DOI:10.3969/j.issn.1005-202X.2019.09.021

医学人工智能

基于深度学习的心血管疾病风险预测模型

安莹[1],黄能军[2],杨荣[3],陈先来[1]
1.中南大学信息安全与大数据学院,湖南 长沙 410083;2.中南大学计算机学院,湖南 长沙 410083;3.中南大学湘雅医院,湖南 长沙 410078

【摘要】心血管疾病的准确预测对其预防工作有着重大的意义,本文提出一种基于电子病历数据挖掘的模型研究心血管疾病的风险预测。该模型利用循环神经网络等技术对患者的历史电子病历数据进行表征学习,不仅能有效捕获电子病历数据中的时序特征,而且其特征工程无需人工干预。此外,在循环神经网络上嵌入的关注机制从每个患者的数据学到了一个上下文向量,该向量能有效增强深度模型的拟合能力和可解释性。为了进一步提高心血管疾病风险预测的准确性,该模型融合了多种类型的临床数据,包括诊断编码序列、实验室数据以及人口学统计数据。该模型利用多个子模块进行表征学习,不仅能充分考虑到数据之间的差异性,还能考虑到它们之间潜在的关联性,最终提高心血管疾病风险预测的性能。实验结果表明,在心血管疾病风险预测的性能方面,该模型相比最新的几种方法具有较高的召回率、F1值和AUC值,其分别可达0.814 9、0.737 8和0.837 5。

【关键词】心血管疾病;风险预测;电子病历;深度学习
【中图分类号】R319;TP391.4 【文献标志码】A 【文章编号】1005-202X(2019)09-1103-10

Deep learning-based model for risk prediction of cardiovascular diseases

AN Ying[1], HUANG Nengjun[2], YANG Rong[3], CHEN Xianlai[1]
1. Information Security and Big Data Research Institute, Central South University, Changsha 410083, China; 2. School of Computer Science and Engineering, Central South University, Changsha 410083, China; 3. Xiangya Hospital, Central South University, Changsha 410078, China

Abstract: The accurate prediction of cardiovascular diseases (CVD) is of great significance for the prevention of CVD. Therefore, a novel model based on electronic medical records (EHR) and data mining is proposed to investigate the risk prediction of CVD. Recurrent neural network is adopted for the representation learning of EHR, which can effectively capture temporal information hidden in EHR and achieve feature engineering without any manual intervention. Meanwhile, a context vector which is obtained *via* attention mechanism embed in recurrent neural network model can improve the fitting performance as well as interpretability of the risk prediction model. To further improve the accuracy of the risk prediction of CVD, the model effectively combines various kinds of clinical data, including diagnostic coding sequence, laboratory data and demographic statistics. The model utilizes several modules for representation learning, which can take full consideration of not only the difference but also the correlation among these clinical data, thus improving the performance in the risk prediction of CVD. Experimental results show that the proposed model outperforms the latest methods in the risk prediction of CVD. The recall rate, F1-score and AUC of the proposed model can reach 0.814 9, 0.737 8 and 0.837 5, respectively.

Keywords: cardiovascular disease; risk prediction; electronic medical record; deep learning

前 言

心血管疾病是一种严重威胁人类健康的常见慢性疾病,在城乡居民总死亡原因中高居首位[1]。准确预测心血管疾病的发病风险对防范心血管疾病的发生有着重大意义。尽管临床上血管造影术可以准确地诊断出心血管疾病,但是血管造影术不仅比较昂贵而且对身体具有创伤性。此外,临床上也常用心电图和一些评分指数来预估心血管的风险,但这些方法需要医生或者执业人员具备丰富的理论知识和

【收稿日期】2019-04-11
【基金项目】国家重点研发计划(2016YFC0901705);湖南省自然科学基金(2018JJ2534);湖南省研究生创新项目(1053320170077);中南大学中央高校基本科研业务费专项(2017zzts721)
【作者简介】安莹,博士,副教授,主要研究方向:大数据分析、机器学习及其应用,E-mail: anying@csu.edu.cn
【通信作者】杨荣,主管护师,现从事临床护理工作,E-mail: cxlyr05-76@163.com

实践经验。近年来，一些研究人员提出利用机器学习等算法对患者电子病历中的风险因素进行建模，从而实现心血管疾病的风险预测。由于这类方法不仅对身体没有创伤性，而且是相对廉价的，因此它已逐渐成为当下一个具有重要现实意义的挑战性课题。

在基于电子病历的心血管疾病风险预测研究中，最大的挑战是如何通过有效的表征学习方法来实现患者画像的准确描绘。电子病历通常包含丰富的患者就诊信息，比如诊断、医嘱、检查检验、生命体征、人口学等数据。传统患者画像的表征方法通常需要大量的人工干预，它们的性能往往受限于研究人员的经验以及特定的电子病历系统，导致其可扩展性和泛化性较差。最近几年，受自动特征学习相关研究成果的鼓舞，很多研究人员成功利用稀疏编码的方式实现了特征表达，比如独热编码（One-Hot）[2]和词袋模型（Bag of Words, BoW）[3]。然而，这些稀疏编码的方式通常无法捕获特征之间的语义性以及电子病历数据中的时序性。近些年，随着深度学习在理论上的突破以及其在生物、金融等众多领域的成功应用，很多研究人员也试图利用深度学习来处理电子病历数据的表征学习。Nguyen等[4]提出将患者的电子病历数据（诊断、药物治疗以及手术记录）表示成一串按时间先后顺序排列的序列，并且利用卷积神经网络（Convolution Neural Network, CNN）对其进行患者的表征提取。但是在时序学习（Temporal Learning）相关任务中，CNN相对来说只能捕获局部特征信息，并且需要假设一份电子病历中的数据是严格按时间顺序排列的。与之相比，基于循环神经网络（Recurrent Neural Network, RNN）的相关算法，比如长短期记忆神经网络（Long Short Term Memory, LSTM），通过不同“门限”来捕捉有用的信息而舍弃没用的信息，从而可以更好地处理带时序性的电子病历数据。Ma等[5]利用双向循环神经网络（Bidirectional Recurrent Neural Network, Bi-RNN）进行电子病历的表征学习，并利用多种关注机制（Attention Mechanism）方法提升模型的表征学习能力和可解释性。尽管该方法已经能有效提升风险预测模型的性能，但是它忽略了电子病历中各数据之间的差异性。相对地，Kim等[6]提出利用相互独立的模块对不同种类的数据（诊断和药物治疗）分别进行表征学习，来提高风险预测的准确性。但实际上，每一种药物治疗的方式在临床上都有与之对应的某一种或多种疾病。尽管该模型实现了有效的风险预测，但是它忽略了疾病和药物治疗两者之间的关联性，因此，预测性能受到了一定的影响。

为了解决以上的问题，本研究提出一个基于RNN和关注机制的心血管风险预测模型（Risk Prediction Model for Cardiovascular, RPMC）。RPMC可以自动从高维、异质、时序的电子病历数据中抽取高质量的表征，用来准确地实现心血管疾病的风险预测。由于关注机制和LSTM的引入，模型不仅能有效增强模型对时序数据的学习能力，还具备一定的可解释性。此外，考虑到不同数据之间的差异性和关联性，模型中不仅存在多个独立的模块负责不同数据的表征学习，还存在一个模块负责融合后数据的特征提取。最后，RPMC结合各个模块学到的表征实现心血管疾病的风险预测。本研究主要的贡献点可以归纳为以下3点：（1）提出一个端对端、易操作、无需医务人员辅助、鲁棒的心血管疾病风险预测模型RPMC；（2）将RNN和关注机制的结合，从而使得RPMC不仅能自动而准确地从高维、异质、时序的电子病历数据抽取潜在的表征，同时还具备良好的可解释性；（3）有效地融合多种不同质的电子病历数据，使用多个子模块进行表征学习，从而使得RPMC不仅能充分考虑到数据之间的差异性，还能考虑到他们之间潜在的关联性，最终提高心血管疾病风险预测的性能。

1 相关研究

1.1 风险预测

在医学领域中，风险预测是一个具有前瞻性和重大现实意义的研究任务。临床上常见的风险预测任务主要有：疾病发病风险预测[4,6]、死亡率预测[7]、再入院风险预测[8]等。在早期的心血管疾病风险预测模型中，很多研究员通过利用队列研究的方法来跟踪患者的状况，从而实现风险预测。Everett等[9]对1 821位心血管疾病患者进行队列研究，结果表明端前脑钠素能有效提高临床上心血管疾病的风险预测能力。此外，Welsh等[10]也利用队列研究的方式发现了更多有助于心脑血管疾病风险预测的医学指标。这些基于队列研究的方法能实现较准确的风险预测，同时还具备一定的医学参考价值和可解释性。但是，这类方法通常需要耗费大量的人力、物力和时间。它们通常基于某一权威的评分标准来进行预测，导致其性能在很大程度上取决于研究员的医学背景和经验。随着医院信息化程度的不断提高，来自医院信息系统的电子病历数据因其提供了极为丰富、完整的患者医疗记录而受到研究人员的关注。因此近年来出现大量基于电子病历的心血管疾病风

险预测模型。Huang等[11]基于患者电子病历数据中的特征，利用回归的方法进行特征学习并能有效提高心血管疾病的风险预测。Jiang等[12]利用电子病历数据中的特征，构建了一个高效的再入院风险预测模型。这些方法的提出不仅有效提高了现有、海量的电子病历数据的利用率，还在相关任务上取得显著的效果。但是，由于电子病历数据包含的特征信息种类繁多，维度庞大，并具有一定的时序性，所以如何对电子病历数据进行表征学习成为这类风险预测任务的主要挑战。

在很多现有的心血管疾病风险预测模型中，特征工程通常需要大量人工干预。比如，Pike等[13]根据Framingham风险评分（Framingham Risk Score, FRS）等标准，从电子病历中抽取出相关的特征，并比较各评分标准的风险预测能力。Kennedy等[14]在FRS的基础上引入了额外的电子病历特征，并实证其具备更好的心血管疾病风险预测能力。这类方法通常都是根据相关的评分标准或权威的文献资料，针对性地从电子病历数据中抽取相关特征。它最大的不足是往往受限于研究人员相关的专业背景和实际经验，同时特征的抽取过程掺杂大量的人力、物力，并不能完全高效地利用海量的电子病历数据。最近几年，很多研究员提出不同的基于机器学习相关方法的心血管疾病风险预测模型[11,15]。这类方法不仅能自动地学习出重要的特征信息，而且还能提高电子病历数据特征的利用率，从而实现高效的风险预测。但是，它们通常无法捕获电子病历数据中的时序信息。为了解决这些问题，很多研究人员利用深度学习来进行电子病历的特征学习工作，并取得了巨大成功。Nguyen等[4]成功利用CNN捕获电子病历数据中的特征信息（包括时序信息），并准确地实现了再入院的风险预测。此外，Ma等[5]利用RNN以及多种关注机制进行电子病历数据的时序特征提取，不仅有效提高模型的准确度，还增强模型的可解释性。Kim等[6]针对不同种类的电子病历数据，分别利用多个独立的RNN模块进行特征学习，并有效提高风险预测的准确性。尽管这些方法大大提高风险预测的准确性，但是它们并没有充分考虑电子病历数据中的多样性与关联性。

1.2 深度学习

近几年，深度学习在理论和应用上都有惊人的突破。深度学习通过组合低层的数据特征形成更加抽象的高层特征，从而发现数据中潜在的、难以被人发现的分布式特征表示[16]。目前，有两种常见的深度学习算法被广泛应用于电子病历和影像数据的特征学习，即CNN和RNN。CNN是一类具有深度结构的前馈神经网络，它通过卷积层和池化层的相关计算完成特征学习，具备平移不变性，因而在图像处理方面有着先天的优势，比如医学图像分割[17]、图片分类[18]等。但是，CNN只能捕获局部的信息，在处理长时间依赖的时序学习任务上略有不足。相比之下，RNN是一类具有记忆功能的深度神经网络，能很好地捕捉数据中的时序信息。因而，RNN被广泛应用于自然语言处理等时序学习任务中[19]。为了增强单向RNN的学习能力，Bi-RNN通过同时从两个方向学习数据的时间依赖信息，从而更全面地捕获数据中上下文的信息[20]。

此外，为了增强模型的可解释性，关注机制被广泛应用于深度学习模型中。关注机制通过计算出一个上下文向量，来捕获序列数据中更多的潜在信息。它不仅能有效提高模型的学习能力，而且还能增强模型的可解释性，尤其是在时序学习[5]、机器翻译[19]等任务中，基于关注机制的深度学习模型能明显优于不带关注机制的模型。

综合已有心血管疾病风险预测模型的优势和不足，本研究提出的模型RPMC利用双向长短期记忆神经网络（Bidirectional Long Short Term Memory, Bi-LSTM）以及关注机制等方法负责电子病历数据的表征学习。考虑到电子数据中的多样性与关联性，RPMC不仅分别利用多个独立的模块来负责不同性质数据的表征学习，同时也单独提供一个独立的模块负责融合后数据的表征学习。从而，RPMC能高效、全面地捕获电子病历数据的特征信息，实现更准确的疾病风险预测。

2 风险预测模型

2.1 数据描述

本研究所使用的实验数据来源于中南大学湘雅医学大数据平台建设项目组整理而成的湘雅医学数据集[21]。目前，该数据集涵盖湘雅3家附属医院近10年的电子病历数据。在湘雅数据集中，每一个疾病编码都遵循第10版国际疾病分类（ICD10）的标准，每一个实验室指标都遵循湘雅医院的规则，并且有特定的正常值参考范围。

本研究所使用的湘雅子数据集共包含322 900位患者的电子病历数据，其中24 615位是心血管疾病的患者。RPMC旨在利用患者的历史数据预测其在接下来一年中患有心血管疾病的风险。因此，每位患者的历史数据（不包含心血管疾病编码）构成观测窗口，而接下来一年的数据构成预测窗口。RPMC

从观测窗口中的电子病历数据捕获特征，并利用预测窗口生成分类标签（二分类，1表示高风险，而0表示非高风险）。

在该数据集上，患者每次医院就诊以7个工作日为单位进行聚合，即将同一患者间隔时间小于一周的不同就诊记录视为同一次医疗就诊。为了保证样本数据有足够的电子病历信息，少于6次医疗就诊的患者被排除在外，从而使得观测窗口中至少包含5次医疗就诊，而预测窗口至少包含1次。为了确认患者在观测窗口之后1年中心血管疾病的患病情况，如果患者有过心血管疾病诊断历史，并且在第一次被诊断有心血管疾病之前至少包含5次医疗就诊记录，同时第一次被诊断为心血管疾病的就诊时间距离上一次就诊时间≤1年，则该患者被标记为高风险样本；如果患者在观测窗口中最后一次医疗就诊后至少1年未诊断出心血管疾病，则被标记为非高风险患者。

除了诊断编码序列数据，RPMC还利用实验室指标数据。根据数据统计结果，出现频次少于100次，以及缺失率高于90%的实验室指标均被剔除在外。此外，为了增强心血管疾病风险预测的准确性，RPMC还结合部分人口学数据，包括年龄、性别、患者类型，就诊次数和手术史。

经过上述的数据筛选过程，最终得到的实验数据集总共包含146 296位患者，其中20 450位属于心血管疾病高风险患者。具体的统计信息如表1所示。

2.2 数据表示

表1 最终数据集的简单描述

Tab.1 A brief description of the final data set

统计指标	统计数量
患者人数	146 296
就诊次数	2 286 267
患者平均年龄/岁	45.07
患者平均就诊次数	15.63
诊断编码数据	817
实验室指标数	687

为方便心血管疾病风险预测模型的描述，数据集中医学编码的集合（包括诊断编码、实验室指标）被表示成 $D=\{d_1,d_2,\cdots,d_M\}$，其中 M 是编码的总数量，任意一个元素 d_j 表示一个医学编码。令 $P=\{p_1,p_2,\cdots,p_N\}$ 表示数据集中的患者集合，其中 N 为患者总数，任意一个元素 p_n 表示一个患者。对于任意患者 p_n，其电子病历数据可以被表示成一个医疗就诊序列 $\left\langle V_1^{p_n},V_2^{p_n},\cdots,V_{T(n)}^{p_n}\right\rangle$，其中 $T(n)$ 表示第 n 个患者的总就诊次数，$V_i^{p_n}$ 表示患者 p_n 的第 i 次就诊记录，是由一个或多个医学编码组成的无序集合。为了将每次就诊记录 V_i 转化成深度模型RPMC输入数据的格式，V_i 被表示成一个一维向量 $\boldsymbol{x}_i$，其中每个维度代表唯一的一种医学编码 d_j。对于二元医学变量，如诊断编码等，只有两种取值（如果 V_i 包含 d_j 则 $\boldsymbol{x}_i$ 中相应位置为1，否则为0）。此外，对于一些有多种取值的医学编码，比如具有连续型取值范围的实验指标数据，则采用如下的赋值策略：如果实验指标的数值在给定的正常值参考范围之内，则 $\boldsymbol{x}_i$ 中相应位置为1；如果实验指标的数值不在给定的正常值参考范围之内，则 $\boldsymbol{x}_i$ 中相应位置为2；否则，$\boldsymbol{x}_i$ 中相应位置为0。

如图1所示，每个患者的电子病历数据都能表示成一条序列。序列中，每个片段代表一次医疗就诊记录，即 V_i，包含一个或多个诊断编码和实验室指标。很显然，图1给出的是一个心血管疾病高风险患者的案例，因为在预测窗口中，该患者出现有心血管疾病的ICD10编码（心绞痛，I20）。假设HIS系统中总共只有9种不同编码（不包括心血管疾病相关编码）：I10、E78、H30、K81、WBC、PDW、FBG、HDL和BP，其中前4个为疾病编码，后5个为实验室指标，那么，每次就诊记录 V_i 都可以被表示成一个向量，其维度为9。比如，在片段1中患者被诊断为I10和E78，同时实验室指标WBC处于正常取值范围内，PDW的取值偏离正常范围，那么该片段 V_1 可以被表示成一个9维的稀疏向量 $\boldsymbol{x}_1=[1,1,0,0,1,2,0,0,0]$。此外，$\boldsymbol{x}_1$ 也可以被拆分成两部分：诊断编码向量 $\boldsymbol{x}_1'=[1,1,0,0]$ 和实验室指标向量 $\boldsymbol{x}_1''=[1,2,0,0,0]$，以便RPMC分别对两者单独进行训练。

在人口学数据中，每一个特征使用One-Hot的方式组织。如图2所示，年龄被拆分成7个阶段（“0-18”、“18-30”、“30-45”、“45-60”、“60-75”和“75+”），性别包括两个特定的值（男和女），患者类型包含3种不同的类别（门诊、急诊和住院），就诊次数被离散成6个片段（“6-12”、“12-18”、“18-24”、“24-30”、“30-36”和“36+”），手术史包括两种状态（“S”和“NS”，分别代表有、无手术史）。需要注意的是，前3个人口学特征（年龄、性别、患者类型）都有一个额外的维度（Unknown，“UK”），表示数据缺失的情况。

2.3 模型架构

如图3所示，RPMC包含4个输入模块，即诊断编码序列、诊断编码+实验室指标序列、实验室指标序

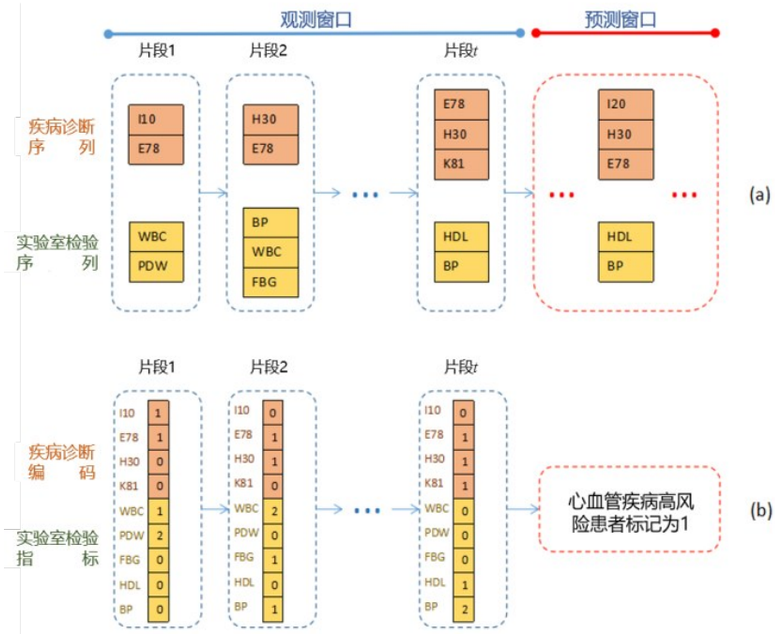

图1 患者序列数据生成示意图

Fig.1 Diagram of sequence data generating process

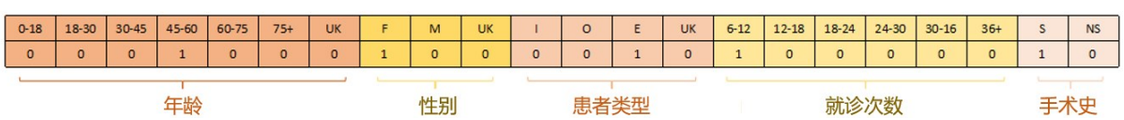

图2 人口学数据向量化

Fig.2 Vectorization of demographic data

列和人口学数据。首先,4种数据分别通过嵌入层的相关技术生成4个嵌入向量 $\boldsymbol{v}^1$、$\boldsymbol{v}^2$、$\boldsymbol{v}^3$ 和 $\boldsymbol{v}^4$;然后,分别利用4个基于关注机制的双向长短期记忆神经网络模块(Attention-based Bi-LSTM, A-LSTM)负责相应嵌入向量的表征学习,并得到相应的4个表征向量 $\boldsymbol{h}^1$、$\boldsymbol{h}^2$、$\boldsymbol{h}^3$ 和 $\boldsymbol{h}^4$;最后,拼接4个表征向量,并用于softmax层进行预测。

在嵌入层中,除了人口学数据模块使用的是词袋模型之外,其他模块所使用的方法都是Med2Vec[21]。Med2Vec利用线性整流单元(Rectified Linear Unit, ReLU)来获取患者就诊记录的嵌入向量。ReLU是神经网络中常见的激活函数,计算公式如下:

$$\boldsymbol{v}_t = \mathrm{ReLU}(\boldsymbol{W}_v \boldsymbol{x}_t + \boldsymbol{b}_c) \tag{1}$$

其中,$\boldsymbol{W}_v \in R^{m \times M}$ 是一个用来衡量每个医学变量重要程度的权重矩阵,m 是嵌入向量 $\boldsymbol{v}_t$ 的大小。

2.4 Bi-RNN

RNN是一类用于处理序列数据的神经网络,它能高效地从序列数据中捕获潜在、深层的语义信息。但是,单向的RNN只能从一个方向捕获序列信息,比如前向循环神经网络(forward RNN)在推断当前节点的状态时,只考虑节点之前的信息,而忽略节点之后的信息。因此,为了保证模型能充分考虑节点的上下文信息,RPMC利用Bi-RNN对嵌入向量进行表征学习。

如图4所示,Bi-RNN由一个前向RNN和一个后向RNN组成,能充分利用当前状态之前和以后的特征信息。前向RNN负责从序列的前端向后端的表征学习任务,而后向RNN正好相反。最后,Bi-RNN将两个单向的RNN所学到的隐藏层特征信息进行融合,得到隐藏层的最终状态。对于两个单向RNN输出的融合方式,常见的有拼接(concatenate)、element-wise等操作。在RPMC中,采用的方法是拼接,因为它通常能取得较好的效果。此外,为了克服梯度消失的问题,RPMC实际采用的是Bi-LSTM。

2.5 关注机制

为了增强Bi-LSTM的表征学习能力,RPMC利用关注机制(Attention Mechanism)来帮助模型捕获更多的上下文信息。如果只单纯利用Bi-LSTM来捕获心血管风险中的时序特征信息 $\langle \boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_t \rangle$,将有

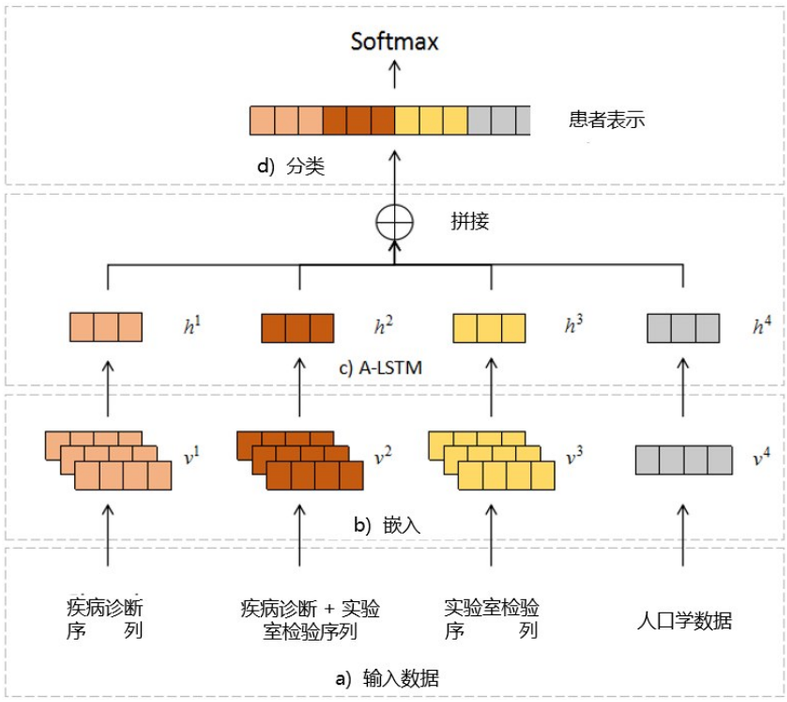

图3 心血管疾病风险预测模型概览图

Fig.3 Framework of risk prediction model of cardiovascular diseases

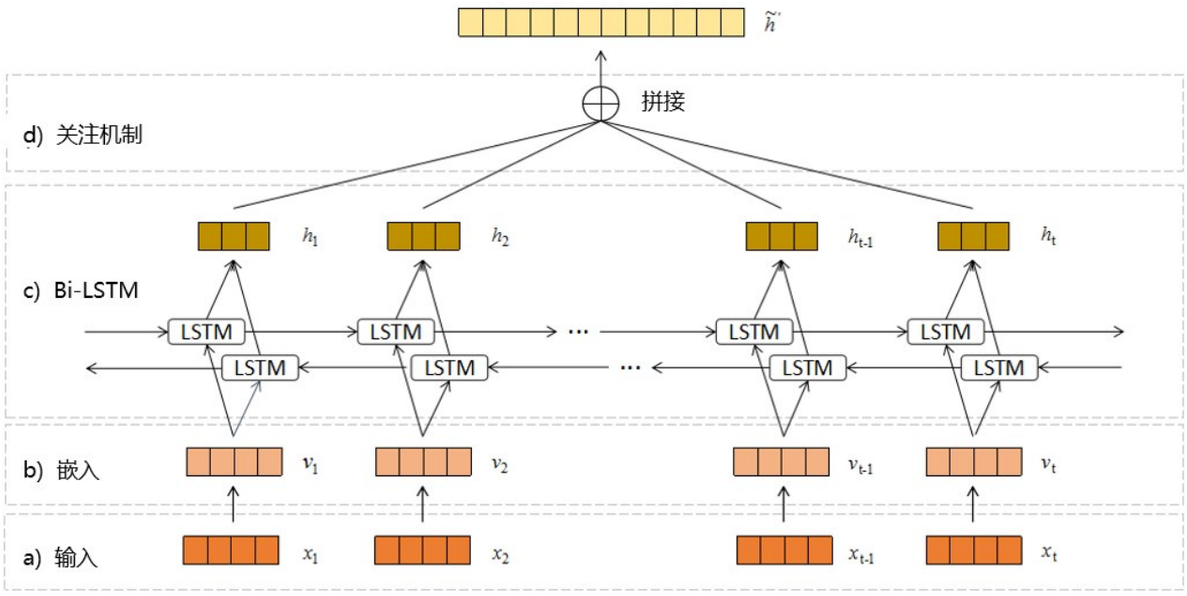

图4 A-LSTM的网络结构图

Fig.4 Network of Bi-LSTM based on attention mechanism (A-LSTM)

可能忽略掉输入的序列数据$\langle \boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_t \rangle$中的一些重要信息。然而,RPMC利用关注机制能学到一个额外的上下文向量$\boldsymbol{c}_t$,这个上下文向量不仅能有效增强模型的预测能力,还能提高模型的可解释性。计算$\boldsymbol{c}_t$方法如式(2)所示:

$$\boldsymbol{c}_t = \sum_{i=1}^{t-1} \boldsymbol{\alpha}_{ti} \boldsymbol{h}_i \tag{2}$$

其中,$\boldsymbol{h}_i$表示第i个隐藏层节点的状态,$\boldsymbol{\alpha}_{ti}$是一个用来衡量当前状态各元素权重的向量,其计算方法如式(3)和式(4)所示:

$$\boldsymbol{\alpha}_{ti} = \boldsymbol{w}_\alpha^{\mathrm{T}} \boldsymbol{h}_i + \boldsymbol{b}_\alpha \tag{3}$$

$$\boldsymbol{\alpha}_t = \text{softmax}([\boldsymbol{\alpha}_{t1}, \boldsymbol{\alpha}_{t2}, \cdots, \boldsymbol{\alpha}_{t(t-1)}]) \tag{4}$$

在式(3)中，$\boldsymbol{W}_\alpha^T \in R^{2p}$ 和 $\boldsymbol{b}_\alpha \in R$ 都是由模型负责学习的参数，分别代表权重矩阵和偏移向量。根据式(4)，RPMC利用softmax函数得到一个权重向量 $\boldsymbol{\alpha}_t$，其中每个元素分别表示与之对应的隐藏层节点在心血管疾病风险预测任务中的重要程度。接着，将隐藏层状态向量 $\boldsymbol{h}_t$ 和上下文向量 $\boldsymbol{c}_t$ 拼接，得到最终表征向量 $\tilde{\boldsymbol{h}}_t$，如式(5)所示：

$$\tilde{\boldsymbol{h}}_t = \tanh(\boldsymbol{W}_c[\boldsymbol{c}_t; \boldsymbol{h}_t]) \tag{5}$$

其中，$\boldsymbol{W}_c \in R^{r \times 4p}$ 是由模型负责学习的权重矩阵。最后将 $\tilde{\boldsymbol{h}}_t$ 输入到softmax层，参与相关计算，如式(6)所示，可以得到类别的概率分布：

$$\hat{y} = \text{softmax}(\boldsymbol{W}_x \tilde{h}_t + \boldsymbol{b}_s) \tag{6}$$

其中，$|y| = 2$ 表示类别的概率分布，$\boldsymbol{W}_x \in R^{2p}$ 和 $\boldsymbol{b}_s$ 分别是由模型负责学习的权重矩阵偏移向量。

2.6 优化目标

为了得到模型参数，RPMC使用预测值与真实值之间的交叉熵作为损失函数，如式(7)所示：

$$L = -\frac{1}{N}\sum_{i=1}^{N} y_i^{\mathrm{T}} \log \hat{y}_i + (1 - y_i)^{\mathrm{T}} \log(1 - \hat{y}_i) \tag{7}$$

其中，y_i 是患者实际的类别标签，1表示心血管疾病高风险患者，而0表示心血管疾病非高风险患者。$\hat{y}_i$ 是RPMC预测出来的类别分布，其中概率值最大的类别为RPMC的最终预测结果。模型采用的优化算法是小批量随机梯度下降算法，由基于TensorFlow和Python 3.5的深度学习框架Keras 2.2.2负责参数的自动计算和更新。

3 实验结果与分析

3.1 基准模型

为了验证Bi-LSTM以及关注机制在时序学习中的优势，首先进行对比实验的基准方法有：逻辑回归(Logical Regression, LR)算法、序列最小优化(Sequential Minimal Optimization, SMO)算法、随机森林(Random Forest, RF)算法、梯度提升决策树算法LightGBM和Bi-LSTM。其中，Bi-LSTM属于时序模型，它能捕获电子病历中的时序信息，而其他4种基准方法是非时序模型。因此，在这4个非时序模型中，都采用BoW模型来表示每个患者的电子病历数据。此外，为了更好地融合不同类型的数据(诊断编码序列和实验室指标序列)，RPMC提出利用多个A-LSTM网络分别对它们进行表征学习。本文还将RPMC与3种最新的深度学习方法(Deepr[4]、Dipole[5]和R-MeHPAN[6])进行性能对比，以进一步证明本研究提出方法的有效性。

3.2 评价指标

在本研究所使用的数据集中，正负样本高度不平衡，比例约为1:6.15。为了客观真实地评估预测模型在不平衡学习问题上的性能，所用到评价指标包括：精准度、召回率、F1值和AUC值。以下是各指标的计算公式：

$$\text{精准率}(P) = \frac{\text{真正例}}{\text{真正例} + \text{假正例}} \tag{8}$$

$$\text{召回率}(R) = \frac{\text{真正例}}{\text{真正例} + \text{假反例}} \tag{9}$$

$$\text{F1值} = \frac{2 \times \text{精准率} \times \text{召回率}}{\text{精准率} + \text{召回率}} \tag{10}$$

其中，真正例、假正例和假反例是根据混淆矩阵计算得来。在心血管疾病风险预测任务中，真正例表示被RPMC正确预测出的心血管疾病高风险患者数目；假正例表示被RPMC预测成为高风险，而实际上是非高风险患者的数目；假反例表示被RPMC预测为非高风险，而实际上是高风险患者的数目。

3.3 模型实现

最终用来实验的数据集被分为3个子集，分别为训练集、验证集和测试集，三者比例为0.7:0.1:0.20。每个预测模型都采用小批量的训练方式，批量大小为1 024。同时为优化模型参数，每个模型迭代100次。此外，为防止过拟合，各模型都采用系数为0.001的二范式正则化方法和早停策略。对于所有基于RNN的预测模型，都统一采用三层隐藏层的网络结构，且各层神经元的个数分别为256、256、128。

3.4 结果分析

3.4.1 时序模型的优势　如表2所示，在基于电子病历的心血管疾病风险预测任务上，时序模型Bi-LSTM和A-LSTM的各项指标基本都优于其他4种非时序模型(逻辑回归、SMO算法、随机森林和LightGBM)。比如，在只利用诊断编码序列数据的情况下，Bi-LSTM的性能可达到0.703 9的召回率、0.654 5的F1值和0.779 8的AUC，这明显优于其他4种中表现最好的LightGBM。之所以Bi-LSTM和A-LSTM能够在心血管疾病风险预测中取得突出的性能，是因为基于LSTM的模型将患者的电子病历数据表示成一个带时间顺序的序列，并且能够从中抽取患者疾病发展过程中的时序性特征。

同时，综合3种不同数据的结果来看，不同类型的数据对心血管疾病风险预测的性能有比较大的影响。基于诊断编码序列数据的预测结果普遍比基于实验室指标数据的要准确。原因之一是实验室指标数据相比诊断编码序列要更稀疏。然而，将两部分数据进行融合之后模型的风险预测结果相比单独使

表2 心血管疾病风险预测模型实验结果

Tab.2 Performances of various models for risk prediction of cardiovascular diseases

模型	诊断序列				实验室检验序列				诊断 & 实验室检验序列			
	P	*R*	F1	AUC	*P*	*R*	F1	AUC	*P*	*R*	F1	AUC
LR	0.567 1	0.672 4	0.615 2	0.757 9	0.524 1	0.612 0	0.564 7	0.728 6	0.591 6	0.696 9	0.640 0	0.772 4
SMO	0.576 4	0.678 6	0.623 3	0.763 2	0.533 3	0.620 7	0.573 7	0.739 0	0.598 0	0.681 0	0.636 8	0.775 7
RF	0.603 4	0.689 8	0.643 7	0.777 1	0.562 0	0.637 8	0.597 5	0.756 7	0.643 7	0.705 3	0.673 1	0.790 4
LightGBM	0.617 1	0.690 1	0.651 5	0.775 9	0.560 6	0.639 3	0.597 4	0.749 3	0.620 0	0.705 1	0.659 8	0.786 1
Bi-LSTM	0.611 5	0.703 9	0.654 5	0.779 8	0.566 8	0.644 2	0.603 1	0.758 7	0.630 5	0.722 0	0.673 1	0.795 4
A-LSTM	0.616 4	0.707 3	0.658 7	0.789 6	0.576 7	0.652 3	0.612 2	0.764 6	0.637 1	0.733 3	0.681 8	0.803 2

用这两种类型数据的预测结果有了显著的提升。当仅使用实验室指标数据时，Bi-LSTM的性能只能达到0.603 1的F1值和0.758 7的AUC。然而，在融合后的数据集上，Bi-LSTM的F1值和AUC分别提升至0.673 1和0.795 4。

此外，A-LSTM在3个数据上的性能都明显优于Bi-LSTM。以基于融合数据的预测模型为例，相比Bi-LSTM，A-LSTM将心血管疾病风险预测的性能从0.630 5的精准率、0.722 0的召回率、0.673 1的F1值和0.795 4的AUC提升到了0.637 1的精准率、0.733 3的召回率、0.681 8的F1值和0.803 2的AUC。这充分表明，关注机制能有效提高心血管疾病风险预测模型的性能，也证明通过关注机制所得到的上下文向量能帮助模型捕获更多潜在的特征信息。

从4种非时序模型在心血管疾病风险预测任务上的性能来看，随机森林和LightGBM明显比其他两种表现突出。主要是因为随机森林和LightGBM属于集成学习框架，分别属于装袋和提升类模型，具备更好的学习能力和泛化能力。

3.4.2 RPMC的优势 如表3所示，相比其他3种深度学习模型，RPMC在心血管疾病风险任务上取得了突出的性能。这表明RPMC能更好地融合不同类型的电子病历数据，因为它充分考虑不同类型数据之间的差异性和关联性。从表中结果可以看出，Deepr的表现相比其他3种较差。原因之一是Deepr是基于CNN的风险预测模型，而相比RNN而言，CNN在电子病历时序学习中只擅长捕获局部信息。另外一个原因是Deepr假设每次医疗就诊记录中的医疗事件都是有时间顺序的，事实上在门诊部门，一次医疗就诊过程持续的时间只有少数几天，期间的医疗事件在EHR中有时并没有严格按照时间顺序进行组织和记录。此外，多个实验室指标通常是同时检测的，并无先后关系，所以Deepr并不是最适合这类医学电子病历的模型。

表3 基于数据融合的心血管疾病风险预测模型的实验结果

Tab.3 Performances of data fusion-based risk prediction models of cardiovascular diseases

模型	无人口学数据				有人口学数据			
	P	*R*	F1	AUC	*P*	*R*	F1	AUC
Deepr	0.600 8	0.719 6	0.654 9	0.776 8	0.634 7	0.732 7	0.680 2	0.798 3
$Dipole_l$	0.602 3	0.732 1	0.660 9	0.780 3	0.631 4	0.784 3	0.699 6	0.801 1
R-MeHPAN	0.663 1	0.769 6	0.712 4	0.812 4	0.676 3	0.773 7	0.721 7	0.811 3
RPMC	0.671 7	0.805 6	0.732 5	0.821 9	0.674 0	0.814 9	0.737 8	0.837 5

值得注意的是，模型R-MeHPAN在心血管疾病风险预测任务上明显比$Dipole_l$表现得好。$Dipole_l$将不同类型的数据融合在一起进行模型的训练，而R-MeHPAN将不同类型的数据分开进行各自的表征学习。这表明，在该实证数据集上两种类型的数据（疾病编码序列和实验室指标序列）之间存在比较大的差异，而且这些差异对心血管疾病的风险预测有很大帮助。

此外，向心血管疾病风险预测模型中加入人口学数据之后，各深度模型的性能普遍都有所提高。以性能提升最突出的Dipole$_l$为例，人口学数据的加入使得它的性能从0.602 3的精准率、0.732 1的召回率、0.660 9的F1值和0.780 3的AUC提升至0.631 4的精准率、0.784 3的召回率、0.699 6的F1值和0.801 1的AUC，各指标分别提升0.029 1、0.052 2、0.038 7和0.020 8。这表明，人口学数据对心血管疾病风险预测的准确性有很大帮助。同时，从表3中还可以看出，RPMC始终是表现最突出的预测模型。

总之，RPMC在心血管疾病风险预测任务上的突出性能表明，RPMC能将不同类型的电子病历数据有效地结合，不仅能充分考虑到它们之间的差异性，还同时能捕获到他们之间潜在的关联性。

3.4.3 模型可解释性分析　由关注机制算出的权重矩阵，能够用来衡量每个患者中每次医疗就诊对心血管疾病风险预测的重要程度。因此，可以有效增强心血管疾病风险预测模型RPMC的可解释性。对每一位患者，关注机制都会算出一个权重向量，选取权重最大的一次就诊记录作为后续统计分析的依据。

如图5所示，根据所有患者最重要的就诊记录集合统计得到关注机制中出现频次最高的前10种疾病诊断编码。该结果表明，糖尿病和高血压对心血管疾病风险预测有着突出的影响，这和临床上的研究结果基本一致[22]。值得注意的是，表中有5种诊断编码是以字母“E”开头的。根据ICD10的分类标准，字母“E”开头的疾病都属于内分泌、营养和代谢疾病种类。已有研究表明，患有代谢综合症状的个体发展成心血管疾病的风险是其他人的两倍左右[23]。根据已有文献的相关结果表明，癫痫和心血管疾病的风险因素（如糖尿病、高血压和高胆固醇等）有着密切的关联[24]。从图5的分析结果也可以明显看出，癫痫（G40）的出现频次排在第5位，说明其对心血管疾病风险预测具有显著的作用。此外，诊断编码K75和K29在ICD10分类标准中属于消化系统疾病种类，对心血管疾病风险预测也有着重要的帮助。Klimenko等[25]在医学上证实消化系统的状态对心血管总体的胆固醇水平有很大影响，而胆固醇是临床上公认的心血管疾病高风险因素之一。

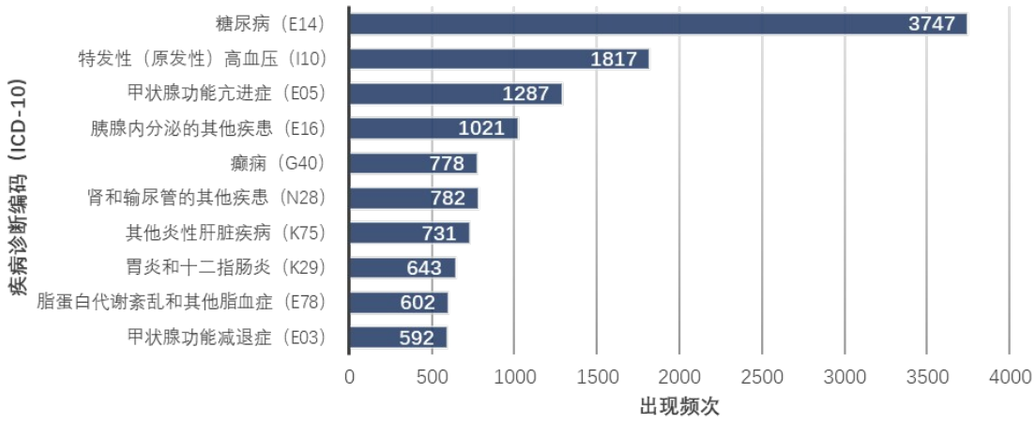

图5　关注机制中出现频次最高的前10种疾病诊断编码

Fig.5 Top 10 frequent diagnosis codes obtained by attention mechanism

4 结 语

在基于电子病历的心血管疾病风险预测任务中，A-LSTM是一个非常合适的表征学习方法，它完全端对端，不需要人工干预，同时又能捕获潜在的时序信息。RPMC分别利用3个A-LSTM对患者的电子病历进行表征学习，能充分考虑不同数据类型之间的差异性和潜在关联性。通过多组实验的比较和多个角度的分析，RPMC在心血管疾病风险预测任务上均获得了相对最佳的性能。

在接下来的工作中，将融入更多类型的电子病历数据，比如临床文本、医嘱信息以及影像数据等。此外，还将进一步验证和优化RPMC的可扩展性和泛化性，以便其能有效处理其他更多的临床问题。

【参考文献】

[1] 陈伟伟, 高润霖, 刘力生, 等. 中国心血管病报告2016[J]. 中国循环杂志, 2014, 32(6): 521-530.
CHEN W W, GAO R L, LIU L S, et al. China cardiovascular disease report 2016[J]. Chinese Circulation Journal, 2014, 32(6): 521-530.

[2] URIARTE-ARCIA A V, LOPEZ-YANEZ I, YANEZ-MARQUEZ C. One-hot vector hybrid associative classifier for medical data classification[J]. PLoS One, 2014, 9(4): e95715.

[3] SIVIC J, ZISSERMAN A. Efficient visual search of videos cast as text retrieval[J]. IEEE Trans Pattern Anal Mach Intell, 2009, 31(4): 591-606.

[4] NGUYEN P, TRAN T, WICKRAMASINGHE N, et al. Deepr: a convolutional net for medical records[J]. IEEE J Biomed Health, 2017, 21(1): 22-30.

[5] MA F, CHITTA R, ZHOU J, et al. Dipole: diagnosis prediction in healthcare *via* attention-based bidirectional recurrent neural networks [C]//ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2017): Halifax, Nova Scotia, Canada, 2017: 13-17.

[6] KIM Y J, LEE Y G, KIM J W, et al. Highrisk prediction from electronic medical records *via* deep attention networks[C]//31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 2017.

[7] XU J, ZHANG Y, AZHAR M, et al. Data mining on ICU mortality prediction using early temporal data: a survey[J]. Int J Inf Tech Decis, 2017, 16(1): 117-159.

[8] HULING J D, YU M, LIANG M, et al. Risk prediction for heterogeneous populations with application to hospital admission prediction: risk prediction for heterogeneous populations [J]. Biometrics, 2018, 74(2): 557-565.

[9] EVERETT B M, BERGER J S, MANSON J A, et al. B-Type natriuretic peptides improve cardiovascular disease risk prediction in a cohort of women[J]. J Am Coll Cardiol, 2014, 64(17): 1789-1797.

[10] WELSH P, HART C, PAPACOSTA O, et al. Prediction of cardiovascular disease risk by cardiac biomarkers in 2 united kingdom cohort studies: does utility depend on risk thresholds for treatment[J]. Hypertension, 2016, 67(2): 309-315.

[11] HUANG Z, GE Z, DONG W, et al. Relational regularized risk prediction of acute coronary syndrome using electronic health records [J]. Inf Sci, 2018, 465(10): 118-129.

[12] JIANG S, CHIN K S, QU G, et al. An integrated machine learning framework for hospital readmission prediction[J]. Knowl-Based Syst, 2018, 146(4): 73-90.

[13] PIKE M M, DECKER P A, LARSON N B, et al. Improvement in cardiovascular risk prediction with electronic health records[J]. J Cardiovasc Transl, 2013, 9(3): 214-222.

[14] KENNEDY E H, WIITALA W L, HAYWARD R A, et al. Improved cardiovascular risk prediction using nonparametric regression and electronic health record data[J]. Med Care, 2013, 51(3): 251-258.

[15] NG K, STEINHUBL S R, DEFILIPPI C, et al. Early detection of heart failure using electronic health records practical implications for time before diagnosis, data diversity, data quantity, and data density[J]. Circ-Cardiovasc Qual, 2016, 9(6): 649-658.

[16] 刘建伟, 刘媛, 罗雄麟. 深度学习研究进展[J]. 计算机应用研究, 2014, 31(7): 1921-1930.
LIU J W, LIU Y, LUO X L. Research and development on deep learning [J]. Application Research of Computers, 2014, 31(7): 1921-1930.

[17] FAKHRY A, ZENG T, JI S. Residual de-convolutional networks for brain electron microscopy image segmentation[J]. IEEE Trans Med Imaging, 2017, 36(2): 447-456.

[18] ZHONG Z, LI J, LUO Z, et al. Spectral-spatial residual network for hyperspectral image classification: a 3-d deep learning framework[J]. IEEE Trans Geosci Remote, 2018, 56(2): 847-858.

[19] LUONG M, PHAM H, MANNING C D. Effective approaches to attention-based neural machine translation[C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP), Lisbon, Portugal, 2015.

[20] HEFFERNAN R, YANG Y, PALIWAL K. Capturing non- local interactions by long short-term memory bidirectional recurrent neural networks for improving prediction of protein secondary structure, backbone angles, contact numbers and solvent accessibility [J]. Bioinformatics, 2017, 33(18): 2842-2849.

[21] CHOI E, BAHADORI 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 (KDD 2016), San Francisco, CA, USA, 2016.

[22] CHOBANIAN A V, BAKRIS G L, BLACK H R, et al. Seventh report of the joint national committee on prevention, detection, evaluation, and treatment of high blood pressure[J]. Blood, 2003, 118(15): 1206-1252.

[23] DEFILIPP Z, DUARTE R F, SNOWDEN J A, et al. Metabolic syndrome and cardiovascular disease following hematopoietic cell transplantation: screening and preventive practice recommendations from CIBMTR and EBMT[J]. Bone Marrow Transpl, 2017, 52(2): 173-182.

[24] VIVANCO-HIDALGO R M, GOMEZ A, MOREIRA A, et al. Prevalence of cardiovascular risk factors in people with epilepsy[J]. Brain Behav, 2017, 7(2): e00618.

[25] KLIMENKO E D, MARTSEVICH M S, MUKHINA A P, et al. Interaction between the cardiovascular and digestive systems in the genesis of experimental atherosclerosis[J]. Bull Exp Biol Med, 1975, 80(4): 1164-1167.

（编辑：陈丽霞）