网络的全局拓扑信息来生成节点嵌入,进一步完善许多网络相关的预测任务。为了学习多层次医学概念的嵌入表示,本文首先构建了一个知识图来建模EHR中医学概念之间的关系。基于CCS,本文利用无向图构造了医学知识图 $G = (V, E)$ ,其中 $V = D \cup C$ 是由EHR中的所有诊断代码和类别代码所构成的顶点集, $E = \{e_1, e_2, \dots, e_{|E|}\}$ 是边集, $|E|$ 表示所有诊断代码和其祖先代码之间的父子关系的数量。此外,未连接的子图通过额外的根节点连接。

我们采用医学知识图来表示序列EHR数据。具体地,每个患者的单次就诊可以由固定结构的无向图 $G = (V, E)$ 来表示,边集 $E$ 可以用对称邻接矩阵 $\mathbf{A} \in \mathbb{R}^{|V| \times |V|}$ 来表示。对于知识图中的每个节点 $V_i$ ,我们分配一个属性标量来指示该节点在单次就诊中的出现,即用1表示其在此次就诊中出现,用0表示未出现,患者 $P$ 的第 $i$ 次就诊即可表示为 $V_i^P \in \mathbb{R}^{|V| \times 1}$ 。一位患者图的输入特征便可以表示为一个维度为 $(C_0, T, N)$ 的张量 $\mathbf{F}$ ,其中, $C_0$ 表示输入通道的数量,即节点特征的维数1; $T$ 表示患者就诊次数; $N$ 表示就诊图中节点的数量。

### 2.2 模型框架

诊断预测的任务是根据患者历史就诊信息 $V_1, V_2, \dots, V_T$ 预测患者第 $T+1$ 次就诊信息,我们提出的诊断预测模型主要包括图嵌入层、图卷积层、循环神经网络层、注意力层和输出层,诊断预测模型的主要框架如图1所示。具体来说,图嵌入层从医学知识图 $G$ 中学习医学概念的嵌入表示,图卷积层从患者的历史就诊中捕获空间特征,循环神经网络层捕获患者就诊的时序特征,注意力层利用多头注意力机制,显式地对历史就诊和高级医学知识之间的相互关系进行建模,自适应地提取所学习到的重要诊断特征来预测下一次就诊信息。

#### 2.2.1 图嵌入层

我们使用两层GCN来学习医学知识图 $G$ 中合理的嵌入表示,该图以顶层或第一层疾病分类节点作为节点分类标签。因为 $G$ 中的节点本身并没有任何属性特征,所以我们将节点所对应的特征矩阵 $\mathbf{X}$ 初始化为单位矩阵。对多层图卷积网络的表述如下:

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

其中, $\sigma(\cdot)$ 表示激活函数, $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}$ 表示具有自连接的医学知识图 $G$ 的邻接矩阵, $\mathbf{I}$ 是单位矩阵。 $\tilde{\mathbf{D}}_{ii} = \sum_j \tilde{\mathbf{A}}_{ij}$ 是图 $G$ 的度矩阵, $\tilde{\mathbf{D}}^{-\frac{1}{2}} \tilde{\mathbf{A}} \tilde{\mathbf{D}}^{-\frac{1}{2}}$ 代表邻接矩阵 $\mathbf{A}$ 的正则化, $\mathbf{W}^{(l)}$ 是第 $l$ 层的可训练权重矩阵, $\mathbf{H}^{(l)}$ 是第 $l$ 层的激活矩阵, $\mathbf{H}^{(0)}$ 表示图 $G$ 节点的初始特征矩阵 $\mathbf{X}$ 。

具体地,本文通过半监督节点分类来学习节点的嵌入

表示,GCN 嵌入模块的具体计算公式如下:

$$\mathbf{X}' = \mathbf{X}\mathbf{W}_{\text{emb}} + \mathbf{b}_{\text{emb}} \quad (2)$$

$$\mathbf{Z} = \text{softmax}(\tilde{\mathbf{A}} \text{ReLU}(\tilde{\mathbf{A}}\mathbf{H}'\mathbf{W}^{(0)})\mathbf{W}^{(1)}) \quad (3)$$

其中,  $\mathbf{W}_{\text{emb}} \in \mathbb{R}^{|\mathcal{C}| \times d_{\text{gen}}}$  和  $\mathbf{b}_{\text{emb}} \in \mathbb{R}^{|\mathcal{C}| \times d_{\text{gen}}}$  是学习的权重矩阵和偏置矩阵,即医学概念嵌入矩阵,  $d_{\text{gen}}$  是嵌入维度的大小;  $\tilde{\mathbf{A}} = \tilde{\mathbf{D}}^{-\frac{1}{2}}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-\frac{1}{2}}$  代表邻接矩阵  $\mathbf{A}$  的正则化;  $\mathbf{W}^{(0)} \in \mathbb{R}^{d_{\text{gen}} \times H}$  是输入层到具有  $H$  个特征的隐藏层的权重矩阵,  $\mathbf{W}^{(1)} \in \mathbb{R}^{H \times H'}$  是隐藏层到输出层的权重矩阵。为了训练嵌入矩阵  $\mathbf{W}_{\text{emb}}$ , 我们使用 Kipf 等<sup>[16]</sup> 使用的交叉熵损失函数:

$$\mathcal{L} = - \sum_{l \in \mathcal{Y}_L} \sum_{f=1} \mathbf{Y}_{lf} \ln \mathbf{Z}_{lf} \quad (4)$$

其中,  $\mathcal{Y}_L$  是有标签的节点索引的集合,  $\mathbf{Y}$  是标签指标矩阵,  $H'$  是输出特征的维度, 它等于类别的数量, 通过上述步骤的训练后, 得到可解释的诊断嵌入矩阵  $\mathbf{W}_{\text{emb}}$ 。

对于患者第  $t$  次的就诊祖先向量  $\mathbf{S}_t \in \{0, 1\}^{|\mathcal{C}|}$ , 将嵌入矩阵  $\mathbf{W}_{\text{emb}}$  与  $\mathbf{S}_t$  相乘, 使用 ReLU 激活函数得到患者第  $t$  次就诊的祖先代码嵌入表示。

$$\mathbf{l}_t = \text{ReLU}(\mathbf{S}_t \mathbf{W}_{\text{emb}}) \quad (5)$$

最后, 就诊次数为  $n$  的患者就诊序列被转换为祖先代码特征矩阵  $\mathbf{L} = [\mathbf{l}_1, \mathbf{l}_2, \dots, \mathbf{l}_n]$ 。

### 2.2.2 图卷积层

将患者历史就诊信息转化为图形式的的数据后, 患者的每一次就诊信息都可以看作是一个图。为了充分利用 EHR 图中的空间信息, 使用 Kipf & Welling<sup>[16]</sup> 提出的近似谱图卷积得到的图卷积作为空间上的卷积操作。具体来说, 以邻接矩阵  $\mathbf{A}$ , 节点特征向量  $\mathbf{F}$  为输入, 通过定义的以下函数可以实现有效高效的卷积运算:

$$\mathbf{F}_{\text{spatial}} = \tilde{\mathbf{D}}^{-\frac{1}{2}} \tilde{\mathbf{A}} \tilde{\mathbf{D}}^{-\frac{1}{2}} \mathbf{F} \mathbf{W} \quad (6)$$

其中,  $\tilde{\mathbf{D}}^{-\frac{1}{2}} \tilde{\mathbf{A}} \tilde{\mathbf{D}}^{-\frac{1}{2}}$  代表邻接矩阵  $\mathbf{A}$  的正则化,  $\mathbf{W}$  是需要训练的权重矩阵。

具体地, 对每个时间步执行空间图卷积操作, 即对维度为  $(C_0, T, N)$  特征向量  $\mathbf{F}$  执行卷积核大小为  $[1, 3]$ 、步长为  $(1, 1)$  的标准 2D 卷积, 得到维度为  $(C_{\text{out}}, T, N)$  的新特征向量  $\mathbf{F}'$ , 2D 卷积相当于输入张量, 是通过将输入矩阵与可学习的权重矩阵  $\mathbf{W}$  相乘并添加偏差  $\mathbf{b}$  获得的。再将得到的  $\mathbf{F}'$  与正则化邻接矩阵  $\tilde{\mathbf{D}}^{-\frac{1}{2}} \tilde{\mathbf{A}} \tilde{\mathbf{D}}^{-\frac{1}{2}}$  相乘实现空间图卷积操作, 得到维度为  $(C_{\text{out}}, T, N)$  的空间特征向量  $\mathbf{F}_{\text{spatial}}$ 。具体描述如下:

$$\tilde{\mathbf{A}} = \tilde{\mathbf{D}}^{-\frac{1}{2}} (\mathbf{A} + \mathbf{I}) \tilde{\mathbf{D}}^{-\frac{1}{2}} \quad (7)$$

$$\mathbf{F}' = \sum \mathbf{F} \cdot \mathbf{W} + \mathbf{b} \quad (8)$$

$$\mathbf{F}_{\text{spatial}} = \text{GCN}(\mathbf{F}) = \tilde{\mathbf{A}} \otimes \mathbf{F}' \quad (9)$$

### 2.2.3 循环神经网络层

我们将得到的特征向量  $\mathbf{F}_{\text{spatial}}$  输入到双向 GRU 神经网络中, 通过一个前向 GRU 神经网络和一个后向 GRU 神经网络来捕捉患者就诊的时序特征, 同时避免了传统 RNNs 梯度消失的问题。具体表述如下:

$$\vec{h}_t = \text{GRU}(\mathbf{F}_{\text{spatial}}, \vec{h}_{t-1}) \quad (10)$$

$$\overleftarrow{h}_t = \text{GRU}(\mathbf{F}_{\text{spatial}}, \overleftarrow{h}_{t-1}) \quad (11)$$

$$\mathbf{h}_t = [\vec{h}_t; \overleftarrow{h}_t] \quad (12)$$

其中,  $\mathbf{h}_{t-1}$  代表隐状态特征向量,  $[\cdot; \cdot]$  代表矩阵的拼接操作。

### 2.2.4 注意力层

为了有效地对祖先代码特征  $\mathbf{l}_t$  和图卷积层得到的患者就诊特征隐状态向量  $\mathbf{h}_t$  进行融合, 并灵活地对就诊特征和多级医学知识之间的相互关系进行建模, 我们设计了一个知识

注意力模块, 该模块利用多头缩放点积注意力<sup>[17]</sup> 来提取关系, 并进一步提高提取和聚合异构特征以进行细粒度诊断预测的能力。我们采用  $h$  头知识注意力来平行关注输入向量的  $h$  个表示子空间, 第  $i$  头知识注意力的映射可以表示为:

$$\mathbf{Q}_i (\text{Query}) = \mathbf{W}_Q \mathbf{H}_i \quad (13)$$

$$\mathbf{K}_i (\text{Key}) = \mathbf{W}_K \mathbf{L}_i \quad (14)$$

$$\mathbf{V}_i (\text{Value}) = \mathbf{W}_V \mathbf{L}_i \quad (15)$$

$$\alpha_i = \text{softmax} \left( \frac{\mathbf{Q}_i \mathbf{K}_i^T}{\sqrt{d_k}} \right) \mathbf{V}_i \quad (16)$$

$$\mathbf{M}_i = \alpha_i \mathbf{V}_i \quad (17)$$

其中,  $d_k$  是隐藏单元的数量;  $\mathbf{W}_Q \in \mathbb{R}^{d \times d_k}$ ,  $\mathbf{W}_K \in \mathbb{R}^{d_{\text{gen}} \times d_k}$ ,  $\mathbf{W}_V \in \mathbb{R}^{d_{\text{gen}} \times d_k}$  分别为知识注意力模块需要学习的参数;  $\alpha_i$  表示第  $i$  头注意力权重;  $\mathbf{M}_i \in \mathbb{R}^{T \times d_k}$  表示 EHR 数据的第  $i$  头知识注意力特征表示,  $T$  表示患者历史就诊次数。

最后, 将  $h$  头注意力并行学习到的特征向量融合在一起构成患者就诊特征的单个向量表示:

$$\mathbf{M} = [\mathbf{M}_1; \mathbf{M}_2; \dots; \mathbf{M}_h] \quad (18)$$

### 2.2.5 输出层

将患者最终的向量表示  $\mathbf{M}$  输入全连接层, 使用 softmax 作为激活函数, 模型的最终输出定义如下:

$$\hat{\mathbf{y}} = \text{softmax}(\mathbf{M} \mathbf{W}_s + \mathbf{b}_s) \quad (19)$$

其中,  $\mathbf{W}_s$  和  $\mathbf{b}_s$  是需要学习的参数。

该诊断预测模型利用 EHR 序列的前  $t$  次就诊, 预测第  $t+1$  次就诊的诊断类别, 我们遵循以往诊断预测的做法, 使用真实值  $\mathbf{y}_{t+1}$  和预测值  $\hat{\mathbf{y}}_{t+1}$  之间的交叉熵来计算每个患者在所有时间步  $T$  下的损失, 具体如下:

$$\mathcal{L}(\mathbf{y}_{t+1}, \hat{\mathbf{y}}_{t+1}) = - \frac{1}{T-1} \sum_{t=1}^{T-1} (\mathbf{y}_{t+1}^T \log(\hat{\mathbf{y}}_{t+1}) + (1 - \mathbf{y}_{t+1}^T) \log(1 - \hat{\mathbf{y}}_{t+1})) \quad (20)$$

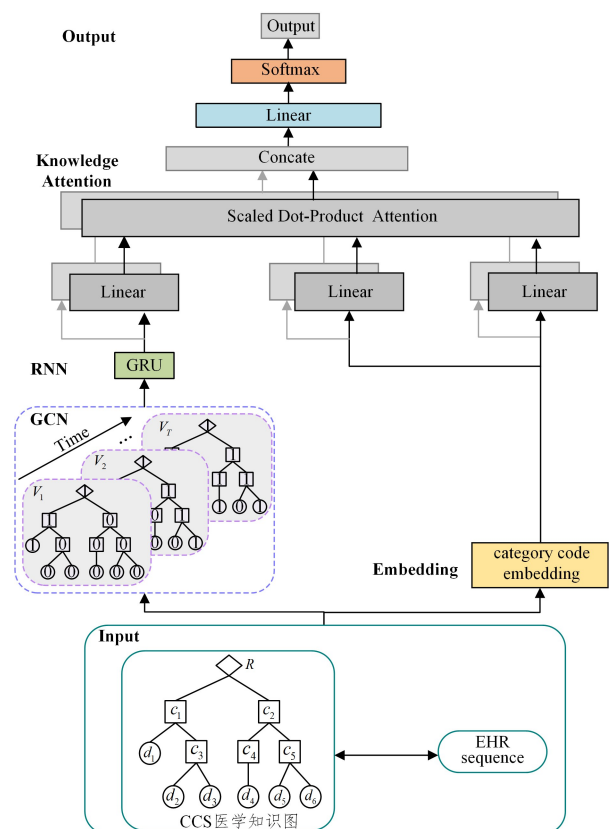


图1 提出的GCAP模型

Fig. 1 The proposed GCAP model

### 3 实验结果与分析

#### 3.1 实验数据集

本文采用两个真实世界公开的数据集来检验本文所提出的诊断模型预测性能。重症监护医学信息数据集三<sup>[18]</sup> (Medical Information Mart for Intensive Care-III, MIMIC-III)是一个大型的 EHR 数据集,它包含了 2001 年 6 月至 2012 年 10 月间在 Beth Israel Deaconess Medical Center 住院的 8645 名成年个体和 7875 名新生儿的 58000 余次住院资料,详细记录了患者在 ICU 治疗期间的几乎所有的数据。重症监护医学信息数据集四<sup>[19]</sup> (Medical Information Mart for Intensive Care-IV, MIMIC-IV)包含了 2008 年至 2018 年之间进入重症监护室的患者的相应数据。

为了通过图神经网络学习医学概念嵌入向量,我们首先利用多级 CCS 分别构建了 MIMIC-III 和 MIMIC-IV 的医学知识图  $G_1$  和  $G_2$ ,两个医学知识图的具体信息如表 1 所列。

表 1 医学知识图信息

Table 1 Statistics of medical knowledge graph

	$G_1$	$G_2$
节点数量	5548	7230
边数量	5547	7229
叶子节点数量	4880	6534
祖先节点数量	668	696
根节点	1	1
一级祖先节点	18	18
二级祖先节点	134	131
三级祖先节点	332	343
四级祖先节点	183	203

对于原始的 MIMIC-III 数据集,本文使用 Retain<sup>[6]</sup> 中提到的数据处理方法来构建按时间排序的患者就诊序列,并筛选出至少有两次就诊记录的患者共 7499 名;随后,通过查询医学知识图  $G_1$ ,得到患者就诊的祖先代码序列,该数据集由较短的就诊信息组成,患者数量较少,可以在训练数据不足的情况下验证模型的性能。对于原始的 MIMIC-IV 数据集,我们筛选出至少有 5 次就诊记录且具有 ICD-9 诊断代码的患者共 8980 名;随后,通过查询医学知识图  $G_2$ ,得到患者就诊的祖先代码序列,该数据集有助于评估预测模型在长 EHR 序列上的表现。筛选后的两个数据集描述信息如表 2 所列。

表 2 数据集信息

Table 2 Datasets statistics

	MIMIC-III	MIMIC-IV
患者数量	7499	8980
就诊次数	19911	77015
每位患者平均就诊次数	2.66	8.58
icd-9 代码数	4880	6534
每次就诊最多 icd-9 代码数	39	39
每次就诊平均 icd-9 代码数	13.06	12.55
icd-9 对应的 CCS 类别数	272	277
每次就诊最多 CCS 类别数	34	37
每次就诊平均 CCS 类别数	11.23	10.79

#### 3.2 评价指标

诊断预测任务可看作多标签分类任务。为了保证诊断预测的细粒度并评估模型的预测性能,本文使用单级 CCS 来

标记每个诊断代码<sup>[8]</sup>。也就是说,预测的标签是单级 CCS 中的诊断类别,而不是具体的诊断代码。为了全面地评估模型的性能,本文采用了在诊断预测中广泛使用的评价指标 Code-level Accuracy@ $k$  和 Visit-level Precision@ $k$ <sup>[10]</sup>。Code-level Accuracy@ $k$  的计算式为:

$$\text{code-level Accuracy}@k = \frac{N_k}{N_{y_{t+1}}} \quad (21)$$

其中, $N_k$  代表第  $t+1$  次就诊的预测结果  $\hat{y}_{t+1}$  中的前  $k$  个正确代码数量, $N_{y_{t+1}}$  代表标签  $y_{t+1}$  中所有患者诊断类别数量。Visit-level Precision@ $k$  则定义为  $N_k$  除以  $\min(k, N_{y_{t+1}})$ ,计算式为:

$$\text{visit-level Accuracy}@k = \frac{N_k}{\min(k, N_{y_{t+1}})} \quad (22)$$

对于所有指标,该值越高,模型的诊断预测性能越好。在本文中,将指标值  $k$  分别取 5, 10, 15 和 20,以分别评估每个模型的性能。

#### 3.3 实验设置

本文采用深度学习框架 TensorFlow2.6 和 Keras 构建神经网络,我们将每个数据集随机划分为训练集(75%)、验证集(10%)和测试集(15%)。所有基线模型的嵌入层中的嵌入矩阵维度和循环神经网络层中的隐状态维度均设置为 128;在图卷积层中,卷积核个数设置为 128,卷积核尺寸设置为  $1 \times 3$ ;循环神经网络层使用 dropout 策略<sup>[20]</sup>来缓解模型过拟合问题,且丢弃率(dropout rate)设为 0.5;多头注意力中的头数设置为 2;使用随机梯度下降 Adam 算法<sup>[21]</sup>作为所有模型的优化器,初始学习率(learning rate)设为 0.0005;使用随机 Mini-batch 算法<sup>[22]</sup>来优化训练过程,其中的小批量训练参数在 MIMIC-III 和 MIMIC-IV 两个数据集中分别设置为 64 和 32。对于基线模型的其他参数,我们按照其论文中推荐的参数进行设置。

#### 3.4 性能比较

为了评估本文所构建诊断预测模型的性能,我们使用 RNN+<sup>[7]</sup>, Dipole<sup>[7]</sup>, GRAM<sup>[8]</sup>, KAME<sup>[9]</sup> 和 CAMP<sup>[10]</sup> 这 5 个基线模型进行对比实验,其中 RNN+ 和 Dipole 两个模型直接根据患者的历史就诊诊断信息进行诊断预测,GRAM, KAME 和 CAMP 这 3 个模型则额外融合了医学知识图中的相关信息进行更准确的诊断预测。

我们实现了上述模型,并获得了它们的 Code-level Accuracy@ $k$  和 Visit-level Precision@ $k$ ,结果如表 3 所列。表 3 中的结果表明,对于基线模型,RNN+ 和 Dipole 在诊断预测任务中都没有使用外部医学知识,只使用 GRU 从 EHR 序列中学习就诊特征,RNN+ 模型在两个数据集上表现最差,Dipole 表现优于 RNN+,但预测性能仍然不理想。GRAM, KAME 和 CAMP 预测性能均优于 RNN+ 和 Dipole,这 3 个模型利用额外的资源来生成医学概念的嵌入,但它们只利用从 CCS 得到的诊断代码之间的共现关系,忽略了同样重要的医学知识图结构。与所有基线模型相比,GCAP 同时利用了双向 GRU 和注意力机制的优点。具体而言,GCAP 的优势在于注意力机制可以提取和聚合长期依赖关系用于进一步诊断,并具有建模就诊特征和多层次医学领域知识之间复杂关系的卓越能力,使 GCAP 在所有评估指标上都优于基线。

表3 GCAP模型与基线模型的性能比较

Table 3 Performance comparison between GCAP model and baseline models

Dataset	Model	Code-level Accuracy@k				Visit-level Precision@k			
		5	10	15	20	5	10	15	20
MIMIC-III	RNN+	0.2727	0.4184	0.5194	0.5958	0.5909	0.5272	0.5487	0.6026
	Dipole	0.2813	0.4321	0.5314	0.6054	0.6092	0.5454	0.5619	0.6124
	GRAM	0.2928	0.4368	0.5341	0.6068	0.6270	0.5503	0.5646	0.6138
	KAME	0.2975	0.4448	0.5428	0.6168	0.6408	0.5614	0.5740	0.6238
	CAMP	0.3038	0.4664	0.5696	0.6362	0.6677	0.5906	0.6014	0.6413
	GCAP	<b>0.3749</b>	<b>0.5911</b>	<b>0.7031</b>	<b>0.7623</b>	<b>0.7782</b>	<b>0.7330</b>	<b>0.7412</b>	<b>0.7723</b>
MIMIC-IV	RNN+	0.3462	0.5146	0.6177	0.6884	0.6880	0.6319	0.6491	0.6951
	Dipole	0.3569	0.5165	0.6102	0.6743	0.7089	0.6344	0.6411	0.6808
	GRAM	0.3581	0.5293	0.6311	0.6971	0.7121	0.6501	0.6631	0.7038
	KAME	0.3618	0.5405	0.6423	0.7079	0.7179	0.6633	0.6750	0.7148
	CAMP	0.3777	0.5598	0.6632	0.7281	0.7267	0.6765	0.6929	0.7337
	GCAP	<b>0.5349</b>	<b>0.8044</b>	<b>0.9159</b>	<b>0.9520</b>	<b>0.9679</b>	<b>0.9582</b>	<b>0.9576</b>	<b>0.9614</b>

### 3.5 消融实验

消融研究可以评估机器学习模型中模块的功效,尤其是在复杂的神经网络中。对于本文提出的GCAP,有两个模块值得研究和分析。第一个是注意力模块,其为GCAP提供就诊特征和医学领域知识聚合能力,通过将注意力模块从

GCAP中去掉,我们得到了第一个消融模型GCAP-A。第二个是图卷积模块,其为GCAP提供捕获患者的历史就诊特征能力,通过去除图卷积模块,我们得到了第二个消融模型GCAP-G。我们在MIMIC-III上进行GCAP的消融实验,具体的诊断预测性能如表4所列。

表4 MIMIC-III数据集上的消融实验结果

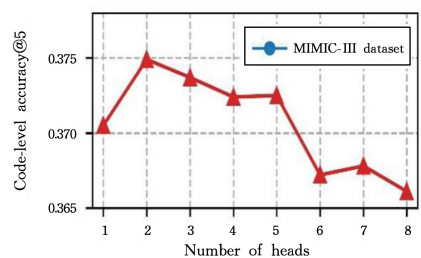
Table 4 Results of ablation experiments on MIMIC-III dataset

Model	Code-level Accuracy@k				Visit-level Precision@k			
	5	10	15	20	5	10	15	20
GCAP	0.3749	0.5911	0.7031	0.7623	0.7782	0.7330	0.7412	0.7723
GCAP-A	0.3445	0.5254	0.6288	0.6959	0.7218	0.6553	0.6631	0.7030
GCAP-G	0.3620	0.5655	0.6768	0.7381	0.7548	0.7037	0.7132	0.7462

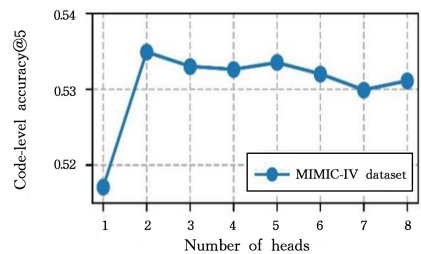
表4中的结果表明,注意力模块和图卷积模块都提高了诊断预测的性能。通过对比GCAP和GCAP-A,我们发现加入知识注意力模块可以在Code-level Accuracy@{5, 10, 15, 20}带来3.04%~7.43%的性能提升,在Visit-level Precision@{5, 10, 15, 20}带来5.62%~7.81%的性能提升。通过比较GCAP和GCAP-G,我们发现,在没有图卷积模块的情况下,GCAP在Code-level Accuracy@{5, 10, 15, 20}上降低了1.29%~2.63%的预测精度,在Visit-level Precision@{5, 10, 15, 20}上降低了2.34%~2.92%的预测精度。此外,通过比较GCAP-A和GCAP-G,我们可以直观地发现注意力模块对提高诊断预测性能的影响更大。

### 3.6 多头注意力的有效性

为了分析多头注意力中的头数对性能的影响,我们在多头注意力中设置了不同的头数,并在MIMIC-III和MIMIC-IV两个数据集上观察所提模型的性能。图2给出了GCAP在两个数据集上的预测性能随着不同头数而变化的情况。图2中的结果表明,随着多头注意力中的头数的增加,GCAP的预测性能整体呈现出先增后减的趋势。具体地,在MIMIC-III和MIMIC-IV两个数据集中,当头数为1时,两个数据集上的模型预测性能都比较差;当头数都设置为2时,GCAP在两个数据集中均实现了最佳的预测性能,随后,头数增加,预测性能并没有得到提高。这也证实了在GCAP模型中多头注意力模块的有效性,以及合理选择头数对GCAP的预测性能的重要性。



(a) MIMIC-III dataset



(b) MIMIC-IV dataset

图2 GCAP在两个数据集上不同头数的Code-level Accuracy@5

Fig. 2 Code-level Accuracy@5 of DCAP on two datasets with different number of heads

**结束语** 诊断预测是个性化医疗的核心任务,但为细粒度的诊断预测设计准确、可靠且可解释的预测模型仍是一个具有挑战性的课题。为了应对这一挑战,本文设计了一个基于图卷积和多头注意力机制的诊断预测模型GCAP。GCAP利用医学领域知识和电子健康记录信息,构建患者的时序图数据,能够进一步将医学领域知识转化为EHR数据的内部

特征,提高诊断预测性能。此外,GCAP 有效地整合了图卷积网络和注意力机制的优势,进一步提高了预测性能。在 MIMIC-III 和 MIMIC-IV 两个公共 EHR 数据集上的实验结果表明,本文模型预测性能优于已有诊断预测模型。但是,EHR 序列包括与患者健康相关的各种背景,如人口统计信息、体检、诊断、测试、程序、治疗、药物治疗、出院、病史等临床信息,而 GCAP 目前只考虑了 EHR 序列中的诊断代码,就临床需要而言,GCAP 需要扩展以考虑更多的临床信息。此外,我们只利用 ICD-9-CM 的临床分类软件(CCS)来构建知识图,GCAP 还可以探索其他合适的医学数据集和医学领域知识进一步提高诊断预测的性能。

## 参考文献

- [1] YADAV P,STEINBACH M,KUMAR V,et al. Mining Electronic Health Records (EHRs): A Survey[J]. ACM Comput. Surv.,2018,50(6):85.
- [2] BOTSI T,HARTVIGSENG G,CHEN F,et al. Secondary use of EHR: data quality issues and informatics opportunities[J/OL]. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3041534/>.
- [3] JENSEN P B,JENSEN L J,BRUNAK S. Mining electronic health records: towards better research applications and clinical care[J]. Nature Reviews Genetics,2012,13(6):395-405.
- [4] KRUSE C S,STEIN A,THOMAS H,et al. The use of Electronic Health Records to Support Population Health: A Systematic Review of the Literature[J]. Journal of Medical Systems, 2018,42(11):214.
- [5] CHOI E,BAHADORI M T,SCHUETZA,et al. Doctor ai: Predicting clinical events via recurrent neural networks[C]// Machine Learning for Healthcare Conference. PMLR, 2016: 301-318.
- [6] CHOI E,BAHADORI M T,SUN J,et al. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism[J]. arXiv:1608.05745,2016.
- [7] MA F,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, 2017:1903-1911.
- [8] CHOI E,BAHADORI M T,SONG L,et al. GRAM: graph-based attention model for healthcare representation learning [C]//Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2017: 787-795.
- [9] MA F,YOU Q,XIAO H,et al. Kame: Knowledge-based attention model for diagnosis prediction in healthcare[C]// Proceedings of the 27th ACM International Conference on Information and Knowledge Management, 2018:743-752.
- [10] GAO J,WANG X,WANG Y,et al. Camp: Co-attention memory networks for diagnosis prediction in healthcare[C]//2019 IEEE International Conference on Data Mining (ICDM), IEEE,2019:

1036-1041.

- [11] LI H,LI W H,CHEN W,et al. Diagnostic prediction based on Node2vec and knowledge attention mechanisms[J]. Computer Science,2021,48(S2):630-637.
- [12] LI Y,QIAN B,ZHANG X,et al. Knowledge guided diagnosis prediction via graph spatial-temporal network[C]//Proceedings of the 2020 SIAM International Conference on Data Mining, 2020:19-27.
- [13] TROTT P. International classification of diseases for oncology [J]. Journal of Clinical Pathology,1977,30(8):782.
- [14] COST H,PROJECT U. Clinical classifications software (CCS) for ICD-9-CM[J]. [www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp](http://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp).
- [15] BENGIO Y,COURVILLE A,VINCENT P. Representation Learning: A Review and New Perspectives[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(8):1798-1828.
- [16] KIPF T N,WELLING M. Semi-supervised classification with graph convolutional networks[J]. arXiv:1609.02907,2016.
- [17] VASWANI A,SHAZEER N,PARMAR N,et al. Attention is all you need[J]. arXiv:1706.03762,2017.
- [18] JOHNSON A E W,POLLARD T J,SHEN L,et al. MIMIC-III, a freely accessible critical care database [J]. Scientific Data, 2016,3(1):1-9.
- [19] JOHNSON A,BULGARELLI L,POLLARD T,et al. MIMIC-IV, a freely accessible electronic health record dataset[J/OL]. <https://physionet.org/content/mimiciv/1.0/>.
- [20] SRIVASTAVA N,HINTON G,KRIZHEVSKY A,et al. Dropout: a simple way to prevent neural networks from overfitting [J]. The journal of machine learning research, 2014, 15(1): 1929-1958.
- [21] KINGMA D P,BA J. Adam: A method for stochastic optimization[J]. arXiv:1412.6980,2014.
- [22] LI M,ZHANG T,CHEN Y,et al. Efficient mini-batch training for stochastic optimization[C]//Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2014:661-670.



**YANG Xianming**, born in 1999, post-graduate. Her main research interests include deep learning and bioinformatics.



**DING Haiyan**, born in 1974, master, associate professor. Her main research interest is intelligent data processing.