计算机科学
COMPUTER SCIENCE

基于多源健康感知数据动静态关系融合的疾病诊断

霍甜媛，顾晶晶

引用本文

霍甜媛，顾晶晶. 基于多源健康感知数据动静态关系融合的疾病诊断[J]. 计算机科学, 2022, 49(11A): 211100241-9.

HUO Tian-yuan, GU Jing-jing. Dynamic and Static Relationship Fusion of Multi-source Health Perception Data for Disease Diagnosis [J]. Computer Science, 2022, 49(11A): 211100241-9.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.211100241

基于多源健康感知数据动静态关系融合的疾病诊断

霍甜媛　顾晶晶
南京航空航天大学计算机科学与技术学院　南京 211106
(huotianyuan@nuaa.edu.cn)

摘　要　疾病诊断是电子健康记录数据挖掘的热门研究领域,也是实现医疗诊断智能化的一个重要环节。但是,电子健康记录中健康感知数据的来源多样、数据结构复杂,且不同类型的数据之间有着潜在的相关性,在进行特征提取和挖掘分析过程中存在着异构数据应该如何融合的问题。只有对医学感测数据、个人体质记录数据、疾病间关系数据进行综合考虑,挖掘其中的相关隐藏特征,才能对多种类别疾病进行更准确的诊断。因此,基于多源健康感知数据动静态关系融合的疾病诊断模型(DSRF)首先通过动静态关系融合算法解决动态医学感测数据和静态体质记录数据的异构性问题并挖掘其相关关系,然后计算多类别疾病的关联矩阵来提取疾病间依赖关系,最后在门控循环单元网络架构的基础上将多种健康感知数据进行融合,完成了多源异构数据的综合分析。在美国 MIMIC-III 临床数据集上的实验结果证明,相比同类型主流模型,该模型可以更准确地对多种类别疾病进行联合诊断。

关键词:多源数据融合;动静态关系融合;疾病诊断;电子健康记录;临床数据挖掘
中图法分类号　TP399

Dynamic and Static Relationship Fusion of Multi-source Health Perception Data for Disease Diagnosis

HUO Tian-yuan and GU Jing-jing
School of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China

Abstract　Disease diagnosis is a field of electronic health record data mining where lots of researchers are interested in, and it is also an important link to realize the intellectualization of medical diagnosis. However, due to the diversity of data sources, complex data structure and potential correlation among different types of health sensing data, there is a problem of how to fuse heterogeneous data in the process of feature extraction and data mining. Therefore, comprehensively considering clinical sensing data, personal physical record data and relationship data between diseases, and mining the latent relevant features can make the diagnosis of multi-category diseases more accurate. Dynamic and static relationship fusion of multi-source health perception data for disease diagnosis(DSRF) is proposed. Firstly, the dynamic and static relationship fusion algorithm is used to extract data correlation features and solve the heterogeneity of dynamic clinical sensing time series data and static personal physical condition data. Then the dependency matrix of multi-category diseases is calculated to extract the correlations among diseases. Finally, various health sensing data is fused based on the gated recurrent unit network. The comprehensive analysis of multi-source heterogeneous data is completed after the above three steps. Experimental results on the real-world American MIMIC-III clinical dataset show that the proposed model outperforms state-of-the-art models and is able to diagnose multiple categories of diseases accurately.

Keywords　Multi-source data fusion, Dynamic and static relationship fusion, Disease diagnosis, Electronic health record, Clinical data mining

1　引言

近年来,传感技术在我国得到了快速发展,医学感测仪器随之崛起,生物传感等高新技术的出现及应用,使得临床监测设备不断向测量速度更快、准确度更高的方向发展[1]。由于我国人口众多,每年入院治疗人数以亿计量,必然会产生大量包含医疗感测数据、住院信息、个人体质状况和疾病诊疗记录等的电子健康记录(Electronic Health Record,EHR)。电子健康记录数据电子化、数字化的特点为医生查阅病人信息、快速做出准确的疾病诊断和合理的治疗决策提供了便利,但其数据的复杂性给这些数据的处理和利用带来了很大的难度。随着人工智能技术的不断进步和计算机计算能力的提升,利用电子健康系统进行疾病诊断和风险评估成为可能,这给医学诊断智能化提供了良好的发展机遇。对电子健康记录的合理使用可以为医生提供重要的临床疾病诊断依据和健康干预措施参考,在有限的资源下提高医务工作者的工作效率与医疗机构的服务质量[2]。

为了更好地利用大量的电子健康记录数据,对不同类型

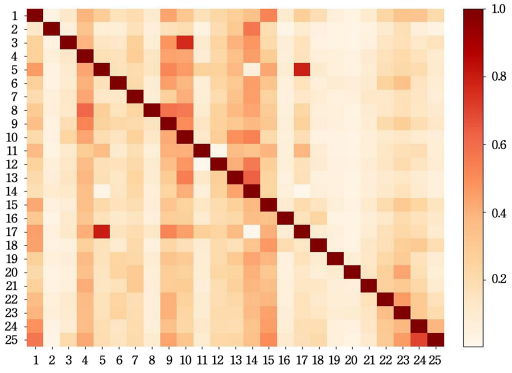
基金项目:国家自然科学基金(62072235)
This work was supported by the National Natural Science Foundation of China(62072235).
通信作者:顾晶晶(gujingjing@nuaa.edu.cn)

结构的数据进行处理和分析十分必要。先前的工作大多仅针对相似结构的数据进行研究，如 Mohammad 等[3]选取了大量的时序体征数据，采用逻辑回归和递归神经网络模型来预测患者在未来 3 个月内是否会出现严重的高血压风险，Ma 等[4]提出了一种可以学习患者生理特征的长期和短期变化作为临床特征的模型，使用医学感测数据来评估患者在不同时间阶段的健康状况。Ayon 等[5]使用深度神经网络对多项诊断测量数据进行学习，将其用于糖尿病的预测。但是，基于健康感知数据对疾病诊断的工作还存在如下问题。

首先，上述工作对不同类型数据分别进行特征提取与学习，而在对多种类型数据进行特征融合以提取数据间的相关关系方面有所欠缺。在本文任务中，由于电子健康记录数据中的医学感测数据是由多种设备获取的动态时序数据，而个人体质记录数据（如年龄、性别等）是静态非时序数据，和医学感测数据之间存在异构性，并且拥有相似体征状况的不同个体所患疾病类型不尽相同。因此，如何对这些异构数据融合建模是疾病诊断任务的关键点。本文针对上述问题提出了一种多源健康感知数据动静态关系融合策略，挖掘多源异构数据间的隐藏特征，以获得更好的模型学习效果。

其次，多类别疾病诊断属于多标签分类任务，直觉上充分利用标签之间的关系可以提升模型性能。以本文选取的疾病为例，如果原发性高血压长期得不到控制，就容易导致各种并发症，因此，原发性高血压和高血压伴并发症一般处于不同的患病阶段，同时出现的可能性较低。而高血压并发症的常见病症之一就是肾脏疾病，因此高血压伴并发症与慢性肾病同时发生的概率较大。

图 1 给出了本文进行多类别疾病诊断任务实验选取的 25 类疾病之间的相关关系，不同编号表示不同类型疾病，疾病相关关系的计算方法将在 3.3 节进行介绍。其中，矩形颜色越深表示两种疾病之间的关系越强，矩形颜色越浅表示两种疾病之间的关系越弱。



注：(1)急性和非特定肾衰竭；(2)急性脑血管病；(3)急性心肌梗死；(4)心律失常；(5)慢性肾病；(6)慢性阻塞性肺疾病；(7)外科/医疗护理的并发症慢性肾病；(8)传导障碍；(9)充血性心力衰竭；(10)冠状动脉粥样硬化及其相关因素；(11)糖尿病及其并发症；(12)无并发症的糖尿病；(13)脂质代谢紊乱；(14)原发性高血压；(15)液体和电解质紊乱；(16)消化道出血；(17)高血压伴并发症；(18)其他肝病；(19)其他下呼吸道疾病；(20)其他上呼吸道疾病；(21)胸膜炎/气胸/肺塌陷；(22)肺炎；(23)呼吸衰竭；(24)败血症；(25)休克

图 1 疾病关系热力图

Fig. 1 Heat map of disease relationship

从图中可以看到，除了每种疾病和自身的关系最强之外，原发性高血压(14)与高血压伴并发症(17)的关系非常弱，而高血压伴并发症(17)和慢性肾病(5)有较强的相关关系，这也符合上述分析。为了能充分利用这一特征，本文设计了将时序特征和疾病关联矩阵融合的模型结构，以进一步提升模型性能。

因此，本文针对多类别疾病诊断任务，挖掘多源健康感知数据间的相关关系与疾病间关系并融合到模型中，构建了基于多源健康感知数据动静态关系融合的疾病诊断模型。具体来说，本文首先对电子健康记录数据中的多源数据进行分析，设计了数据动静态关系融合算法来提取异构数据的关系特征，从而解决多源健康感知数据的异构性问题和关系表示问题；然后构建了疾病关联矩阵来提取疾病间的相关特征；最后提出了一种基于多源健康感知动静态关系数据融合的疾病诊断模型，依次对医学感测数据、个人体质记录数据和疾病关系数据进行融合，从而提高分类任务的准确性。本文的主要研究内容如下。

(1)针对电子健康记录中感知数据来源存在多样性、数据类型具有复杂性，易导致疾病检测任务中不同类型数据之间相关性难以挖掘的问题，本文提出了一种多源健康感知数据动静态关系融合模型。该模型通过添加掩码向量作用于医学感测数据，在门控循环单元网络架构的基础上对医学感测数据和个人体质记录数据进行融合学习，挖掘多源异构数据的隐藏特征，从而实现多类别疾病的诊断。

(2)针对疾病诊断普遍存在的个人体质记录数据对医学感测数据的隐含关系，本文挖掘了个人体质记录静态数据和医学感测动态数据之间的相关性，并在提取动静态数据相互作用关系的同时保留原始数据特征，以帮助模型根据不同患者的体质状况对其所患疾病进行差异性诊断。

(3)针对多类别疾病诊断任务间的疾病依赖和疾病互斥关系难以联合挖掘的问题，本文构建了关联矩阵来提取疾病间关系，并将疾病关系数据与医学感测数据、个人体质记录数据特征进行融合，从而提高了模型的分类准确度。

(4)通过在真实的美国 MIMIC-III 临床数据集[6]上进行实验，证明了本文模型可以准确地对多种类别疾病进行诊断，并优于最好的基准模型。

本文第 1 节阐述了疾病诊断任务存在的问题与意义；第 2 节介绍了目前通过 EHR 数据进行疾病诊断工作的研究现状；第 3 节详细讲解了本文提出的基于多源健康感知数据动静态关系融合模型框架与实现算法；第 4 节将本文模型与基准模型进行实验对比，证明了本文模型的有效性；最后总结全文并展望未来。

2 相关工作

由于电子健康记录数据结构的多样性和复杂性，近年来大量研究人员针对其不同特点进行研究。以下将从电子健康记录数据的适用任务、多源健康感知数据融合算法以及现有的疾病诊断模型 3 个方面来阐述国内外相关工作的具体研究内容。

2.1 电子健康记录数据的适用任务

目前，电子健康记录(EHR)系统被应用在各大医院的

医疗保健服务中，为广大医生和患者提供便利。为了保证记录的完整性和准确性，EHR 系统支持多种类型的记录方式和不同结构的医学数据，例如，MIMIC-III 数据库[6]中的数据包括生命体征、服用药物、检验结果、手术代码、诊断代码、影像报告、住院时间等，给相关研究提供了许多方向。

研究工作[7-9]表明，机器学习模型在死亡率预测和重症监护室住院时间预测方面取得了良好的效果。Pirracchio[10]提出了超级学习器算法，该算法利用一种机器学习模型来增强重症监护患者院内死亡率的预测性能。Sanjay 等[11]对 MIMIC-III 数据集进行了多方面的分析，提出了一组基准任务，包括住院死亡率预测、住院时间预测和 ICD-9 代码组预测，并使用深度学习模型、几种机器学习模型和 ICU 评分系统对 3 种临床预测任务进行了基准评估。在此基础上，Harutyunyan 等[12]使用深度模型对 MIMIC-III 数据集上的 4 项临床预测基准任务做了对比实验，增加的基准任务为用 24 h 内死亡率表示的得失代偿预测。Lee 等[13]提出了多尺度间隔模式感知网络（Multi-Scal Interval Pattern-Aware Network，MSIPA），通过卷积网络挖掘时序 EHR 数据中短、中、长时间间隔模式信息，利用缩放点积注意力机制获取与 3 种时间间隔模式对应的上下文，使用 Transformer 进行 ICU 病房转移预测。

2.2 多源电子健康记录数据融合模型

电子健康记录（EHR）的一个重要特点是其多源健康感知数据结构的复杂性。EHR 通常同时具有稀疏和不规则特征的结构化（代码）和非结构化（自由文本）数据。在深度学习中，不同模态表示的数据应该如何融合在一起是一个难题。

Choi 等[14]研究了在 EHR 数据上执行监督预测任务的同时共同学习多源 HER 数据隐藏结构的可能性。具体而言，文章分析了学习隐藏的 EHR 结构的合适模型，并提出了使用数据统计方法的图卷积 Transformer 模型。该模型在仿真数据和公共可用 EHR 数据的图构建、再入院预测和死亡率预测等任务中取得了不错的效果，是一种有效的多模态数据融合方法。Xu 等[15]扩展了最先进的神经结构搜索（Neural Architecture Search，NAS）方法，并提出了多模态融合架构搜索（Mutimodal Fusion Architecture Search，MUFASA）方法，同时搜索多模态融合策略和模型的特定架构。

2.3 基于机器学习和深度学习的疾病诊断模型

传统的多类别疾病研究一般通过基于专家预定义标准的数据统计分析来完成，疾病判别主要使用简单的医学定义，如 ICD-9 代码。电子健康记录系统的采用使得人们对多类别疾病预测的机器学习方法越来越感兴趣，包括分类[16-17]和聚类[18-19]方法。

随着深度网络模型的逐渐成熟和计算机计算能力的不断提高，多类别疾病诊断成为了近年来深度学习研究者的热门方向。Che 等[20]首先提出将前馈神经网络应用于临床时间序列数据中生理特征模式的探索和疾病检测，并对标准神经网络训练进行了修改，充分利用医学数据的独有特性。Choi 等[21]开发了基于递归神经网络时间模型的人工智能医生，可以从大量患者的时间序列记录中学习有效的特征表示，预测患者未来可能患的疾病。Razavian 等[22]针对早期疾病发作检测任务，提出了两个新的卷积神经网络模型和一个长短期记忆递归神经网络的大规模应用模型。上述研究表明，这些基于自动特征学习的深度学习方法明显优于手动提取临床特征的逻辑回归算法。Ma 等[23]提出的 ConCare 通过具有时间感知注意力机制的多通道 GRU 提取临床特征，获取静态基线信息和动态特征之间的相互依赖关系，构建健康上下文并重新编码临床信息，以根据病史预测患者的临床健康状况。

3 基于多源健康感知数据动静态关系融合的疾病诊断模型

本文深入研究了多源异构数据的处理及融合技术、疾病相关特征提取技术和多类别疾病诊断技术等，进而提出一种基于多源健康感知数据动静态关系融合的疾病诊断模型。该模型通过数据动静态关系融合算法提取异构数据的关系特征，解决了多源健康感知数据的异构性问题，并有效提取异构数据间关系特征，在门控循环单元网络架构的基础上依次对动态医学感测数据、静态个人体质记录数据和疾病相关关系数据进行融合，充分挖掘多源异构数据的隐藏特征，从而提高分类任务的准确性。

本文模型框架主要包括 4 个部分：医学感测数据处理、个人体质记录数据处理、疾病关联矩阵计算和多类别疾病诊断，如图 2 所示。

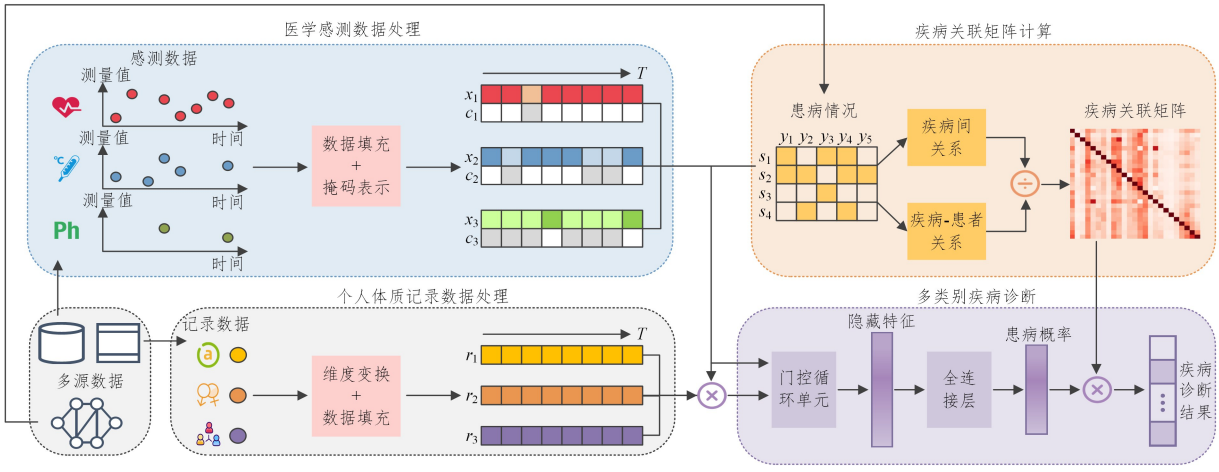


图 2 本文模型框架

Fig. 2 Framework of the proposed model

(1)在处理感测数据的过程中,由于数据的采样频率不同,为兼顾数据长度和真实数据比例,本文以 1 h 为间隔对数据进行划分,对于采样数不足的数据,按优先顺序采用向后填充或特定值填充的方法进行时间补全,并添加掩码特征来标记真实数据和填充数据,即图 2 的感测数据处理模块中白色和灰色方框分别表示该时间的数据为真实数据或填充数据。由于性别、年龄等个人体质记录数据会对体征的正常值产生影响,因此将其融合到感测数据的每个时间步中。

(2)记录数据的填充方式与感测数据类似,将记录数据扩充到和感测数据等长的时间维度,并使用真实值进行补全。随后,对于全部的感测数据和记录数据,本模型采用标准差标准化方法处理连续型数据,采用独热编码方式表示离散型数据。

(3)疾病关联矩阵计算模块通过统计不同疾病的患病人数和任意两类疾病的共患病人数,提取疾病间关系和疾病-患者间关系特征,计算得到疾病关联矩阵,表示任意两类疾病的相关关系,如依赖关系和互斥关系。

(4)在多类别疾病诊断模块中,本文对医学感测数据、个人体质记录数据和疾病关系数据 3 种异构数据在模型的不同层次上进行融合。首先将感测数据和记录数据在时间维度上进行特征提取与融合,并将其送入门控循环单元网络进行训练,提取数据的时间依赖,得到隐藏特征的向量表示,完成了第一层次的数据融合。然后全连接层将提取的数据特征向量转换成和疾病类别数相同的维度,表示待分类疾病的先验概率。最后将基于患者的患病情况数据提取的疾病关联矩阵和基于感测数据及记录数据提取的待分类疾病的先验概率相乘,进行第二层次的数据融合,选择合适的激活函数映射得到最终的多类别疾病诊断结果。

3.1 多类别疾病诊断任务建模

本文对多类别疾病诊断任务进行数学建模。原始数据集 $X=\{X_M,X_R,X_D\}$,其中 $X_M\in\mathbb{R}^{T\times d_m}$ 表示医学感测数据,该数据具有 d_m 种属性,如心率、血压等;T 为时间序列长度;$X_R\in\mathbb{R}^{d_r}$ 表示个人体质记录数据,该数据具有 d_r 种属性,如性别、年龄等;$X_D\in\mathbb{R}^K$ 表示数据集中患者的患病情况数据,K 为待分类的疾病数。假设多类别疾病诊断问题为 F,分类结果 $Y\in\mathbb{R}^K$,因此,这一问题可建模为如式(1)所示:

$$\{X_M;X_R;X_D\}\xrightarrow{F}Y \tag{1}$$

3.2 基于掩码结构的数据填充

为了解决医学感测数据采样频率不一致造成的数据长度不匹配的问题,设定统一的时间间隔为 1 h,并在规则间隔内对时间序列数据重新采样,如果在同一时间间隔内存在同一特征的多个测量值,则使用最后一个测量值。由于部分特征原始采样频率远大于 1 h,按统一的时间间隔重新采样会产生数据稀疏的问题,因此需要采取有效措施对缺失数据进行填充。为此,本文采取的填充方法是,如果某缺失值在其先前时间内存在测量值,以先前最近的测量值代替缺失值,否则使用预先设定值。此外,为每个特征提供一个二进制掩码向量,用于表示其不同时间对应的特征值是否为真实测量值。数据填充和掩码向量构建完成后,采用独热编码表示离散型特征,使用标准差标准化方法处理连续型特征。对 t 时刻的医学感测数据 $X_M^{(t)}$ 进行标准差标准化处理得到 x_t 的计算式如下:

$$\mu=\frac{1}{T}\sum_{t=1}^{T}X_M^{(t)} \tag{2}$$

$$\sigma=\sqrt{\frac{1}{T-1}\sum_{t=1}^{T}(X_M^{(t)}-\mu)^2} \tag{3}$$

$$x_t=\frac{X_M^{(t)}-\mu}{\sigma} \tag{4}$$

经过独热编码、标准化和掩码表示等数据预处理步骤后,将感测数据向量和二进制掩码向量进行拼接,可获得动态时间序列数据 $X_{dy}\in\mathbb{R}^{T\times d_{in}}$:

$$X_{dy}=[(x_1,c_1),(x_2,c_2),\cdots,(x_m,c_m)] \tag{5}$$

其中,$x_m(t=1,2,\cdots,d_m)$ 为属性 m 的独热编码表示或标准化特征值,$\boldsymbol{c}_m(m=1,2,\cdots,d_m)$ 为属性 m 对应的二进制掩码向量,T 为时间序列长度,d_{in} 为数据特征维度与掩码向量维度之和。

3.3 基于条件概率的疾病相关关系挖掘

由于疾病及其症状表现的复杂性和多样性,医学上采用 6 位字母数字组合的 ICD 编码[6]来表示不同的疾病,而对所有疾病进行分类无疑会带来巨大的时间和空间开销。为此,本文按照文献[12]所述方式选取了发病率较高的一些疾病,并将其归为 K 类,判断患者患有其中的哪些疾病。经统计,患者的平均患病数为 4.13,超过 83%的患者患有至少两种以上上述疾病。因此,有效提取疾病间的相互影响关系并将其结合到模型中,可以提高模型的学习能力。本文利用患者的患病情况数据,构建疾病关联矩阵 $\boldsymbol{I}\in\mathbb{R}^{K\times K}$。疾病关联矩阵的计算过程如下。

首先,统计各类疾病的患病人数:

$$S_k=\sum_{n=1}^{N}y_{nk} \tag{6}$$

其中,y_{nk} 为第 n 名患者是否患有疾病 k 的二进制值,1 表示患有此类疾病,0 表示没有患此类疾病。因此可以得到各类疾病的患病率 $P(S_k)$:

$$P(S_k)=\frac{S_k}{N} \tag{7}$$

K 类疾病的患病率共同组成了疾病-患者间关系向量 $\boldsymbol{w}$:

$$\boldsymbol{w}=\{P(S_1),P(S_2),\cdots,P(S_K)\} \tag{8}$$

然后统计同时患有任意两类疾病的人数,令 S_{jk} 表示同时患有疾病 j 和疾病 k 的患者数,则:

$$S_{jk}=\sum_{n=1}^{N}y_{nj}y_{nk} \tag{9}$$

$$P(S_{jk})=\frac{S_{jk}}{N} \tag{10}$$

其中,y_{nj} 和 y_{nk} 分别为第 n 名患者是否患有疾病 j 和疾病 k 的二进制值,$P(S_{jk})$ 表示疾病 j 和疾病 k 的共同患病概率,得到了疾病间关系矩阵 $\boldsymbol{v}$:

$$\boldsymbol{v}=\{P(S_{jk})\mid 1\leqslant j,k\leqslant K\} \tag{11}$$

由上述计算公式可以看出,$\boldsymbol{v}$ 为对称矩阵。

最后是计算疾病关联矩阵 $\boldsymbol{I}$。由于不同疾病的发病率差异较大,为了避免发病率低的疾病在计算相关关系时数值较小的问题,采用条件概率计算疾病 j 对疾病 k 的影响 I_{jk},即:

$$I_{jk}=P(k\mid j)=\frac{P(S_{jk})}{P(S_j)} \tag{12}$$

$$\boldsymbol{I}=\{I_{jk}\mid 1\leqslant j,k\leqslant K\} \tag{13}$$

图 1 给出了通过该算法计算得到的疾病关联矩阵的热力图表示。

3.4 基于 GRU 的多源健康感知数据动静态关系融合

GRU 可以学习时间序列中的趋势特征，包括长期依赖和短期依赖，且具有计算时间开销较小的优势，非常适用于大量长时间医学感测数据的特征提取与分析。医学感测数据依赖其时间趋势特征在体征测量值变化明显的急性病判断方面具有一定优势，但患者个体的年龄、性别等体质条件会对体征的整体数值产生影响，进而影响模型的判断。因此引入了个人体质记录数据，采用 GRU[24] 将其与医学感测数据进行融合，融合方法如图 3 所示。

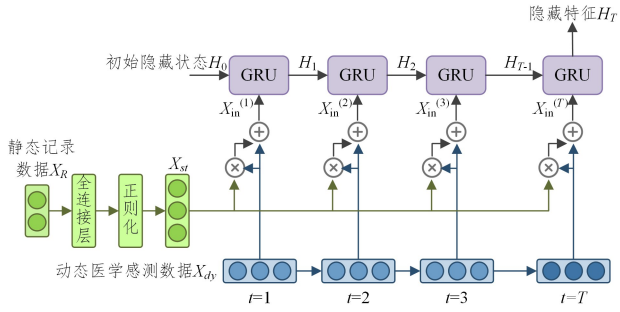


图 3 基于 GRU 的动静态关系融合

Fig. 3 Dynamic and static relationship fusion based on gated cyclic unit

对于记录数据，同样使用独热向量编码和标准化处理。在每个时间步 t，对记录数据与感测数据进行融合，首先通过全连接网络对记录数据进行维度变换，并进行正则化处理，获得静态数据权重，然后通过 Hadamard 乘积融合动静态数据，并将融合数据与动态数据进行拼接，即：

$$X_{st}=Norm(f_r(X_R)) \tag{14}$$

$$X_{in}^{(t)}=[X_{st}\circ X_{dy}^{(t)},X_{dy}^{(t)}],t=1,2,\cdots,T \tag{15}$$

其中，$X_R\in\mathbb{R}^{d_r}$ 为个人体质记录数据，$X_{dy}^{(t)}\in\mathbb{R}^{d_{in}}$ 为 t 时刻的动态时间序列数据，f_r 为全连接网络，$Norm()$ 为正则化操作，$\circ$ 表示 Hadamard 乘积。在这一步中，使用全连接层提取了个人体质记录数据的浅层特征并将其映射到与医学感测数据相同的维度，通过 Hadamard 乘积计算个人体质状况与对体征变化的影响，得到了 GRU 输入 $X_{in}\in\mathbb{R}^{T\times d_{in}}$。

图 4 给出了 GRU 网络结构。假设隐藏单元个数为 d_h，给定上一时间步隐藏状态 $H_{t-1}\in\mathbb{R}^{d_h}$，令 t 时刻的输入数据为 $X_{in}^{(t)}\in\mathbb{R}^{d_{in}}$，则重置门 $R_t\in\mathbb{R}^{d_h}$ 和更新门 $Z_t\in\mathbb{R}^{d_h}$ 的计算公式如下：

$$R_t=\sigma(X_{in}^{(t)}W_{xr}+H_{t-1}W_{hr}+b_r) \tag{16}$$

$$Z_t=\sigma(X_{in}^{(t)}W_{xz}+H_{t-1}W_{hz}+b_z) \tag{17}$$

其中，W_{xr}，W_{hr}，W_{xz}，W_{hz}，b_r，b_z 为可学习参数，σ 为 sigmoid 函数，将变量映射到(0,1)之间。然后通过重置门计算当前时间步的候选隐藏状态。时间步 t 的候选隐藏状态 $\tilde{H}_t\in\mathbb{R}^{d_h}$ 定义如下：

$$\tilde{H}_t=\tanh(X_{in}^{(t)}W_{xh}+(R_t*H_{t-1})W_{hh}+b_h) \tag{18}$$

其中，W_{xh}，W_{hh}，b_h 为可学习参数，tanh 激活函数将候选隐藏状态 $\tilde{H}_t$ 的值映射到(−1,1)中。之后使用更新门计算当前时间步 t 的隐藏状态 H_t：

$$H_t=Z_t*H_{t-1}+(1-Z_t)*\tilde{H}_t \tag{19}$$

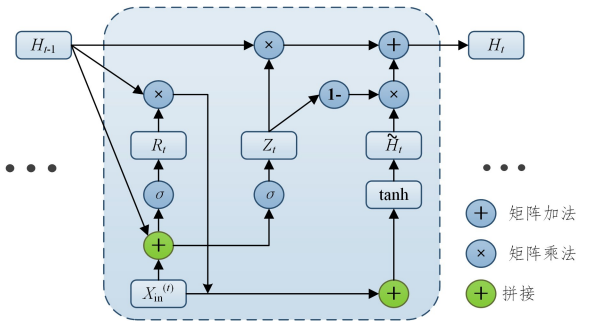


图 4 GRU 网络结构

Fig. 4 Network structure of gated cyclic unit

初始化隐藏状态 $\boldsymbol{H}_0\in\mathbb{R}^{d_h}$ 为零向量，经过门控循环单元 T 时间步的不断更新，可以获得医学感测数据和个人体质记录数据融合后的隐藏特征 H_T。在数据融合过程中，为了避免模型训练出现过拟合，采用 dropout 正则化方法随机让网络的一些节点失效，即：

$$O_T=Dropout(H_T) \tag{20}$$

3.5 疾病诊断算法描述

在 3.3 节计算得到了疾病关联矩阵 $\boldsymbol{I}\in\mathbb{R}^{K\times K}$；在 3.4 节获得了医学感测数据和个人体质记录数据融合后的 GRU 输出 $O_T\in\mathbb{R}^h$。由于两者的结构并不一致，需要对数据做进一步的处理和融合。

首先通过全连接网络将 GRU 输出 O_T 转换为 K 类疾病的患病概率 C：

$$C=f_c(W_cO_T+b_c) \tag{21}$$

其中，f_c 为全连接神经网络，W_c 和 b_c 为可学习参数，患病概率 $C\in\mathbb{R}^K$。

然后对患病概率 C 和疾病关联矩阵 $\boldsymbol{I}$ 做乘法运算，将疾病依赖关系融合到模型中，并使用 sigmoid 激活函数得到最终的多类别疾病诊断结果 $\hat{Y}$：

$$\hat{Y}=\sigma(C\times\boldsymbol{I}) \tag{22}$$

本模型采用的损失函数为平均二进制交叉熵损失 L：

$$L=-\frac{1}{K}\sum_{k=1}^{K}(y_k\log(\hat{y}_k)+(1-y_k)\log(1-\hat{y}_k)) \tag{23}$$

其中，y_k 为疾病 k 的真实患病情况，$\hat{y}_k$ 为疾病 k 的诊断结果。

算法 1 DSRF 模型训练过程

输入：医学感测数据 X_M；个人体质记录数据 X_R；患者患病情况数据 X_D

输出：25 类疾病诊断结果 Y

1. 医学感测数据预处理 $X_{dy}=[(x_1,c_1),(x_2,c_2),\cdots,(x_T,c_T)]$
2. 计算疾病关联矩阵 $\mathbf{I}_{jk}=P(k|j)=P(S_{jk})/P(S_j)$
3. 个人体质记录数据维度变换 $X_{st}=Norm(f_r(X_R))$
4. 对于每个时间步 t，t=1,2,…,T：
5. 　　动静态数据融合 $X_{in}^{(t)}=[X_{st}\circ X_{dy}^{(t)},X_{dy}^{(t)}]$
6. 　　计算 GRU 隐藏特征 H_t
7. 正则化 $O_T=Dropout(H_T)$
8. 计算各疾病患病概率 $C=f_c(W_cO_T+b_c)$
9. 疾病诊断结果 $\hat{Y}=\sigma(C\times\mathbf{I})$

10. 计算分类损失 L
11. 从第三步开始迭代训练模型至设定迭代次数或分类结果达到停止标准

4 实验与分析

本文将基于多源健康感知数据动静态关系融合的疾病诊断模型(DSRF)与基准分类模型进行对比,以评估模型的有效性和参数敏感性,以及模型的不同结构对疾病诊断性能的影响。

4.1 数据集与参数设置

本文使用美国大型医疗临床数据库 MIMIC-III(V1.4)[6]进行实验,从原始数据库中筛选出有效的医学感测数据、个人体质记录数据和患病情况数据,并按照 8:1:1 的比例进行训练集、验证集和测试集的划分。

模型选择 Adam 优化算法,截取时间序列长度为 480,设置训练批次大小为 64,学习率为 0.001,GRU 层使用 256 个隐藏神经单元,dropout 参数为 0.4,设置迭代次数为 50。

4.2 基准模型与评价指标

为了验证基于多源健康感知数据动静态关系融合的疾病诊断模型的性能,与以下基准模型进行实验对比。

LR[25]:使用 Lipton 等描述的方法进行特征提取。

LSTM[26]:标准长短期记忆网络。

Doctor AI[21]:使用 RNN 将医学记录数据用于预测后续就诊和药物类别诊断。

Med2Vec[27]:采用可扩展的两层神经网络来学习医学数据的低维向量表示。

Crist'obal[28]:使用 RNN 和前馈神经网络将序列数据和静态数据相结合。

SAnD[29]:采用掩码自注意力机制、位置编码和密集插值策略建模临床时间序列数据。

AdaCare[4]:学习患者生理特征的长短期变化,在多个时间尺度上描述患者的健康状况。

评估指标:本文采用 Micro AUROC, Macro AUROC 和 Weighted AUROC 3 种评估指标评估多类别疾病诊断情况。给定分类结果 $\hat{Y}$ 和真实标签 Y,定义函数 $Neg(y)$、$Pos(y)$ 分别表示样本 y 是否为正例、是否为负例,$Num(y_1, y_2)$ 表示样本 y_1 的分类概率是否大于 y_2。3 种评估指标的计算式如下:

$$Micro\ AUROC=\frac{\sum_{i=1}^{K}\sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{n=1}^{N}Pos(y_{nk})Neg(y_{ij})Num(\hat{y}_{nk},\hat{y}_{ij})}{\sum_{k=1}^{K}\sum_{n=1}^{N}Pos(y_{nk})\sum_{k=1}^{K}\sum_{n=1}^{N}Neg(y_{nk})} \tag{24}$$

$$Macro\ AUROC=\frac{1}{K}\sum_{k=1}^{K}\frac{\sum_{i=1}^{N}\sum_{n=1}^{N}Pos(y_{nk})Neg(y_{ik})Num(\hat{y}_{nk},\hat{y}_{ik})}{\sum_{n=1}^{N}Pos(y_{nk})\sum_{n=1}^{N}Neg(y_{nk})} \tag{25}$$

$$Wighted\ AUROC=\sum_{k=1}^{K}\frac{\sum_{i=1}^{N}\sum_{n=1}^{N}Pos(y_{nk})Neg(y_{ik})Num(\hat{y}_{nk},\hat{y}_{ik})w_k}{\sum_{n=1}^{N}Pos(y_{nk})\sum_{n=1}^{N}Neg(y_{nk})} \tag{26}$$

其中,K 是待分类疾病类别总数,N 是患者样本总数,w_k 为疾病 k 的患病人数占所有疾病患病人数总和的比例:

$$w_k=\frac{\sum_{n=1}^{N}P(y_{nk})}{\sum_{k=1}^{K}\sum_{n=1}^{N}P(y_{nk})} \tag{27}$$

4.3 对比实验结果分析

表 1 列出了不同模型基于 MIMIC-III 数据集的多类别疾病诊断任务的评估结果。从表中可以看出,本文模型 DSRF 与其他模型相比在所有评估标准中取得了最好的指标值。

表 1 不同模型下多类别疾病诊断任务结果比较

Table 1 Comparison of disease diagnosis tasks with different models

(单位:%)

评估指标	*Micro AUROC*	*Macro AUROC*	*Weighted AUROC*
LR	80.12	74.06	73.23
LSTM	82.10	77.04	75.69
DoctorAI	82.21	77.40	76.54
Med2Vec	82.24	77.51	76.62
Crist'obal	82.44	77.81	76.98
SAnD	81.64	76.58	75.39
AdaCare	82.59	77.76	76.72
Ours	**83.34**	**78.64**	**77.57**

对于医学感测数据中的时间序列数据,LR 模型采用统计学方法,在不同变量的给定时间序列的 7 个不同子序列上计算 6 个不同的样本统计特征来表示数据的时间特征,但这种手动提取特征的方法忽略了不同时间的先后关系。SAnD 模型采用了掩码自注意力机制对临床时间序列数据建模,并使用位置编码和密集插值策略来学习数据的时间特征,但该方法在提取长时间特征方面存在局限性。与以上两种模型相比,本模型使用的门控循环神经网络 GRU 在提取时间趋势特征和长短期依赖方面具有较大优势。

在多源健康感知数据融合方面,LSTM 和 AdaCare 模型仅使用了医学感测时序数据进行疾病诊断。虽然 AdaCare 模型中的扩张卷积层可以学习患者生理特征的长期和短期变化,并且采用了尺度自适应特征提取和重校准方法对临床特征之间的相关性进行建模,但该模型没有考虑到个人体质状况对患者生理特征的影响和多种疾病之间的相互影响。本节提出的医学感测数据和个人体质状况数据的融合使模型可以学习到两者之间的潜在关系,并且构建了疾病关联矩阵,充分利用了疾病间依赖关系这一特征,使多类别疾病诊断任务效果有了明显提升。

DoctorAI, Med2Vec 和 Crist'obal 等提出的模型利用循环神经网络、前馈神经网络及其组合的方法将医学感测数据和个人体质记录数据两种异构数据映射到相同的维度进行融合,但 3 种模型均对两类数据分别进行处理后融合,并未提取数据间关系。因此,本文在数据关系提取方面进行了改进,采用线性层提取个人体质记录数据的浅层特征并将其映射到与医学感测数据相同的维度,使用 Hadamard 积计算个人体质状况与对体征变化的影响,与感测数据原始特征共同作为门控循环单元的输入。该动静态关系融合方法不仅可以解决异构数据的融合问题,而且能够提取异构数据的关系特征,帮助模型根据不同患者的体质状况对其所患疾病进行差异性诊断。表 2 列出了急性、混合性、慢性疾病 3 类疾病的患病比例和在这 3 类疾病上的 Macro AUROC 和 Weighted AUROC 评估结果。从表 2 可以看出,本模型对 3 类疾病的分类效果

均优于最佳的基准模型，这表明所提出的数据融合模型对不同类型的疾病分类都有不错的效果。

在所有基准模型中，急性病的分类效果最好且模型差距小，这是因为急性病的时序趋势特征明显，更容易被学习到。在这种情况下，使用扩张卷积网络提取模型趋势特征的 AdaCare 模型明显优于其他模型，但其对疾病间相关性问题缺乏考虑，导致在急性病诊断性能上略低于本文 DSRF 模型。慢性病的趋势特征较弱，因此，将医学感测数据与个人体质数据进行融合处理的 DoctorAI，Med2Vec 和 Crist'obal 模型均取得了相对较好的分类效果，但以上 3 种基准模型的慢性病诊断性能仍差于本文模型，这证实了本文模型对医学感测动态数据与个人体质记录数据相关性进行挖掘的有效性。

表 2　疾病分类别评估结果比较

Table 2　Comparison of disease classification evaluation results

（单位：%）

疾病类型（占比）	急性病(38.90%)			慢性病(39.35%)			混合型疾病(21.75%)		
评估指标	*Micro AUROC*	*Macro AUROC*	*Weighted AUROC*	*Micro AUROC*	*Macro AUROC*	*Weighted AUROC*	*Micro AUROC*	*Macro AUROC*	*Weighted AUROC*
LR	80.73	74.81	74.41	79.48	73.24	72.16	80.19	73.97	73.03
LSTM	82.72	77.47	76.20	81.49	76.51	75.13	82.07	77.09	75.79
DoctorAI	82.86	77.89	76.95	81.56	76.98	76.12	82.22	77.35	76.56
Med2Vec	82.88	78.04	77.03	81.59	77.05	76.18	82.25	77.38	76.66
Crist'obal	82.97	78.65	77.74	81.84	77.13	76.27	82.56	77.54	76.91
SAnD	82.54	77.09	75.84	80.97	76.13	75.04	81.23	76.46	75.23
AdaCare	83.65	79.02	78.12	81.49	76.67	75.53	82.64	77.45	76.35
Ours	**84.05**	**79.14**	**78.07**	**82.63**	**78.12**	**77.05**	**83.35**	**78.66**	**77.62**

为了研究模型中医学感测数据和个人体质记录数据动静态关系融合、疾病关联矩阵融合两个模块的效果，本文用拼接方法代替式(14)、式(15)得到拼接融合模型($DSRF_{noFu}$)，屏蔽模型中的疾病关联矩阵，获得无疾病关系模型($DSRF_{noRe}$)，并将其结果与 DSRF 模型进行比较，如图 5 所示。

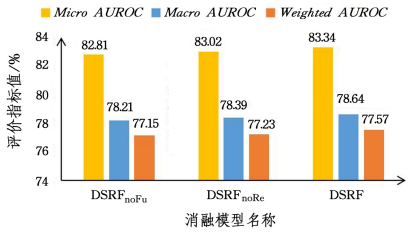


图 5　消融实验结果比较

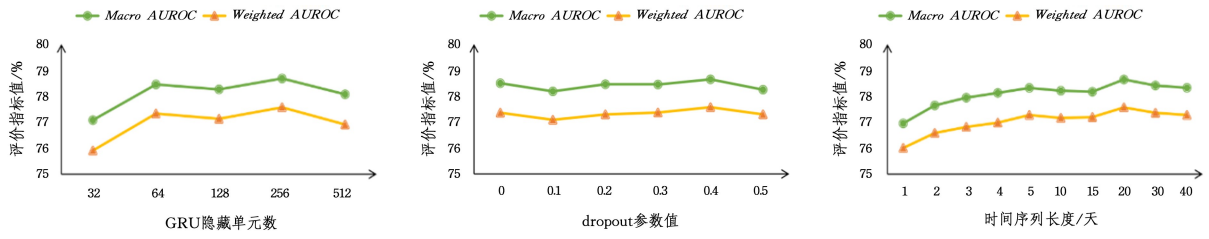
Fig. 5　Results comparison of ablation experiment

由图 5 可以看出，使用医学感测数据与个人体质记录数据融合、医学感测数据与疾病关系数据融合的模型性能评估结果均低于所提出的 DSRF 模型，这意味着动静态数据关系和疾病相关关系在多类别疾病诊断任务中发挥着重要作用，模型中的多源健康感知数据融合方法和疾病相关关系挖掘方法是有效的，可以更好地完成多类别疾病诊断任务。

4.4　参数选择与样例分析

本节对模型中的 GRU 隐藏单元数、dropout 正则化参数和数据的时间序列长度进行研究。首先，在基于 GRU 的多源健康感知数据融合过程中，GRU 中的隐藏单元可以用来提取和记忆输入数据中的时序特征，越多的隐藏单元 d_h 可以带来越复杂的计算能力，但更容易造成模型的过拟合。为了避免这一问题，本节所述模型采用 dropout 正则化方法（见式(20)）随机让网络的一些节点失效。因此，GRU 隐藏单元数和 dropout 参数是影响模型学习能力的两个重要参数。接下来，本节将通过实验结果来评估不同 GRU 隐藏单元数不同 dropout 参数值和不同时间序列长度对模型分类性能的影响。

图 6(a)和图 6(b)分别给出了不同 GRU 隐藏单元数和 dropout 参数对模型分类性能的影响，具体的评估指标包括 Macro AUROC 和 Weighted AUROC。其中，指标值越高，说明分类结果越接近真实值，意味着模型的性能越好；反之，指标值降低表示模型性能有所下降。从图中可以看出，当 GRU 网络的隐藏单元数大于等于 64 时，模型获得了较好的疾病诊断效果，Macro AUROC 和 Weighted AUROC 指标分别保持在 0.78 和 0.77 以上。当 GRU 网络的隐藏单元数为 256 时，模型达到最佳性能。因此，本节将模型中 GRU 隐藏单元参数设置为 256。对于 dropout 参数来说，当 dropout 参数从 0.1 向 0.4 变化过程中分类效果逐渐变好，在 0.4 时达到顶峰，之后逐渐下降。参数值为 0.4 时两种指标值均大于不使用 dropout 正则化方法（即 $dropout=0$）的指标值，这说明该方法对于改善模型过拟合是有效的。



(a)GRU 隐藏单元数对模型性能的影响　(b)dropout 参数对模型性能的影响　(c)时间序列长度对模型性能的影响

图 6　GRU 隐藏单元数、dropout 参数和时间序列长度对模型性能的影响

Fig. 6　Effect of hidden units number, dropout parameters and sequence length on model performance

为了进一步评估所提出模型的性能，本文将数据处理成不同长度的时间序列，以评估模型的鲁棒性，图 6(c)给出了

DSRF 模型在不同长度时间序列上的分类性能评估曲线。从图中可以看出，模型获得最佳性能所需要的时间序列长度为 20 d 即 480 h。从第 5 天开始，模型性能上升趋势趋于平缓，Macro AUROC 和 Weighted AUROC 指标分别保持在 0.78 和 0.77 以上，这说明所提方法对较短时间序列的医疗数据依然有效，具有较好的泛化性能。

另外，本文还选取了部分患者，将 DSRF 模型和基准模型 GRU 进行比较，来验证疾病关联矩阵模块的有效性。图 7、图 8 分别给出了一名患者在不同模型下的疾病诊断概率与真实患病情况，矩形颜色越深表示模型认为患者患有该疾病的概率越大，反之颜色越浅。每一列表示一种疾病，3 行从上到下依次为 GRU 模型诊断概率、DSRF 模型诊断概率和代表患者是否患病的二进制值。

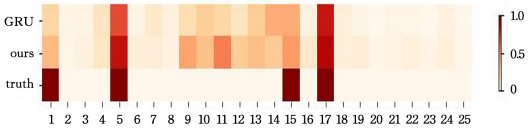


图 7 患者 1 的疾病诊断概率与真实患病情况

Fig. 7 Patient 1's disease diagnosis probability and real disease condition

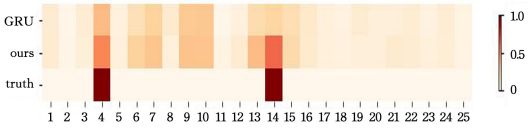


图 8 患者 2 的疾病诊断概率与真实患病情况

Fig. 8 Patient 2's disease diagnosis probability and real disease condition

从图 7 可以看出，与 GRU 相比，DSRF 模型判断这名患者有较大概率患有慢性肾病(5)，因为其患有高血压伴并发症(17)的概率较高，这与前文提到的高血压伴并发症和慢性肾病存在依赖关系相符。关联矩阵根据两种疾病之间的关系对患者的疾病诊断结果进行了修正，成功地诊断出慢性肾病。从图 8 可以看出，有了疾病关联矩阵的作用，原发性高血压(14)与高血压伴并发症(17)的互斥关系使得两者的患病概率呈现一高一低的特点，DSRF 模型对原发性高血压进行了正确的诊断。

结束语 本文提出了一种基于多源健康感知数据动静态关系融合的疾病诊断模型(DSRF)。该模型使用数据动静态关系融合算法提取异构数据的关系特征，构建疾病关联矩阵来提取疾病间的相关关系特征，并利用门控循环单元网络先后融合了动态医学感测数据、静态个人体质记录数据和疾病关系数据。最后在 MIMIC III 真实公开数据集上进行评估。实验结果表明，DSRF 模型优于最好的基准模型，可以很好地对多种类型疾病进行诊断。

在未来工作中，可以对疾病关系的提取和融合工作做进一步的改进，尝试对不同类型的疾病进行更有针对性的诊断和预测，并将其应用于更多的相关领域中。

参 考 文 献

[1] KOREN A, PRASAD R. IoT Health Data in Electronic Health Records(EHR): Security and Privacy Issues in Era of 6G[J]. Journal of ICT Standardization, 2022, 10(1): 63-84.

[2] MAHAJAN P, RANA D. Investigating Clinical Named Entity Recognition Approaches for Information Extraction from EMR [M]// Tracking and Preventing Diseases with Artificial Intelligence. Springer, Cham, 2022: 153-175.

[3] MOHAMMADI R, JAIN S, AGBOOLA S, et al. Learning to identify patients at risk of uncontrolled hypertension using electronic health records data[J]. AMIA Summits on Translational Science Proceedings, 2019, 2019: 533-542.

[4] MA L, GAO J, WANG Y, et al. AdaCare: Explainable Clinical Health Status Representation Learning via Scale-Adaptive Feature Extraction and Recalibration [C] // Proceedings of the AAAI Conference on Artificial Intelligence, 2020: 825-832.

[5] AYON S I, ISLAM M M. Diabetes Prediction: A Deep Learning Approach[J]. International Journal of Information Engineering and Electronic Business, 2019, 11(2): 21-27.

[6] 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.

[7] WONG M S, WELLS M, PARRINELLA K, et al. EHR phenotyping by Natural Language Processing improves detection of patients at risk for preeclampsia[J]. American Journal of Obstetrics & Gynecology, 2022, 226(1): S65-S66.

[8] LIAO B, JIA X, ZHANG T, et al. DHDIP: An Interpretable Model for Hypertension and Hyperlipidemia Prediction Based on EMR Data[J/OL]. SSRN. http://dx. doi. org/10. 2139/ssrn. 4022954.

[9] GUTIERREZ G. Artificial intelligence in the intensive care unit [J]. Critial Care, 2020, 24(1): 1-9.

[10] PIRRACCHIO R. Mortality prediction in the ICU based on MIMIC-II results from the super ICU learner algorithm(SICULA) project[J/OL]. Secondary Analysis of Electronic Health Records, 2016, 2016: 295-313. https://doi. org/10. 1007/978-3-3-19-43742-2_20.

[11] SANJAY P, CHUIZHENG M, ZHENGPING C, et al. Benchmarking deep learning models on large healthcare datasets[J]. Journal of Biomedical Informatics, 2018, 83: 112-134.

[12] HARUTYUNYAN H, KHACHATRIAN H, KALE D C, et al. Multitask learning and benchmarking with clinical time series data[J]. Scientific Data, 2019, 6(1): 1-18.

[13] LEE W, SHI Y, SUN H, et al. MSIPA: Multi-Scale Interval Pattern-Aware Network for ICU Transfer Prediction [J]. ACM Transactions on Knowledge Discovery from Data (TKDD), 2021, 16(1): 1-17.

[14] CHOI E, XU Z, LI Y, 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.

[15] XU Z, SO D R, DAI A M. MUFASA: Multimodal Fusion Architecture Search for Electronic Health Records[C] // Proceedings of the AAAI Conference on Artificial Intelligence. 2021: 10532-10540.

[16] AGARWAL V, PODCHIYSKA T, BANDA J M, et al. Learning statistical models of phenotypes using noisy labeled training data [J]. Journal of the American Medical Informatics Association, 2016, 23(6): 1166-1173.

[17] HALPERN Y,HORNG S,CHOI Y,et al. Electronic medical record phenotyping using the anchor and learn framework[J]. Journal of the American Medical Informatics Association,2016,23(4):731-740.

[18] MARLIN B M,KALE D C,KHEMANI R G,et al. Unsupervised pattern discovery in electronic health care data using probabilistic clustering models[C]// Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium. 2012:389-398.

[19] HO J C,GHOSH J,SUN J. Marble:high-throughput phenotyping from electronic health records via sparse nonnegative tensor factorization[C]// Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2014:115-124.

[20] CHE Z,KALE D,LI W,et al. Deep computational phenotyping[C]// Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2015:507-516.

[21] CHOI E,BAHADORI M T,SCHUETZ A,et al. Doctor ai:Predicting clinical events via recurrent neural networks[C]// Machine Learning for Healthcare Conference. PMLR,2016:301-318.

[22] RAZAVIAN N,MARCUS J,SONTAG D. Multi-task prediction of disease onsets from longitudinal laboratory tests[C]// Machine Learning for Healthcare Conference. PMLR,2016:73-100.

[23] MA L,ZHANG C,WANG Y,et al. Concare:Personalized clinical feature embedding via capturing the healthcare context[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2020:833-840.

[24] CHUNG J,GULCEHRE C,CHO K H,et al. Empirical evaluation of gated recurrent neural networks on sequence modeling[J]. arXiv:1412.3555,2014.

[25] LIPTON Z C,KALE D C,ELKAN C,et al. Learning to diagnose with LSTM recurrent neural networks[J]. arXiv:1511.03677,2015.

[26] HOCHREITER S,SCHMIDHUBER J. Long short-term memory[J]. Neural Computation,1997,9(8):1735-1780.

[27] CHOI E,BAHADORI M T,SEARLES E,et al. Multi-layer representation learning for medical concepts[C]// proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016:1495-1504.

[28] ESTEBAN C,STAECK O,BAIER S,et al. Predicting clinical events by combining static and dynamic information using recurrent neural networks[C]// 2016 IEEE International Conference on Healthcare Informatics(ICHI). IEEE,2016:93-101.

[29] SONG H,RAJAN D,THIAGARAJAN J J,et al. Attend and diagnose:Clinical time series analysis using attention models [C]// Thirty-second AAAI Conference on Artificial Intelligence. 2018.

HUO Tian-yuan, born in 1997, postgraduate, is a member of China Computer Federation. Her main research interests include machine learning and data mining.

GU Jing-jing, born in 1986, Ph.D, professor, is a member of China Computer Federation. Her main research interests include mobile computing and data mining.