ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 38,No. 4,Jul. 2023,pp. 777－791
DOI:10. 16337/j. 1004-9037. 2023. 04. 003

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

基于知识表示向量的可解释深度学习模型及其疾病预测应用

徐　鹤[1,2]，郑群力[1,2]，谢作玲[3]，程海涛[1,2]，李　鹏[1,2]，季一木[1,2]

(1. 南京邮电大学计算机学院/软件学院/网络空间安全学院,南京 210023;2. 江苏省高性能计算与智能处理工程研究中心,南京 210023;3. 东南大学附属中大医院内分泌科,南京 210009)

摘　要：近年来,深度学习方法广泛应用于各种疾病预测任务,甚至在其中一些方面超过了人类专家。然而,算法的黑盒性质限制了其临床应用。对此,本文结合知识表示学习和深度学习方法构建了一种融入知识表示向量的可解释深度学习模型。该模型首先依据体检指标正常范围构建体检指标与检测值之间的关系图,并通过基于知识表示学习的深度学习模型对人体体检指标与检测值关系图进行编码,然后将患者体检数据表示为向量,输入到构建的自注意力机制和卷积神经网络构建的分类器中来实现疾病预测。将模型应用于糖尿病预测实验中,其准确率和召回率均优于对比的机器学习方法。与表现较优的随机森林算法相比,模型的准确率和召回率分别提升了0.81%和5.21%。实验结果表明,通过可解释性方法将知识表示学习和深度学习技术融合应用于糖尿病预测,可以达到对糖尿病的早期发现与辅助诊断的目的。

关键词：疾病预测;知识表示学习;深度学习;自注意力机制;卷积神经网络;可解释性

中图分类号：TP391　　**文献标志码**:A

Interpretable Deep Learning Model Based on Knowledge Representation Vectors and Its Application in Disease Prediction

XU He[1,2], ZHENG Qunli[1,2], XIE Zuoling[3], CHENG Haitao[1,2], LI Peng[1,2], JI Yimu[1,2]

(1. School of Computer Science/School of Software/School of Cyberspace Security, Nanjing University of Posts and Telecommunications, Nanjing 210023, China; 2. Jiangsu HPC and Intelligent Processing Engineer Research Center, Nanjing 210023, China;3. Department of Endocrinology, Zhongda Hospital Southeast University,Nanjing 210009, China)

Abstract: In recent years, deep learning methods have been widely applied to various disease prediction tasks, even surpassing human experts in some aspects. However, the black box nature of the algorithm limits its clinical application. In this paper, the knowledge representation and reasoning learning and deep learning methods are combined to build an interpretable deep learning model incorporating knowledge representation and reasoning vectors. The model first builds a relationship graph between physical examination indicators and test values according to the normal range of physical examination indicators, and the relationship graph between physical examination indicators and test values is coded through the deep learning model based on knowledge representation and reasoning learning. Then, the patients'

基金项目:江苏省科技支撑计划项目(BE2019740);江苏省六大人才高峰高层次人才项目(RJFW-111)。
收稿日期:2022-04-19;**修订日期**:2023-07-06

physical examination data are expressed as vectors, which are input into the self-attention mechanism and the classifier constructed by convolutional neural network to realize the disease prediction. When the model is applied to the prediction experiment of diabetes, the accuracy and recall of the model are better than those of the comparative machine learning methods. Compared with the random forest algorithm, the accuracy and recall are also improved by 0.81% and 5.21%, respectively. Experimental results show that the application of knowledge representation and reasoning learning and deep learning technological convergence to diabetes prediction through interpretable methods can achieve the purpose of early detection and auxiliary diagnosis of diabetes.

Key words: disease prediction; knowledge representation learning; deep learning; self-attention mechanism; convolutional neural network; interpretability

引　言

深度学习是机器学习的一个重要分支领域[1]。近年来,随着海量数据的可用性和计算机算力的提高,基于深度学习的智能系统在很多领域已经达到甚至超过了人类专家的水平,例如在语音识别[2]、图像分类[3]、自然语言处理[4]等领域。随着模型复杂度的增加,深度学习算法缺乏可解释性,产生的结果变得难以预测和不可解释。此外,在一些领域应用中,将深度学习算法训练出的模型看作是黑盒,阻碍了深度学习在该领域的应用[5-6],例如金融、医疗和自动驾驶等领域。

深度学习在医疗领域当前主要用于医学影像的处理[7]和电子健康记录(Electronic health record, EHR)处理[8-9],并在视网膜疾病检测[10]、糖尿病预测[11]及阿尔茨海默病分类[12]等医学任务上取得了显著成果。尽管如此,基于深度学习的方法却尚未在临床上取得广泛应用。这是由于深度学习算法潜在的黑盒性质造成的。深度学习算法提供的视觉和文本解释似乎合理,但是算法的决策细节并未清晰的暴露出来,具有不透明性[13]。虽然研究人员了解这些系统的体系结构以及生成用于分类的模型的过程,但模型本身对患者或医生来说难以理解。此外,缺乏可解释性的自动医疗诊断模型会给患者诊断出错误的治疗方案,甚至严重威胁患者的生命安全[14],而且将医疗决策权交给黑箱系统也有违临床医生的道德责任[15]。

由于临床医学的复杂性以及深度学习算法的黑盒性,使得任何深度模型都不可能实现完美的决策。对此,临床医生将解释性视为在模型预测的背景下证明其临床决策合理性的一种手段,并指出相关模型需反映出与医学决策制定方法类似的分析过程[16]。例如,在实际临床诊断过程中,医生通常会以患者的体检数据作为参考,根据体检指标的正常值范围来做出相应的判断。

此外,大多数深度学习模型使用数据驱动的方法,但大部分数据具有不确定性,这种不确定性可能来自嘈杂、缺失的数据或数据中存在的固有不确定性。因此,为了增强深度学习模型的可解释性,可采用知识驱动的方法,向模型中嵌入外部人类知识[6]。例如,将人类的领域知识表示成领域的知识图谱[17]形式,并采用知识表示学习方法[18]对领域知识进行编码,与深度学习等技术相融合以构建具有可解释性的深度学习模型。

综上所述,为了增强深度学习在医疗领域的可解释性,模型需反映出与医学诊断类似的分析过程,或者嵌入外部人类知识。因此,本文将医学中体检指标的正常范围作为外部知识,模拟医生依据体检数据的诊断过程,提出了一种融入知识表示向量的可解释深度学习模型,并将其应用于糖尿病预测中,其主要贡献总结如下:

(1) 依据常规体检指标正常值范围,并采用知识表示学习方法,构建了糖体检指标和检测值表示向量。该表示向量能准确地描述体检指标和检测值之间存在的偏高、偏低等关系,可提高一些疾病预测模型的可解释性。

(2) 提出了一种融入知识表示向量的可解释深度学习模型,该模型使用构建好的体检指标和检测值表示向量,得到体检数据的关系矩阵,然后通过自注意力机制关联每个体检指标,并使用卷积神经网络进行特征提取,从而应用于糖尿病的预测。

(3) 将本文构建的融入知识表示向量的可解释深度学习模型与经典的支持向量机、随机森林等机器学习模型进行对比实验。结果表明,本文模型在准确率和召回率两方面均优于对比的机器学习模型,说明本文提出的可解释深度学习模型具有较好的疾病预测效果。

1　相关工作

1.1　医疗领域可解释性研究进展

在医学领域,可解释问题包含其他领域不考虑的因素,例如风险和责任[19]。医疗决策往往伴随着生命风险,将如此重要的决策交给无法提供责任且缺乏解释性的机器,无异于推卸责任,且可能导致灾难性后果[13]。因此,大量研究人员开始进行面向医学领域的可解释性深度学习模型的研究[15-16,20]。

早期可解释模型通过讨论输入变量对输出的作用及意义来增强可解释性。例如,Haufe等[21]讨论了用于估计大脑状态的不同线性模型,包括它是如何被误解的;比较了前向模型和后向模型,并提出了对线性模型的改进建议。Caruana等[22]通过逻辑回归模型发现了哮喘与肺炎死亡风险降低之间的关系,在回归模型中,哮喘作为风险预测因子的权重为负。Varol等[23]使用生成判别机(Generative discriminator model, GDM)结合普通最小二乘回归和岭回归处理阿尔茨海默病和精神分裂症数据集中的混杂变量。其中,GDM参数被认为是可解释的,因为它们是临床变量的线性组合。虽然这类简单模型更易于解释,但是通常会牺牲模型的性能。文献[24]中指出,与可解释性模型相比,复杂模型(如深度神经网络)通常可获得更高的性能,因此多数情况下,更倾向于使用这些复杂模型。

为了权衡模型的性能和可解释性,研究人员开始致力于解释复杂的深度黑盒模型,其中大部分人首先从医学图像的可解释性入手,进行了一系列的研究[25-28]。比如Van等[26]尝试通过可视化学习到的特征图来解开皮肤病领域卷积神经网络的黑盒。他们发现,在某种程度上,卷积神经网络关注的特征与皮肤科医生用于诊断的特征相似。但该方法存在的问题是无法解释模型检测到的特征与其输出之间的因果关系,不具有通用性。此外,注意力机制也常作为可解释的医学图像分析的深度学习工具。如文献[27]提出了一种新的测试概念激活向量(Testing concept activation vectors,TCAV)方法,用人类可理解的概念向领域专家解释不同层次学习的特征。TCAV 使用显著图方法解释了糖尿病视网膜病变水平并实现检测视网膜中存在的微动脉瘤和动脉瘤。

然而,仅仅针对医学图像的研究缺乏医学专业知识的支撑,对此,Zhang等[29]提出了一个融合语义和视觉可解释的医学图像诊断网络MDNet,为可解释深度学习技术在医疗图像诊断中应用提供了一个新的视角:生成诊断报告和与报告对应的网络关注,借助于注意力机制使得网络诊断和决策过程具有语义和视觉上的可解释性。此外,为了嵌入外部知识,知识图谱在医疗领域中的应用也越来越多[30]。例如,刘勘等[31]结合知识图谱、表示学习和深度神经网络等方法构建了一种可解释的并发症辅助诊断模型。

从上述研究中可以看出,医疗领域的可解释研究大多针对于医学图像数据,而且缺乏专业知识的支撑。对此,本文将结合知识表示学习和深度学习技术构建一种融入知识表示向量的可解释深度学习

模型,针对医疗领域的体检数据,进行高精度且可解释的疾病辅助诊断。

1.2 知识表示学习模型

通常,传统的知识图谱是以三元组(h, r, t)表示,其中h表示头实体,t表示尾实体,r表示关系。知识表示学习将研究对象(实体和关系)表示稠密低维实值向量[32]。研究者提出了多种知识表示模型,本文将介绍目前性能比较稳定的TransE[33]、TransH[34]和TransR[35]模型,模型架构如图1所示。

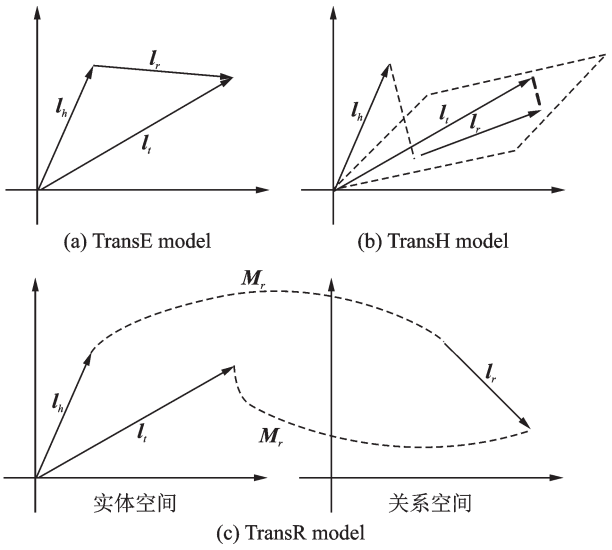

图1 3种知识表示模型

Fig.1 Three knowledge representation models

TransE模型[33]用关系r的向量$\boldsymbol{l}_r$作为头实体向量$\boldsymbol{l}_h$和尾实体向量$\boldsymbol{l}_t$之间的平移,它们之间满足

$$\boldsymbol{l}_h+\boldsymbol{l}_r\approx\boldsymbol{l}_t \tag{1}$$

其损失函数可表示为

$$f_r(h,t)=|\boldsymbol{l}_h+\boldsymbol{l}_r-\boldsymbol{l}_t|_{L_1/L_2} \tag{2}$$

即向量$\boldsymbol{l}_h+\boldsymbol{l}_r$和$\boldsymbol{l}_t$的$L_1$或$L_2$距离。

TransE模型参数较少,计算复杂度较低,可扩展性强,但因为模型简单,在处理复杂关系时性能显著降低。例如,在一对多的关系中,假设知识库中有两个三元组,分别是(糖尿病,并发症,糖尿病肾病)和(糖尿病,并发症,糖尿病足),如果使用TransE模型,会使得糖尿病肾病和糖尿病足的向量变得相同,这显然不符合事实。针对TransE处理复杂关系的不足,Wang等[34]和Lin等[35]分别提出了改进的TransH和TransR模型。

TransH模型[34]首先将头实体向量$\boldsymbol{l}_h$和尾实体向量$\boldsymbol{l}_t$沿法线投影到关系r对应的超平面上,可分别用$\boldsymbol{l}_{h_r}$和$\boldsymbol{l}_{t_r}$表示为

$$\boldsymbol{l}_{h_r}=\boldsymbol{l}_h-\boldsymbol{w}_r^{\mathrm{T}}\boldsymbol{l}_h\boldsymbol{w}_r \tag{3}$$

$$\boldsymbol{l}_{t_r}=\boldsymbol{l}_t-\boldsymbol{w}_r^{\mathrm{T}}\boldsymbol{l}_t\boldsymbol{w}_r \tag{4}$$

其损失函数可表示为

$$f_r(h,t)=\left\|\boldsymbol{l}_{h_r}+\boldsymbol{l}_r-\boldsymbol{l}_{t_r}\right\|_{L_1/L_2} \tag{5}$$

TransR模型[35]通过定义投影矩阵$\boldsymbol{M}_r\in\mathbf{R}^{d\times k}$,实现将实体向量投影到其关系$r$的子空间,可分别用

$\boldsymbol{l}_{h_r}$和$\boldsymbol{l}_{t_r}$表示为

$$\boldsymbol{l}_{h_r}=\boldsymbol{l}_h\boldsymbol{M}_r \tag{6}$$

$$\boldsymbol{l}_{t_r}=\boldsymbol{l}_t\boldsymbol{M}_r \tag{7}$$

然后使$\boldsymbol{l}_{h_r}+\boldsymbol{l}_r\approx\boldsymbol{l}_{t_r}$,其损失函数为

$$f_r(h,t)=\left\|\boldsymbol{l}_{h_r}+\boldsymbol{l}_r-\boldsymbol{l}_{t_r}\right\|_{L_1/L_2} \tag{8}$$

式(5)和式(8)中的L_1、L_2表示向量$\boldsymbol{l}_h+\boldsymbol{l}_r$和$\boldsymbol{l}_t$的$L_1$或$L_2$距离。

2　模型架构

本文提出的融入知识表示向量的可解释深度学习模型,旨在模拟医生依据患者体检数据进行疾病诊断的过程,其核心思想主要是利用知识表示学习模型和外部体检知识构建体检指标实体和检测值实体的表示向量,然后得到患者体检数据的矩阵表示,并输入到深度学习模型中,从而实现对疾病的预测。

融入知识表示向量的可解释深度学习模型架构如图2所示,主要分为3个部分:

(1) 依据体检指标检测值的正常范围,构建体检指标与检测值的关系图,然后利用知识表示学习模型,获取体检指标和检测值的表示向量。

(2) 获取患者的体检数据,根据(1)中体检指标和检测值的表示向量,得到所有体检指标与对应检测值之间的关系向量,并拼接成关系矩阵。

(3)将其关系矩阵输入到自注意力机制(Self-attention)和卷积神经网络(Convolutional neural networks,CNN)构建的分类器中,得出糖尿病的预测结果。

本文将提出的模型简称为TH-SAC,即TransH-Self-Attention-CNN。

2.1　体检指标与检测值的表示向量

在疾病实际临床诊断中,医生常会结合患者体检数据和已有的体检知识来做出判断。例如,在糖尿病的临床诊断中,空腹血糖值的正常范围为3.9～6.1 mmol/L[36],当患者空腹血糖值大于7.0 mmol/L时,则考虑可能患有糖尿病。本文考虑在模型中嵌入医学领域的专业知识,首先将体检指标与检测值之间的关系划分为以下7类:严重偏低、一般偏低、轻微偏低、正常、轻微偏高、一般偏高和严重偏高,并将这些体检知识转化成三元组的形式,例如(空腹血糖,轻微偏高,7.1 mmol/L),(空腹血糖,正常,6.0 mmol/L)等。

由于体检指标与对应的检测值之间存在一对多和多对一的复杂关系,本文选择的TransH模型符合这种关系表示。因此,将体检知识转换成三元组的形式,使用TransH知识表示模型表示。该模型采用$\boldsymbol{l}_h+\boldsymbol{l}_h\approx\boldsymbol{l}_t$为基本思想,并使用平移向量$\boldsymbol{l}_r$和超平面的法向量$\boldsymbol{w}_r$来表示关系$r$。根据式(3,4)计算得到实体向量$\boldsymbol{l}_h$和$\boldsymbol{l}_t$在关系$r$所在的超平面上的投影向量$\boldsymbol{l}_{h_r}$和$\boldsymbol{l}_{t_r}$,再根据式(5)得到体检指标和检测值实体低维稠密表示向量$\boldsymbol{e}_H$。

2.2　体检指标与检测值的关系向量

得到体检知识实体的向量表示后,为了能够在模型中体现出体检指标与其对应检测值之间存在的关系,本文基于知识表示学习模型的基本思想$\boldsymbol{l}_h+\boldsymbol{l}_r\approx\boldsymbol{l}_t$,用每个体检指标实体向量与其对应的检测值实体向量之差来表示它们之间的关系,即

$$\boldsymbol{e}_r=\boldsymbol{e}_v-\boldsymbol{e}_c \tag{9}$$

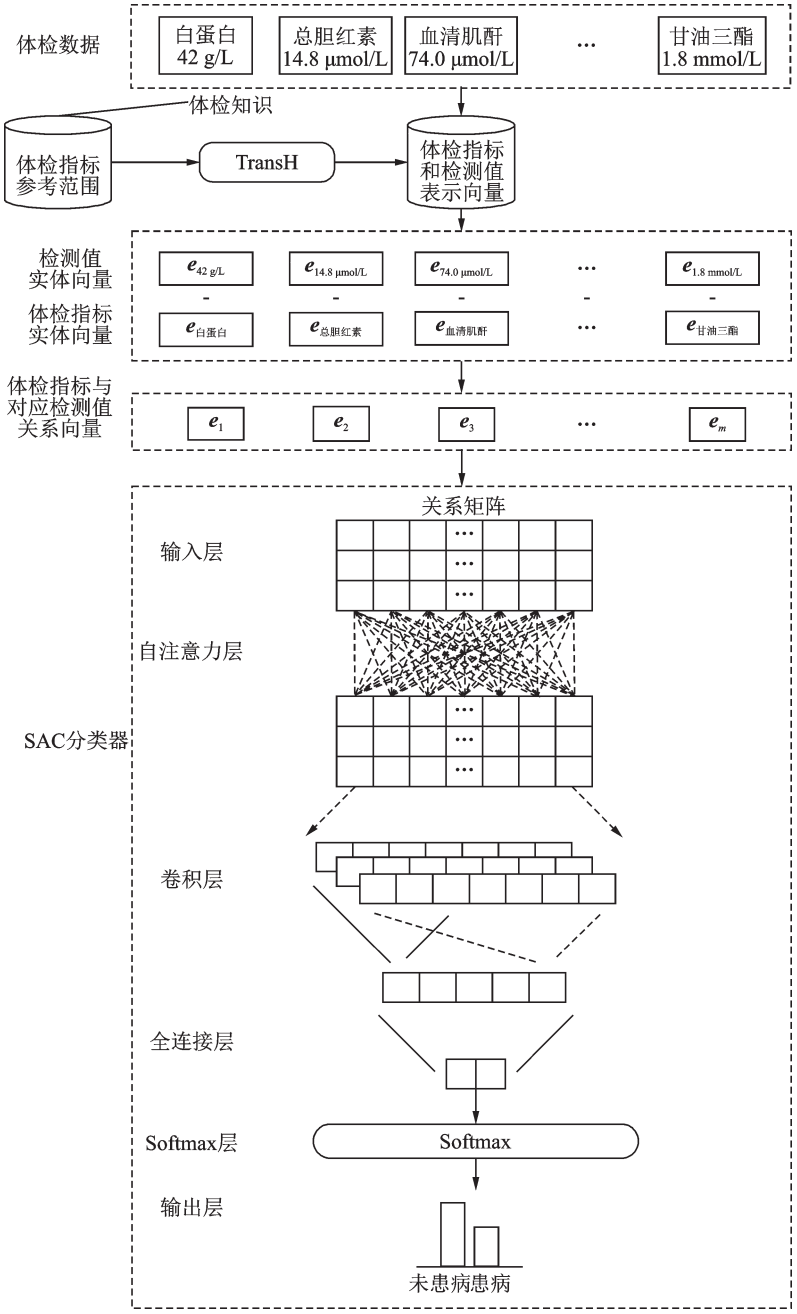

图2 融入知识表示向量的可解释深度学习模型架构图

Fig.2 Architecture diagram of interpretable deep learning model incorporating knowledge representation vectors

式中：$\boldsymbol{e}_v$为检测值实体向量；$\boldsymbol{e}_c$为体检指标实体向量。

例如体检指标空腹血糖实体向量$\boldsymbol{e}_{\text{空腹血糖}}$和检测值“7.1 mmol/L”实体向量$\boldsymbol{e}_{7.1\ \text{mmol/L}}$之间的关系表示为$\boldsymbol{e}_{7.1\ \text{mmol/L}}-\boldsymbol{e}_{\text{空腹血糖}}$。将所有体检指标与其对应检测值之间的关系向量拼接起来，可构成患者体检指标与检测值之间的关系矩阵，可表示为

$$\boldsymbol{E}^{m\times k}=\left[\boldsymbol{e}_r^1,\boldsymbol{e}_r^2,\boldsymbol{e}_r^3,\cdots,\boldsymbol{e}_r^m\right] \tag{10}$$

式中:k为实体向量的维度;m为体检指标的个数。

2.3 SAC分类器

SAC分类器为图1中TH-SAC模型的下侧部分,主要由以下几层组成:

(1) 输入层:将所有体检指标与对应检测值之间的关系向量拼接起来得到的关系矩阵$\boldsymbol{E}^{m\times k}$即为该分类器的输入。

(2) 自注意力层:由于每个体检指标之间相互关联,所以将关系矩阵$\boldsymbol{E}^{m\times k}$进一步输入自注意力层中,使得每个体检指标获取全局信息,这符合当前的医学诊断经验。本文自注意力层采用的层数为2缩放点积注意力机制。在注意力层中,每个关系向量$\boldsymbol{e}_r^i$被线性映射到3个不同的空间,得到查询向量$\boldsymbol{q}_i$、键向量$\boldsymbol{k}_i$和值向量$\boldsymbol{v}_i$。对于每个查询向量$\boldsymbol{q}_i$,根据式(11)计算输出向量$\boldsymbol{e}_{\text{attn}}$。

$$\boldsymbol{e}_{\text{attn}}^i=\sum_{j=1}^{m}a_{ij}\boldsymbol{v}_i \tag{11}$$

式中:a_{ij}表示第i个输出关注到第j个输入的权重,可表示为

$$a_{ij}=\text{softmax}(s(\boldsymbol{k}_j,\boldsymbol{q}_i)) \tag{12}$$

$$s(\boldsymbol{k}_j,\boldsymbol{q}_i)=\frac{\boldsymbol{k}_j^{\mathrm{T}}\boldsymbol{q}_i}{\sqrt{D_k}} \tag{13}$$

式中:softmax(·)为按列进行归一化的函数;D_k为$\boldsymbol{q}_i$的维度。

为了同时计算关系矩阵$\boldsymbol{E}^{m\times k}$中每个关系向量对应的输出向量,可将查询向量$\boldsymbol{q}_i$、键向量$\boldsymbol{k}_i$和值向量$\boldsymbol{v}_i$分别合并成查询矩阵$\boldsymbol{Q}$,键矩阵$\boldsymbol{K}$,值矩阵$\boldsymbol{V}$,然后根据式(14)得到自注意力层输出矩阵。

$$\boldsymbol{E}_{\text{attn}}=\boldsymbol{V}\text{softmax}\left(\frac{\boldsymbol{K}^{\mathrm{T}}\boldsymbol{Q}}{\sqrt{D_k}}\right) \tag{14}$$

(3) 卷积层:通过自注意力层获取了全局信息后,为了更深层次挖掘关系矩阵中的信息,将自注意力层的输出矩阵$\boldsymbol{E}_{\text{attn}}$输入到卷积神经网络。假设$\boldsymbol{W}^f\in\mathbf{R}^{h\times d}$,$h$为滤波器窗口大小,$d$表示输入向量的维度。对于输入的从$i$行到$i+k-1$行的局部特征$\boldsymbol{e}_{\text{attn}}^{i:(i+h-1)}$,卷积滤波器所提取的特征子矩阵的第$i$个特征值可表示为

$$c_i=f(\boldsymbol{w}^f\cdot\boldsymbol{e}_{\text{attn}}^{i:i+h-1}+b) \tag{15}$$

式中:$f(\cdot)$为非线性激活函数Relu(·);b为偏置值。

因此,注意力层得到的输出矩阵$\boldsymbol{E}_{\text{attn}}$的局部特征矩阵为

$$\boldsymbol{C}=[c_1,c_2,c_3,\cdots,c_{m-h+1}] \tag{16}$$

接着对特征映射进行最大池化操作,即

$$\hat{c}=\max\{\boldsymbol{C}\} \tag{17}$$

最终,得到体检数据最终的表示向量的表达式为

$$\boldsymbol{Z}_{tj}=[\hat{c}_1,\hat{c}_2,\hat{c}_3,\cdots,\hat{c}_n] \tag{18}$$

(4)全连接层和Softmax层:将上述得到的体检数据的表示向量经全连接层的变换后得到患者是否患有糖尿病的得分向量$\boldsymbol{s}$,全连接层的隐藏单元个数为2,即患有糖尿病和未患糖尿病,最后将得分向量$\boldsymbol{s}$输入到Softmax层,使其转化成一个条件概率分布,即

$$p_i(\boldsymbol{s})=\frac{\exp(s_i)}{\sum_{j=1}^{2}\exp(s_j)}\qquad i=1,2 \tag{19}$$

整个模型采用交叉熵损失函数来衡量糖尿病预测概率分布与真实概率分布之间的差距,并通过反向传播算法来训练和更新模型的参数。损失函数可表示为

$$\text{loss}=-\frac{1}{N}\sum_{i}\left[y_i\cdot\lg p_i+(1-y_i)\cdot\lg(1-p_i)\right] \tag{20}$$

式中:N表示样本数目;y_i表示样本i的真实标签,患有疾病为1,未患有疾病为0。

3 实验与分析

3.1 实验数据

实验中使用到的数据主要有:

(1)用于构建体检指标和检测值实体表示向量的外部体检知识,来源于某三甲医院提供的糖尿病体检指标检测值的参考范围,如表1所示,给出了部分体检指标的检测值参考范围。依据这些体检知识,本文共构建了5 518个相关实体,7种关系实体(严重偏低、一般偏低、轻微偏低、正常、轻微偏高、一般偏高和严重偏高)以及9 410个三元组关系。实体类型及其数量如表2所示,关系类型及其数量如表3所示,其中,由于无法预知实际中每个体检指标的临界值,故将大于(小于)实验中设定的最大值(最小值)的检测值实体统一当作是异常偏高<HIGEST>实体(异常偏低<LOWEST>实体)。此外,所有的缺失值项均用未知实体<UNK>来代替。

表1 部分体检指标检测值参考范围

Table 1 Reference range of detection value of some physical examination indexes

体检指标	参考范围
血清谷丙转氨酶/(IU•L)	9～50
血清谷草转氨酶/(IU•L)	15～40
白蛋白/(g•L^{-1})	40.0～55.0
总胆红素/(μmol•L^{-1})	2.0～20.0
血尿素氮/(mmol•L^{-1})	3.6～9.5
总胆固醇/(mmol•L^{-1})	2.86～6.10
甘油三酯/(mmol•L^{-1})	0.45～1.81
低密度脂蛋白/(mmol•L^{-1})	0.00～3.37
高密度脂蛋白/(mmol•L^{-1})	1.16～1.42
…	…

表2 实体类型及其数量

Table 2 Entity type and its quantity

实体类型	举例	实体数量
体检指标	甘油三酯	16
检测值	1.62 mmol/L	5 499
异常偏高	<HIGHEST>	1
异常偏低	<LOWEST>	1
未知	<UNK>	1

表3 关系类型及其数量

Table 3 Relationship type and its quantity

关系类型	实体数量
严重偏低	337
一般偏低	343
轻微偏低	457
正常	1 558
轻微偏高	2 663
一般偏高	2 005
严重偏高	2 017

(2)采用一家大型公司提供的糖尿病患者的体检数据,其中包含血清谷丙转氨酶、血清谷草转氨酶、和白蛋白等11个常规体检指标,总共有48 887条数据,其中训练集用80%的数据,测试集用20%的数据,具体如表4所示。

表4　体检数据集分布

Table 4　Distribution of physical examination data set

患病标签	训练集	测试集
糖尿病	3 815	954
非糖尿病	35 924	8 824
总数	39 109	9 778

3.2　实验设置

实验中主要使用的是Pytorch深度学习框架和OpenKE知识表示学习框架,本文模型的具体参数设置如表5所示。

表5　模型参数设置

Table 5　Model parameter setting

参数	数值
优化器	Adam
Batch_size	32
Epoch	100
Dropout	0.5
学习率	0.000 2
体检数据实体向量维度	256
卷积滤波器窗口大小	2,3,4
每种窗口大小卷积滤波器个数	100
自注意力层层数	2

3.3　评价指标

采用使用准确率(Accuracy)和召回率(Recall)作为结果的评价指标。此外,选取Mean rank(MR)和Hit@10作为知识表示模型的评价指标。

(1) Mean rank

在评估知识表示学习模型性能时,会对每个评测的三元组(h,r,t),移去头部实体,依次替换成知识库中的其他实体,构建错误的三元组实体(h',r,t)。利用关系函数$f_r(h,t)$计算头部实体和尾部实体的相似度,得到所有的三元组(包括正确的三元组和错误的三元组)头部实体和尾部实体的相似度后,按照升序排序。所有正确三元组排序位置的平均值即为Mean rank。对于一个好的知识图谱表示来说,正确三元组的得分(即头部实体和尾部实体的关系函数值)会小于错误三元组的得分,排名会比较靠前。因此,Mean rank值越小,知识图谱表示向量越好,具体如下

$$\mathrm{MR}=\frac{1}{N_{\mathrm{T}}}\sum_{i=1}^{N_{\mathrm{T}}}\mathrm{rank}_i \tag{21}$$

式中:N_{T}表示正确三元组的个数;rank_i表示正确三元组的排名。

(2) Hit@10

上述排序中排名前10中所包含正确三元组的个数占正确三元组总数的比例即为Hit@10值。所以,Hit@10值越大,知识图谱表示向量越好,具体如下

$$\mathrm{Hit@10}=\frac{N_{\mathrm{T}}^{\mathrm{rank}\leqslant 10}}{N_{\mathrm{T}}}\times 100\% \tag{22}$$

式中$N_{\mathrm{T}}^{\mathrm{rank}\leqslant 10}$表示正确三元组中在排名前十的个数。

(3) 准确率

在预测任务中,给定样例集$D=\{(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)\}$,其中$y_i$是示例$x_i$的真实标记。一般通过对比模型预测结果$f(x)$与真实标记$y$的差异,来评估模型$f$的性能。

准确率(acc)是分类任务最常用的性能度量,即分类正确的样本数占样本总数的比例,可定义为

$$\mathrm{acc}=\frac{1}{m}\big(f(x_i)=y_i\big) \tag{23}$$

(4) 召回率

对于二分类问题,可将样例根据真实类别与模型预测类别划分为真正例(TP)、假正例(FP)、真反例(TN)、假反例(FN)四种情形。召回率R可定义为

$$R=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{24}$$

3.4 实验设计与结果分析

(1) 知识表示模型对比分析

首先,分析不同知识表示模型的性能,结果如表6、7所示。如表6所示,综合MR指标和Hit@10指标来看,TransH模型进行知识表示的效果最好。这说明TransH能更好地处理体检与检测值之间存在的“一对多”和“多对一”的复杂关系,弥补了TransE的不足。TransR模型虽然考虑到了这些复杂关系,但是体检指标与检测值之间只存在偏高、偏低等类似的关系,不同关系关注的是实体的相似属性,所以TransR模型进行知识表示的效果并不好。

表6 不同知识表示模型的MR和Hit@10

Table 6 MR and Hit@10 of different knowledge representation models

模型	MR	Hit@10/%
TransE	**623.0**	44.9
TransH	711.6	**47.9**
TransR	897.8	19.0

表7 不同知识表示模型的准确率和召回率

Table 7 Accuracy and recall rate of different knowledge representation models

模型	准确率/%	召回率/%
TransE-SAC	97.11	87.16
TH-SAC	97.18	87.32
TransR-SAC	**97.03**	**86.89**

此外,从表7中可以看出,TransH模型的表现均优于TransE模型和TransR模型,在准确率上分别提高了0.07%、0.15%,召回率上分别提高了0.16%、0.43%。这也进一步说明,针对本文中依据体检知识构建的三元组,TransH模型的表示方式更加合理,也使得预测模型的性能更优。

(2) 本文模型与其他模型对比分析

为了验证本文提出的TH-SAC模型在糖尿病预测任务上的优势,选取了一些相关糖尿病预测模型进行对比实验。TH-SAC模型是通过知识表示学习将体检数据表示成向量,采用深度学习的方法进行预测。首先选取在糖尿病预测任务上效果良好的机器学习方法以及深度神经网络(Deep neural network,DNN)进行比较,结果如表8所示。从表8可以看出本文提出的TH-SAC模型相较于机器学习中效果最好的随机森林方法,其准确率和召回率分别提升了0.81%和5.21%。这是因为在基于本文构建的融合知识表示的可解释深度学习方法中,模型架构更加“窄而深”,能更好地挖掘出体检数据中所包含的信息。相比单纯采用DNN,其准确率和召回率分别提升了6.97%和28.7%,说明本文通过知识表示学习将体检数据表示为向量的方法比单纯使用检测值效果更优。嵌入的外部知识不仅提高了模型的可解释性,也对模型的性能具有提升作用。

此外,TH-SAC模型中使用的分类器是融合了自注意力机制(Self-attention)[37]和卷积神经网络(Convolutional neural networks,CNN)[38]进行设计并实现。因此,本文还进行了与以下方法的对比实验:单独使用Self-Attention和CNN、结合Self-Attention和双向长短期记忆网络(Bi-directional long short-term memory,BiLSTM)[39]的方法,结果如表8所示。可以看出,与单一Self-Atttention、CNN、Self-Atttention-BiLSTM相比,SAC分类器在准确率和召回率方面都有更好的性能。这是因为分类器通过获取全局的信息和对局部特征的提取,比单独使用Self-Attention或CNN性能更优。此外,体检数据中并不存在时序信息,所以使用BiLSTM效果并不是很好。

表8　不同糖尿病预测模型的准确率和召回率

Table 8　Accuracy and recall rate of different diabetes prediction models

模型	准确率/%	召回率/%
逻辑回归(Logistic regression,LR)	90.29	49.9
支持向量机(Support vector machine,SVM)	90.59	51.60
朴素贝叶斯(Naive Bayes,NB)	87.48	53.94
随机森林(Random forest,RF)[40]	96.37	82.11
XGBoost[41]	92.42	61.64
深度神经网络(Deep neural network,DNN)	90.21	58.62
TH-Self-Atttention	96.05	86.15
TH-CNN	96.26	84.24
TH-Self-Atttention-BiLSTM	93.90	78.20
TH-SAC	**97.18**	**87.32**

(3) 知识表示与随机表示对比分析

为了验证融入外部体检知识的有效性,本文对体检指标实体和检测值实体的随机表示和知识表示进行对比,其中随机表示指对所有实体进行one-hot编码,然后与一个随机生成的矩阵相乘得到。对比结果如表9和图3、4所示。从表9可以看出,在预测性能上知识表示明显优于随机表示模型。这说明本文中通过体检指标和检测值之间的这种关系构建的实体向量发挥了良好的作用。此外,图3、4分别为两种模型在训练过程中前100批次的准确率和召回率,可以看出,结合了知识表示学习的模型训练时间更短,更快达到收敛状态。

表9　不同表示方式的准确率和召回率

Table 9　Accuracy and recall rates of different representations

模型	准确率/%	召回率/%
Random-SAC	96.72	86.76
TH-SAC	**97.18**	**87.32**

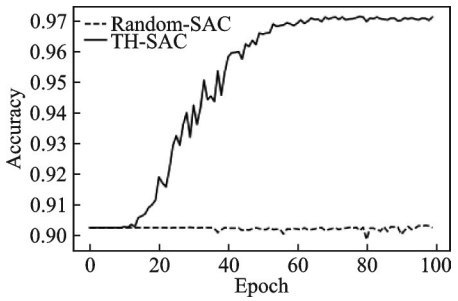

图3　模型训练过程中的准确率

Fig.3　Accuracy during model training

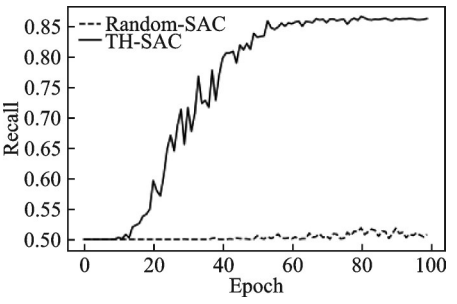

图4　模型训练过程中的召回率

Fig.4　Recall during model training

(4) 不同维度的表示向量比较分析

在知识表示学习的过程中,如果向量维度选择得过小或者过大,也会存在过拟合/欠拟合的风险。为了选取较优的表示向量维度,分别对比了200、256、300和512这4个维度,结果如图5、6所示。从图5、6可以看出,在较低的200维表示向量时因为所包含的信息不全面,其准确率和召回率相比其他的要低一些。但维度越高,模型参数越复杂,训练时间也越长。综合考虑准确率、召回率和模型参数复杂度,最终选取表示向量维度为256。

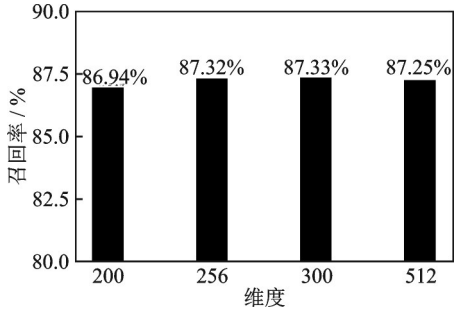

图5 不同维度的表示向量召回率

Fig.5 Representation vector recall rate at different dimensions

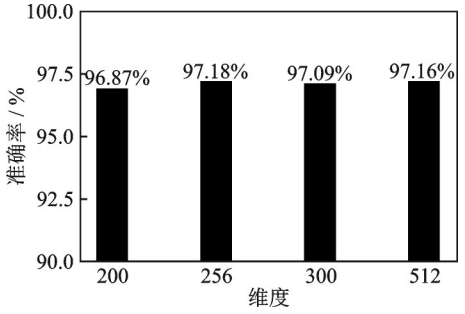

图6 不同维度的表示向量准确率

Fig.6 Accuracy of representation vectors at different dimensions

3.5 讨 论

表10展示了本文所提模型与不同深度学习模型异同点比较情况。本文融入知识表示向量的方法具有以下优势：

(1)提高了深度学习模型在疾病预测方面的性能。通常，数据表征的质量会影响深度学习模型在下游任务上的表现。本文通过知识表示学习将体检数据表示为向量形式，作为深度学习模型的输入，进行疾病预测。文中实验结果表明，融入该向量后，预测性能明显提升。

(2)提高深度学习模型的可解释性。本文基于体检指标正常参考范围，构建体检指标与测量值的关系图，并通过知识表示学习转换成向量形式，输入深度学习模型中。通过该方式，将医学专业知识嵌入深度学习中，增强了模型的可解释性。

表10 不同深度学习模型异同点比较

Table 10 Comparison of differences and similarities among different deep learning models

模型	是否嵌入知识	是否可解释
DNN	否	否
TH-Self-Atttention	是	否
TH-CNN	是	否
TH-Self-Atttention-BiLSTM	是	否
TH-SAC	是	是

4 结束语

针对传统人工智能方法在疾病预测领域应用缺乏可解释性的问题，本文提出了一种融入知识表示向量的可解释深度学习模型并应用于糖尿病预测。依据体检指标与检测值之间的关系，通过TransH模型，构建了体检知识实体的向量表示，进而得到患者体检数据的关系矩阵；然后通过构建的自注意力机制和卷积神经网络进行特征提取，从而设计并实现了一种面向糖尿病预测的可解释深度学习模型。实验与分析结果验证了引入知识表示向量后的深度学习模型的有效性和可解释性。该模型因为采用了外部体检知识，并且符合医学领域专业知识的诊断结果，所以具有良好的可解释性。但是，本文中所使用的体检知识并不全面，未考虑到体检指标正常范围与年龄和性别之间的关系，此外，本文模型仅使

用了体检数据,未考虑到患者的相关症状,与实际临床诊断存在差别。在下一步工作中,将引入上述关系对糖尿病预测模型进行改进,并构建计算机辅助诊断系统。

参考文献:

[1] 余凯,贾磊,陈雨强,等. 深度学习的昨天、今天和明天[J]. 计算机研究与发展, 2013, 50(9): 1799-1804.
YU Kai, JIA Lei, CHEN Yuqiang, et al. Deep learning: Yesterday, today, and tomorrow[J]. Journal of Computer Research and Development, 2013, 50(9): 1799-1804.

[2] BENZEGHIBA M, DE MORI R, DEROO O, et al. Automatic speech recognition and speech variability: A review[J]. Speech Communication, 2007, 49(10/11): 763-786.

[3] NATH S S, MISHRA G, KAR J, et al. A survey of image classification methods and techniques[C]//Proceedings of 2014 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT). [S.l.]: IEEE, 2014: 554-557.

[4] NADKARNI P M, OHNO-MACHADO L, CHAPMAN W W. Natural language processing: An introduction[J]. Journal of the American Medical Informatics Association, 2011, 18(5): 544-551.

[5] 陈珂锐,孟小峰. 机器学习的可解释性[J]. 计算机研究与发展, 2020, 57(9): 1971-1986.
CHEN Kerui, MENG Xiaofeng. Interpretation and understanding in machine learning[J]. Journal of Computer Research and Development, 2020, 57(9): 1971-1986.

[6] 成科扬,王宁,师文喜,等. 深度学习可解释性研究进展[J]. 计算机研究与发展, 2020, 57(6): 1208-1217.
CHENG Keyang, WANG Ning, SHI Wenxi, et al. Research advances in the interpretability of deep learning[J]. Journal of Computer Research and Development, 2020, 57(6): 1208-1217.

[7] SHEN D, WU G, SUK H. Deep learning in medical image analysis[J]. Annual Review of Biomedical Engineering, 2017, 19: 221-248.

[8] SHICKEL B, TIGHE P J, BIHORAC A, et al. Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis[J]. IEEE Journal of Biomedical and Health Informatics, 2017, 22(5): 1589-1604.

[9] RAJKOMAR A, OREN E, CHEN K, et al. Scalable and accurate deep learning with electronic health records[EB/OL]. (2018-05-11). https://doi.org/10.48550/arXiv.1801.07860.

[10] SENGUPTA S, SINGH A, LEOPOLD H A, et al. Ophthalmic diagnosis using deep learning with fundus images—A critical review[J]. Artificial Intelligence in Medicine, 2020, 102: 101758.

[11] SISODIA D, SISODIA D S. Prediction of diabetes using classification algorithms[J]. Procedia Computer Science, 2018, 132: 1578-1585.

[12] JO T, NHO K, SAYKIN A J. Deep learning in Alzheimer's disease: Diagnostic classification and prognostic prediction using neuroimaging data[J]. Frontiers in Aging Neuroscience, 2019. DOI: 10.3389/fnagi.2019.00220.

[13] TJOA E, GUAN C. A survey on explainable artificial intelligence (XAI): Toward medical XAI[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 32(11): 4793-4813.

[14] 纪守领,李进锋,杜天宇,等. 机器学习模型可解释性方法、应用与安全研究综述[J]. 计算机研究与发展, 2019, 56(10): 2071-2096.
JI Shouling, LI Jinfeng, DU Tianyu, et al. Survey on techniques, applications and security of machine learning interpretability[J]. Journal of Computer Research and Development, 2019, 56(10): 2071-2096.

[15] TONEKABONI S, JOSHI S, MCCRADDEN M D, et al. What clinicians want: Contextualizing explainable machine learning for clinical end use[EB/OL]. (2019-08-07). https://doi.org/10.48550/arXiv.1905.05134.

[16] LONDON A J. Artificial intelligence and black-box medical decisions: Accuracy versus explainability[J]. The Hastings Center Report, 2019, 49(1): 15-21.

[17] JI S, PAN S, CAMBRIA E, et al. A survey on knowledge graphs: Representation, acquisition, and applications[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 33(2): 494-514.

[18] 刘知远,孙茂松,林衍凯,等. 知识表示学习研究进展[J]. 计算机研究与发展, 2016, 53(2): 247-261.
LIU Zhiyuan, SUN Maosong, LIN Yankai, et al. Knowledge representation learning: A review[J]. Journal of Computer Research and Development, 2016, 53(2): 247-261.

[19] XIE Y, GAO G, CHEN X. Outlining the design space of explainable intelligent systems for medical diagnosis[EB/OL]. (2019-02-16). https://doi.org/10.48550/arXiv.1902.06019.

[20] HOLZINGER A, LANGS G, DENK H, et al. Causability and explainability of artificial intelligence in medicine[J]. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2019, 9(4): e1312.

[21] HAUFE S, MEINECKE F, GÖRGEN K, et al. On the interpretation of weight vectors of linear models in multivariate neuroimaging[J]. Neuroimage, 2014, 87: 96-110.

[22] CARUANA R, LOU Y, GEHRKE J, et al. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission[C]//Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. [S.l.]: ACM, 2015: 1721-1730.

[23] VAROL E, SOTIRAS A, ZENG K, et al. Generative discriminative models for multivariate inference and statistical mapping in medical imaging[C]//Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention. [S.l.]: Springer, 2018: 540-548.

[24] RIBEIRO M T, SINGH S, GUESTRIN C. "Why should I trust you?" explaining the predictions of any classifier[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. [S.l.]: ACM, 2016: 1135-1144.

[25] PEREIRA S, MEIER R, ALVES V, et al. Automatic brain tumor grading from MRI data using convolutional neural networks and quality assessment[M]//Understanding and Interpreting Machine Learning in Medical Image Computing Applications. Cham, Switzerland: Springer, 2018: 106-114.

[26] VAN MOLLE P, DE STROOPER M, VERBELEN T, et al. Visualizing convolutional neural networks to improve decision support for skin lesion classification[M]//Understanding and Interpreting Machine Learning in Medical Image Computing Applications. Cham, Switzerland: Springer, 2018: 115-123.

[27] BIFFI C, OKTAY O, TARRONI G, et al. Learning interpretable anatomical features through deep generative models: Application to ardiac remodeling[C]//Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention. [S.l.]: Springer, 2018: 464-471.

[28] KIM B, WATTENBERG M, GILMER J, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV)[C]//Proceedings of the 35th International Conference on Machine Learning. Stockholm, Sweden: [s.n.], 2018: 2668-2677.

[29] ZHANG Z, XIE Y, XING F, et al. MDNet: A semantically and visually interpretable medical image diagnosis network[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2017: 6428-6436.

[30] 侯梦薇,卫荣,陆亮,等. 知识图谱研究综述及其在医疗领域的应用[J]. 计算机研究与发展, 2018, 55(12): 2587-2599.
HOU Mengwei, WEI Rong, LU Liang, et al. Research review of knowledge graph and its application in medical domain[J]. Journal of Computer Research and Development, 2018, 55(12): 2587-2599.

[31] 刘勘,张雅荃. 基于医疗知识图谱的并发症辅助诊断[J]. 中文信息学报, 2020, 34(10): 85-93.
LIU Kan, ZHANG Yaquan. Medical knowledge graph based auxiliary diagnosis of complications[J]. Journal of Chinese Information Processing, 2020, 34(10): 85-93.

[32] 钱虹. 糖尿病的研究现状及进展[J]. 医学综述, 2017, 21(13): 2418-2420.
QIAN Hong. Current situation and progress of study on diabetes[J]. Medical Recapitulate, 2017, 21(13): 2418-2420.

[33] BORDES A, USUNIER N, GARCIA-DURAN A, et al. Translating embeddings for modeling multi-relational data[C]//

Proceedings of the 27th Conference on Neural Information Processing Systems (NIPS 2013). [S.l.]: [s.n.], 2013: 26-34.

[34] WANG Z, ZHANG J, FENG J, et al. Knowledge graph embedding by translating on hyperplanes[C]//Proceedings of AAAI Conference on Artificial Intelligence. [S.l.]: AAAI, 2014: 1112-1119.

[35] LIN Y, LIU Z, SUN M, et al. Learning entity and relation embeddings for knowledge graph completion[C]//Proceedings of the Twenty-ninth AAAI Conference on Artificial Intelligence. [S.l.]: AAAI, 2015: 2181-2187.

[36] 中华医学会糖尿病学分会. 中国2型糖尿病防治指南(2020年版)[J]. 中华内分泌代谢杂志, 2021, 37(4): 311-398.
Chinese Diabetes Society. Guideline for the prevention and treatment of type 2 diabetes mellitus in China (2020 edition)[J]. Chinese Journal of Endocrinology and Metabolism, 2021, 37(4): 311-398.

[37] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[EB/OL]. (2017-12-06). https://doi.org/10.48550/arXiv.1706.03762.

[38] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[39] HUANG Z, XU W, YU K. Bidirectional LSTM-CRF models for sequence tagging[EB/OL]. (2015-08-09). https://doi.org/10.48550/arXiv.1508.01991.

[40] HO T K. Random decision forests[C]//Proceedings of the 3rd International Conference on Document Analysis and Recognition. [S.l.]: IEEE, 1995, 1: 278-282.

[41] CHEN T, GUESTRIN C. Xgboost: A scalable tree boosting system[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. [S.l.]: ACM, 2016: 785-794.

作者简介:

徐鹤(1985-),男,博士,教授,研究方向:大数据、物联网技术、知识表示与推理等,E-mail: xuhe@njupt.edu.cn。

郑群力(1998-),男,硕士研究生,研究方向:大数据、知识表示与推理。

谢作玲(1970-),女,硕士,副主任医师,研究方向:糖尿病及其急性、慢性并发症、性腺疾病等的诊治。

程海涛(1986-),男,博士,讲师,研究方向:时空大数据、知识图谱、知识表示与推理。

李鹏(1979-),**通信作者**,男,博士,教授,研究方向:医学数据处理、物联网技术、网络安全等,E-mail: lipeng@njupt.edu.cn。

季一木(1978-),男,博士,教授,博士生导师,研究方向:高性能计算、大数据和人工智能。

(编辑:王静)