Hans Journal of Data Mining 数据挖掘, 2020, 10(1), 30-38
Published Online January 2020 in Hans. http://www.hanspub.org/journal/hjdm
https://doi.org/10.12677/hjdm.2020.101003

Hans汉斯

A Risk Prediction Model Based on Prior Medical Knowledge

Qian Lu[1], Min Liang[1], Ningning Li[1], Dong Lin[2], Yuchang Mo[1]

[1]Fujian Province University Key Laboratory of Computational Science, School of Mathematical Sciences, Huaqiao University, Quanzhou Fujian
[2]College of Acupuncture, Fujian University of Traditional Chinese Medicine, Fuzhou Fujian
Email: yuchangmo@sina.com

Received: Dec. 19th, 2019; accepted: Dec. 27th, 2019; published: Jan. 3rd, 2020

Abstract

The task of predicting potential disease risks through electronic health records is a hot research topic in the medical field in recent years. With the rapid development of machine learning research and application, the classical machine learning model is gradually unable to meet the growing data volume and complex data analysis needs, while the neural network model in deep learning can solve the problem that machine learning cannot or is difficult to solve. There is no explicit consideration of prior medical knowledge in existing disease prediction work. We propose a new, generic framework called the risk prediction task, which successfully applies discrete prior medical knowledge to all the most advanced prediction models using posterior regularization techniques. In this paper, the convolution neural network is used to establish the risk prediction model, and prior medical knowledge is added. Gradient descent algorithm was used for optimization. Experiments prove that this model can effectively improve the accuracy of risk prediction compared with the convolution neural network in traditional deep learning.

Keywords

Electronic Health Record, Deep Learning, Prior Medical Knowledge

基于先验医学知识的风险预测模型

陆　迁[1]，梁　敏[1]，李宁宁[1]，林　栋[2]，莫毓昌[1]

[1]华侨大学，数学科学学院，计算科学福建省高校重点实验室，福建　泉州
[2]福建中医药大学，针灸学院，福建　福州
Email: yuchangmo@sina.com

收稿日期：2019年12月19日；录用日期：2019年12月27日；发布日期：2020年1月3日

文章引用: 陆迁, 梁敏, 李宁宁, 林栋, 莫毓昌. 基于先验医学知识的风险预测模型[J]. 数据挖掘, 2020, 10(1): 30-38.
DOI: 10.12677/hjdm.2020.101003

摘　要

通过电子健康记录预测潜在疾病风险任务是近年来医疗领域的研究热点。随着机器学习研究与应用的快速发展，经典机器学习模型渐渐无法满足日益增长的数据量和复杂的数据分析需求，而深度学习中神经网络模型可以解决机器学习无法解决或难以解决的问题。现有疾病预测工作中没有对先验医学知识的明确考虑。本文提出了一种新的、通用的框架，称为风险预测任务，它可以使用后验正则化技术成功地将离散的先验医学知识应用到所有最先进的预测模型中。本文以卷积神经网络建立风险预测模型，并加入先验医学知识，以梯度下降算法进行优化。实验证明，与传统深度学习中的卷积神经网络相比该模型能有效提高风险预测的准确率。

关键词

电子健康记录，深度学习，先验医学知识

1. 引言

先验医学知识在医疗领域起着重要作用。当一个病人去看医生时，医生首先会检查病人目前的症状，然后会仔细检查病史，如药物、吸烟史、饮酒史、家族史等，这些都是疾病的危险因素。然而，将先验医学知识直接应用于 EHR 数据是极其困难的。一方面，医学知识具有随意性或异质性。一些疾病可能与年龄有关(持续价值)，而另一些则是由吸烟或饮酒等习惯引起(分类价值)。另一方面，几乎所有的医学知识都用规则来表示。因此，将离散的任意医疗规则转化为连续的真实价值是一个发人深省的问题，即使我们能够得到先验医学知识的实值表示，如何将先验医学知识与预测模型合理结合仍然是一个挑战。

本文提出框架 PRIME (Prior Medical)，使用后验正则化技术成功地将离散的先验医学知识应用到所有最先进的预测模型中。PRIME 将先验医学知识建模为后验正则化，并使用对数线性模型学习期望的后验分布，它能够区分不同先验知识对风险预测的重要性。在三个医疗数据集上的实验结果表明，提出的 PRIME 框架对于风险预测任务是有效的。

2. 相关工作

2005 年，美国医疗卫生信息与管理系统协会在年会上提出了电子健康概念为：电子健康档案是深度数字化的、上下文关联的病人终身医疗记录，从时间跨度上覆盖个人从生到死整个生命周期，从内容上强调个人信息[1]。目前我国 EHR 技术正处于起始阶段。在过去的 10 年里，美国医院采用电子健康档案(EHR)系统的数量激增，部分原因是 2009 年《卫生信息技术促进经济和临床卫生(HITECH)法案》(health Information Technology for Economic and Clinical health, Act of 2009)为医院和医生采用电子健康档案系统提供了 300 亿美元的激励[2]。根据 2016 年美国国家卫生信息技术协调员办公室的报告，近 84%的医院至少采用了基本的 EHR 系统，比 2008 年增加了 9 倍[3]。此外，医生采用基本的和经过认证的电子病历的比例从 42%增加了一倍多，达到 87%。虽然最初的设计是为了从操作的角度提高医疗效率，但许多研究已经发现临床信息学应用的更多用途。特别是 EHR 系统中包含的患者数据已经被用于医学概念提取、

患者轨迹建模、疾病推断、临床决策支持系统等任务。我国电子健康记录处于发展阶段，已广泛应用于医疗机构中帮助医疗人员处理各种医疗信息[4]。

用于分析丰富 EHR 数据的技术大多基于传统的机器学习和统计技术，如 logistic 回归[5]、支持向量机(SVM) [6]、随机森林[7]。近年来，深度学习技术通过深层次特征构造和以有效方式捕获数据中的远程依赖性在许多领域都取得了巨大成功。深度学习是机器学习的重要分支，它是在机器学习基础上发展起来的，深度学习中很多新技术在继承机器学习优点、克服其不足的基础上发展起来的，如循环神经网络(recurrent neural network, RNN)及其变体门控循环单元(gated recurrent unit, GRU)，相比于传统医学研究所使用的 logistic 回归模型，对于腹膜透析临床预后预测具有更佳效果，可能有助于医生早期干预，提高医疗质量，具有很强的临床应用价值[8]。随着深度学习方法的普及和越来越多的患者数据的增加，对于临床信息学任务应用到深度学习 EHR 数据的出版物数量也有所增加，这会比传统方法产生更好的性能，而且需要进行预处理和特征工程的时间会更少[9]。

3. 基本概念

3.1. EHR 数据描述

EHR 数据包括患者按时间排序的访问记录。设 $\mathcal{P}$ 表示所有患者的集合，$|\mathcal{P}|$ 是 EHR 数据中患者的数量。对每个患者 $p\in\mathcal{P}$，有 T_p 时间访问顺序 $V_1^{(p)},V_2^{(p)},\cdots,V_{T_p}^{(p)}$。设 $|C|$ 为唯一诊断代码的数量，令 $C=\left\{c_1,c_2,\cdots,c_{|C|}\right\}$ 为所有诊断代码的集合。每一个访问 $V_t^{(p)}$ 包含诊断代码 c_i。预测模型的输入是第 p 位患者的 EHR 记录，定义为 $X^{(p)}=\left\{x_t^{(p)}\right\}_{t=1}^{T_p}\in\mathbb{R}^{T_p\times|C|}$。为简明起见，我们删除上标签$(p)$。

3.2. 卷积神经网络(CNN)

由于输入 $X^{(p)}$ 过于稀疏且具有高维性，因此需要学习它的低维和有意义的嵌入。将每个输入 x_t 嵌入到访问状态中 $v_t\in\mathbb{R}^k$ 中：

$$v_t=W_v x_t+b_v \tag{1}$$

其中，$W_v\in\mathbb{R}^{k\times|C|}$ 和 $b_v\in\mathbb{R}^k$ 是需要学习的参数，$k=256$ 是隐藏层大小。

对 $V^{(p)}=\left\{v_t^{(p)}\right\}_{t=1}^{T_p}\in\mathbb{R}^{T_p\times k}$ 应用卷积运算，使用具有不同窗口大小的 $m=100$ 个滤波器组合。设 l 表示时间窗口的大小，然后 $v_{t:t+l-1}$ 表示从 v_t 到 v_{t+l-1} 的 l 次的连接。一个滤波器 $W_f\in\mathbb{R}^{l\times k}$ 应用于 l 次访问的窗口来产生一个新的特征 $f_t\in\mathbb{R}$，使用线性整流函数 ReLU 激活功能：$f_t=\mathrm{ReLU}\left(W_f v_{t:t+l-1}+b_f\right)$，其中 b_f 是偏置项，以及 $\mathrm{ReLU}(f)=\max(f,0)$。这个过滤器适用于整个描述 $\left\{v_{1:l},v_{2:l+1},\cdots,v_{T_p-l+1:T_p}\right\}$ 中的每个可能的访问窗口，以生成如下的一个特征：$f=\left[f_1,f_2,\cdots,f_{T_p-l+1}\right]$。为了获得最重要的特征，在特征上使用了最大池化技术，即 $\hat{f}=\max(f)$。使用 s 个不同窗口大小的 m 个过滤器，通过将所有提取的特征连接起来，就可以得到第 p 位患者的最终向量表示，如 $z^{(p)}\in\mathbb{R}^{ms}$。最后，应用一个全连接的 softmax 层以产生预测概率，如下：

$$\hat{y}_p=\mathrm{softmax}\left(W_y z^{(p)}+b_y\right) \tag{2}$$

其中，$W_y\in\mathbb{R}^{N\times ms}$ 和 $b_y\in\mathbb{R}^N$ 是可学习的参数，N 是目标疾病的数量。在这次实验中，我们专注于二元预测任务，即 $N=2$。设 θ 为卷积神经网络中所有参数的合集，则预测概率 $\hat{y}_p$ 也可以由后验分布 $P\left(y_p\mid X^{(p)};\theta\right)$ 表示，其中 y_p 是真实值。

3.3. 随机梯度下降法

在梯度法中，函数的取值从当前位置沿着梯度方向前进一定距离，然后在新的地方重新求梯度，再沿着新梯度方向前进，如此反复，不断地沿梯度方向前进。像这样，通过不断地沿梯度方向前进，逐渐减小函数值的过程就是梯度法。随机梯度下降法步骤如下：

步骤 1：从训练数据中随机选出一部分数据，这部分数据称为 mini-batch。目标是减小 mini-batch 的损失函数的值。

步骤 2：为了减小 mini-batch 的损失函数值，需要求出各个权重参数的梯度，梯度表示损失函数的值减小最多的方向。

步骤 3：将权重参数沿梯度方向进行微小更新。

步骤 4：重复以上 3 步骤，直到损失函数收敛。

这个方法通过梯度下降发更新参数，不过因为这里使用的数据是随机选择的 min batch 数据，所以又称为随机梯度下降法。

4. 基于先验医学知识的风险预测模型

在本节中，我们描述先验医学知识中 5 种风险因素的数学建模，然后在对数线性模型的基础上建立风险预测模型。

4.1. 先验医学知识

后验正则化[10]是通过对潜在变量的后验分布进行结构约束而引入间接监督(即先验医学知识)的方法。后验正则化的目标是使用先验知识来限制模型后验的空间，以引导模型朝向期望的参数分布。设 $q\left(y_p\right)$ 表示患者 p 的期望分布。$\mathcal{Q}$ 是后验信息约束的集合，定义为：

$$\mathcal{Q}=\left\{q\left(\boldsymbol{y}_p\right):\mathbb{E}_q\left[\boldsymbol{\phi}\left(\boldsymbol{X}^{(p)},\boldsymbol{y}_p\right)\right]\le\boldsymbol{b}\right\} \tag{3}$$

其中 $\boldsymbol{\phi}\left(\boldsymbol{X}^{(p)},\boldsymbol{y}_p\right)$ 是约束特征的集合，$\boldsymbol{b}$ 是约束特征期望的(已知)界限。

由于不同的疾病具有不同的风险因素，在医学领域，医学将风险因素分为五大类：患者特征，潜在疾病，疾病持续时间，遗传学和家族史。

4.1.1. 患者种族和年龄特征

给出患者 p 的人口统计信息 $g^{(p)}=\left[g_e^{(p)},g_a^{(p)}\right]$ 和相应的标签 y_p，关于种族的特征被定义为如下：

$$\phi_e\left(X^{(p)},y_p\right)=\begin{cases}1 & g_e^{(p)}\in\mathcal{E}\\ 0 & \text{其他}\end{cases} \tag{4}$$

其中 $\mathcal{E}$ 定义为与预测相关的种族集合。因为 ϕ_e 的值为 1 或 0，因此种族向量 $\boldsymbol{\phi}_e=[1,1]$ 或 $[0,0]$。为了模拟对案例和控制的不同重要性，引入约束特征种族的置信度向量 γ_e。

对于大多数疾病，随着患者年龄的增长，风险会增加。因此，引用常用的逻辑函数来模拟年龄的影响如下：

$$\phi_a\left(X^{(p)},y_p;w_y^{(a)}\right)=\frac{1}{1+\exp\left\{-w_y^{(a)}\left(g_a^{(p)}-\psi\right)\right\}} \tag{5}$$

其中 $w_y^{(a)}\in\mathbb{R}$ 是疾病特异性参数，用于模拟年龄对风险预测的影响。如果疾病对年龄不敏感，则 $w_y^{(a)}\to\infty$。

ψ 是预定义的标量。在本文，使用年龄组而不是患者的真实年龄，设置 $\psi=9$ (即，年龄是 40 到 45 岁)。年龄特征向量 $\boldsymbol{\phi}_a=\left[\phi_a\left(w_0^{(a)}\right),\phi_a\left(w_1^{(a)}\right)\right]$，$\boldsymbol{\gamma}_a$ 是对应的置信向量。

4.1.2. 潜在疾病约束特征

对于潜在疾病约束特征，首先得到每个风险预测任务的潜在疾病，记为 $\mathcal{U}$，然后计算这些潜在疾病在 p 患者就诊的频率，用 u_p 表示。原因是频率越高，风险越高。另外，不同潜在疾病的对于最终疾病预测的影响是不同的。因此，潜在疾病的约束特征设计如下：

$$\phi_u\left(X^{(p)},y_p;w_y^{(u)}\right)=\begin{cases}\dfrac{1}{1+\mathrm{e}^{-w_y^{(u)}\cdot u_p}} & sum\left(u_p\right)>0\\ 0 & sum\left(u_p\right)=0\end{cases} \tag{6}$$

其中，$w_y^{(u)}\in\mathbb{R}^{|\mathcal{U}|}$ 是代表不同潜在疾病的不同影响的倾斜参数，$|\mathcal{U}|$ 是潜在疾病的数量，$\mathrm{sum}\left(u_p\right)$ 是 u_p 的总和。潜在疾病的向量是 $\boldsymbol{\phi}_u=\left[\phi_u\left(w_0^{(u)}\right),\phi_u\left(w_1^{(u)}\right)\right]$，它的置信度向量是 $\boldsymbol{\gamma}_u$。

4.1.3. 潜在疾病持续时间特征

为了获得潜在疾病的持续时间，首先从患者 p 的就诊记录中找到某一潜在疾病 d 的起始时间 $t_d^{(p)}$，然后使用 $T_p-t_d^{(p)}$ 计算持续时间。最后，疾病的持续时间记为 d_p。基于 d_p，疾病持续时间的约束特征定义如下：

$$\phi_d\left(X^{(p)},y_p;w_y^{(d)}\right)=\begin{cases}\dfrac{1}{1+\mathrm{e}^{-w_y^{(d)}\cdot d_p}} & sum\left(d_p\right)>0\\ 0 & sum\left(d_p\right)=0\end{cases} \tag{7}$$

其中，$w_y^{(d)}\in\mathbb{R}^{|\mathcal{U}|}$ 类似于 $w_y^{(u)}$ 来模拟潜在疾病中的差异，以及 $\boldsymbol{\phi}_d=\left[\phi_d\left(w_0^{(d)}\right),\phi_d\left(w_1^{(d)}\right)\right]$ 与置信向量 $\boldsymbol{\gamma}_d$。

4.1.4. 遗传病特征约束

对于遗传病，首先收集了一组与目标疾病相关的遗传疾病 $\mathcal{G}$。设 $C^{(p)}$ 表示患者 p 访问 $X^{(p)}$ 中的所有诊断代码。当 $C^{(p)}$ 和 $\mathcal{G}$ 的交集不为空，约束特征值为 1。正式公式如下：

$$\phi_g\left(X^{(p)},y_p\right)=\begin{cases}1 & C^{(p)}\cap\mathcal{G}\neq\varnothing\\ 0 & \text{其他}\end{cases} \tag{8}$$

与种族约束特征相似，ϕ_g 的值为 1 或 0。因此，$\phi_g=[1,1]$ 或 $[0,0]$，γ_g 为置信向量。

4.1.5. 家族史特征约束

一些疾病与整个家庭的疾病史有关，例如慢性肾病。我们收集了一系列家族史疾病 $\mathcal{H}$，然后提供如下的约束特征函数：

$$\phi_h\left(X^{(p)},y_p\right)=\begin{cases}1 & C^{(p)}\cap\mathcal{H}\neq\varnothing\\ 0 & \text{其他}\end{cases} \tag{9}$$

其中，$\phi_h=[1,1]$ 或 $[0,0]$，置信向量为 γ_h。

4.2. 分析预测模型

使用前述五种风险因素的权重组合进行预测，定义 $\boldsymbol{\Gamma}\cdot\boldsymbol{\phi}\left(\boldsymbol{X}^{(p)},y_p;\boldsymbol{\mathcal{W}}\right)$：

$$\boldsymbol{\Gamma}\cdot\boldsymbol{\phi}\left(\boldsymbol{X}^{(p)},y_p;\boldsymbol{\mathcal{W}}\right)=\boldsymbol{\gamma}_e\odot\boldsymbol{\phi}_e+\boldsymbol{\gamma}_a\odot\boldsymbol{\phi}_a+\boldsymbol{\gamma}_u\odot\boldsymbol{\phi}_u+\boldsymbol{\gamma}_d\odot\boldsymbol{\phi}_d+\boldsymbol{\gamma}_g\odot\boldsymbol{\phi}_g+\boldsymbol{\gamma}_h\odot\boldsymbol{\phi}_h \tag{10}$$

其中，$\boldsymbol{\Gamma}$ 是根据先验医学知识不同的约束特征类别的可学习置信矩阵，$\mathcal{W}$ 为参数集。

使用对数线性模型学习先验医学知识编码的期望分布 $\tilde{y}_p$，定义如下：

$$\tilde{y}_p = \mathcal{Q}\left(y_p \mid X^{(p)}; \boldsymbol{\Gamma}, \mathcal{W}\right) = \frac{\exp\left\{\Gamma \cdot \phi\left(X^{(p)}, y_p; \mathcal{W}\right)\right\}}{\sum_{y'_p} \exp\left\{\Gamma \cdot \phi\left(X^{(p)}, y'_p; \mathcal{W}\right)\right\}} \tag{11}$$

用交叉熵损失作为目标函 $\mathcal{J}(\boldsymbol{\theta}, \boldsymbol{\Gamma}, \mathcal{W})$，其中 $\mathcal{L}(\boldsymbol{\theta})$ 为真实值 y_p 和预测值 $\hat{y}_p$ 之间的交叉熵平均值，$\mathcal{L}'(\boldsymbol{\Gamma}, \mathcal{W})$ 真实值 y_p 和期望分布 $\tilde{y}_p$ 之间的交叉熵平均值，α, β 为用于平衡模型之间损失的超参数，$KL(\cdot \| \cdot)$ 用于测量期望分布和预测值之间差异的 Kullback-Leibler 散度。预测模型如图 1 所示。

$$\mathcal{J}(\boldsymbol{\theta}, \boldsymbol{\Gamma}, \mathcal{W}) = \mathcal{L}(\boldsymbol{\theta}) + \beta \mathcal{L}'(\boldsymbol{\Gamma}, \mathcal{W}) + \alpha \frac{1}{|\mathcal{P}|} \sum_{p=1}^{|\mathcal{P}|} \min_{q \in \mathcal{Q}} KL\left(\tilde{y}_p \| P\left(y_p \mid X^{(p)}; \theta\right)\right) \tag{12}$$

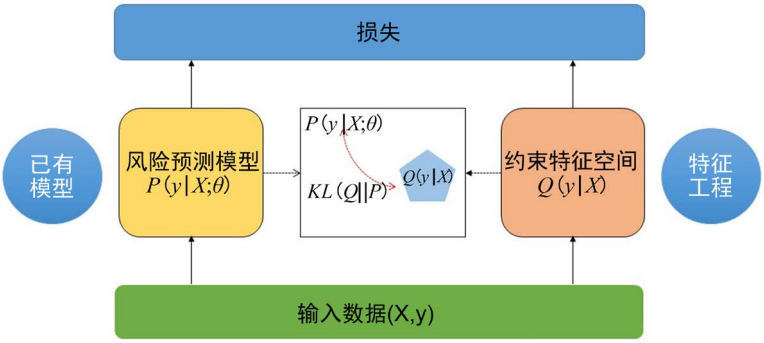

Figure 1. Risk prediction model framework
图 1. 风险预测模型框架

5. 实证研究

前文以卷积神经网络为基本模型，加入先验医学知识约束，构造了风险预测模型。为了验证所提出的风险预测框架的性能，本节以卷积神经网络为基本预测模型，加入先验医学知识约束，实现基本预测模型与提出的预测模型，并进行对比。

5.1. 参数设置

实验设置诊断代码总数 $|C| = 1016$，$k = 256$，即隐藏层为 256，超参数 $\alpha = \beta = 0.01$。对于 CNN，设置过滤器窗口的大小为[2, 3, 4, 5]，其中滤波器个数为 $m = 100$，使用正则化(系数为 0.001 的范数)和 dropout (退出率为 0.5)抑制过拟合。

5.2. 数据集

对于训练模型，实验使用 Adadelta，mini-batch 大小为 50。我们以 0.75:0.10:0.15 的比例将数据集随机划分为训练、验证和测试集。训练集用于模型拟合的数据样本，验证集是模型训练过程中单独留出的样本集，它可以用于调整模型的超参数和用于对模型的能力进行初步评估，测试集用来评估模最终模型的泛化能力。重复程序 10 次，并报告最优性能。

5.3. 实验结果

模型在 python3.6.5 上运行，参数设置如表 1 所示。

Table 1. Parameter settings for the prediction model in python
表 1. 预测模型在 python 中的参数设置

参数名称	参数值	参数名称	参数值
beta	0.01	visit_size	256
alpha	0.01	log_eps	1e−08
base_age	9.0	n_epoch	10
n_underlying	5	L2_reg	0.001
n_features	5	dropout_rate	0.5
use_gpu	False	batch_size	10
cuda_id	0	filter_hs	[2, 3, 4, 5]

实验结果如表 2 所示。mean_cost 为训练集损失平均值，validate_cnn_loss、test_cnn_loss 为真实值 y_p 和预测值 $\hat{y}_p$ 之间的交叉熵平均值，validate_loss、test_loss 为加入了先验医学知识损失的目标损失函数，对加入损失一起进行训练，由式(12)可知 validate_loss、test_loss 均大于未加入先验医学知识的交叉熵损失。由结果可知在第九次迭代时验证集损失 validate_loss 最小，获得最优结果，训练集损失为 0.551314，测试集损失为 0.641656，验证集损失为 0.533692。Loss1、2 两种情况下为验证集与训练集损失之差，由表可知 Loss2 小于 Loss1，则加入先验医学知识的预测模型优于卷积神经网络。

Table 2. The result
表 2. 实验结果

epoch	mean_cost	validate_cnn_loss	test_cnn_loss	validate_loss	test_loss	Loss1	Loss2
0	0.685308	0.629910	0.674288	0.633043	0.677421	0.055398	0.052265
1	0.646248	0.601331	0.665307	0.604463	0.668440	0.044917	0.041785
2	0.613947	0.560590	0.654358	0.563723	0.657491	0.053357	0.050224
3	0.657491	0.560678	0.644346	0.563810	0.647478	0.096813	0.093681
4	0.595203	0.560958	0.673584	0.564091	0.676716	0.034245	0.031112
5	0.596476	0.551949	0.629526	0.555081	0.632659	0.044527	0.041395
6	0.590181	0.559197	0.630622	0.562330	0.633755	0.030984	0.027851
7	0.586710	0.553262	0.657255	0.556394	0.660387	0.033448	0.030316
8	0.574634	0.561785	0.604853	0.564917	0.607986	0.012849	0.009717
9	**0.551314**	0.530559	0.638523	**0.533692**	**0.641656**	0.020755	0.017622

使用已有模型与提出的 PRIME 模型进行实验对比。输入数据为每次就诊出现的所有诊断代码的频率。使用以下用于分类方法的传统基线模型：

1) logistic 回归(LR)；2) 支持向量机(SVM)；3) 随机森林(RF)。

使用以下用于深度学习的基线模型：

1) GRU 神经网络；2) 长短期记忆网络(LSTM)；3) RETAIN；4) 卷积神经网络(CNN)。对于 GRU，LSTM 和 RETAIN，潜在表征大小设为 256。对于 CNN，设置过滤器窗口(l)的大小为 2 到 5，其中过滤器映射为 $s = 100$。

使用以下 PRIME 模型：

1) 使用 LSTM 为基本预测模型且加入先验医学知识的 PRIME_r，设置 $\alpha = \beta = 0.01$；

2) 使用 CNN 作为基本预测模型且加入先验医学知识的 PRIME_c，设置 $\alpha = 0.01, \beta = 0.1$；

3) 使用 LSTM 为基本预测模型且不加入先验医学知识的 PRIME_{r^-}，设置 $\alpha = \beta = 0.01$；

4) 使用 CNN 作为基本预测模型且不加入先验医学知识的 PRIME_{c^-}，设置 $\alpha = 0.01, \beta = 0.1$。

与已有模型对比结果如表 3 所示。在心力衰竭数据集上，传统的 LR、RF 和 SVM 方法的整体性能都比基于深度学习的方法差。这说明采用深度学习技术对高维稀疏的 EHR 数据进行建模对于风险预测任务是有效的。在四个基于深度学习的基线中，GRU 和 LSTM 的表现优于 RETAIN 和 CNN。由于 RETAIN 采用了注意机制，因此培训 RETAIN 需要大量的 EHR 数据。心力衰竭数据集的大小相对较小，因此 RETAIN 的性能较 GRU 和 LSTM 差。CNN 的优势在于捕捉当地时间的重要特征。然而，心力衰竭是一种慢性疾病，需要捕捉疾病演化的长期特征。基于 RNN 的模型可以正确识别心衰数据集上的这些特征，这使得与 CNN 相比性能更好。对于提出的四种方法，PRIME_r 取得了最好的性能。我们可以看到，和的性能都优于基本的预测模型 LSTM。同样，所有这些措施的值 PRIME_c 和 PRIME_{c^-} 高于在 CNN 中的值。这些观察结果有力地证实了先验医学知识可以帮助预测模型提高性能。

在 COPD 数据集中，RETAIN 的性能优于 GRU 和 LSTM，说明在所有基线中，注意力机制开始发挥作用，CNN 的性能最好。即使对拟议中的 PRIME_c 和 PRIME_{c^-}，所有的测量值都小于 CNN。原因在于，与某些疾病不同，COPD 有明确的病因，这与吸烟直接相关。CNN 具有出色的能力来捕捉这些局部的重要特征，即的诊断代码，有关吸烟在访问。因此，与其他方法相比，它取得了更好的性能。然而，在使用后验正则化整合先验医学知识后，即与 CNN 相比，PRIME_c 提出的方法有了显著的改进。这再次证实了考虑先前的医学知识对风险预测任务是有效的。

由于肾脏疾病患者的特点非常明确，传统的分类方法 RF 可以达到与深度学习相似的性能。即使在简单的数据集上，结合先前的医学知识仍然可以提高预测性能。在肾脏疾病数据集上，我们也观察到基本模型 LSTM 的性能与所提出的 PRIME_r 的性能相当。这是因为我们不调整最佳超参数 α 和 β。这两个参数对数据集非常敏感。尽管如此，在肾病数据集上，提议的 PRIME_c 优于其他方法。

Table 3. The results compared with the existing models
表 3. 与已有模型对比结果

Model		Heart Failure			COPD			Kidney Disease		
		AUROC	F1Score	Accuracy	AUROC	F1Score	Accuracy	AUROC	F1Score	Accuracy
Traditional Classification	LR	0.8810	0.8383	0.9048	0.8940	0.8559	0.9206	0.9147	0.8922	0.9335
	RF	0.8755	0.8444	0.9173	0.8801	0.8478	0.9202	0.9235	0.9145	0.9491
	SVM	0.8424	0.7734	0.8590	0.8400	0.7711	0.8715	0.8940	0.8545	0.9067
Deep Learning	GRU	0.9047	0.8854	0.9357	0.9014	0.8772	0.9349	0.9263	0.9146	0.9485
	RETAIN	0.8913	0.8661	0.9251	0.9110	0.8925	0.9431	0.9225	0.9133	0.9485
	LSTM	0.9034	0.8827	0.9339	0.9041	0.8812	0.9370	0.9267	0.9164	0.9498
	CNN	0.8994	0.8712	0.9260	0.9181	0.8968	0.9444	0.9284	0.9161	0.9491
This Work	PRIME_c	0.9059	0.8881	0.9374	0.9084	0.8859	0.9399	0.9258	0.9107	0.9455
	PRIME_{c^-}	0.8944	0.8709	0.9278	0.9204	0.9005	0.9464	0.9331	0.9201	0.9511
	PRIME_r	0.9126	0.8955	0.9410	0.9052	0.8868	0.9403	0.9276	0.9118	0.9459
	PRIME_{r^-}	0.9070	0.8788	0.9295	0.9211	0.9014	0.9468	0.9362	0.9236	0.9530

6. 总结

电子健康档案包含大量纵向、时间戳的临床数据，用机器学习算法处理此类数据通常需要将其转换为表格格式。本研究提出了一种基于子序列的时间序列符号化表示方法。该方法允许直接应用任何标准机器学习算法，同时与基于单一表示的方法相比，它在一定程度上能够获取时序信息，因而显著地提高了预测性能。

基金项目

国家自然科学基金项目(61572442)；福建省高校创新团队发展计划，福建省研究生导师团队，泉州市高层次人才团队项目(2017ZT012)；华侨大学研究生科研创新基金资助项目。

参考文献

[1] Lin, L. (2007) Analysis of Electronic Medical Record and Related Concepts. *China Medical Record*, No. 4, 40-41.

[2] Birkhead, G.S., Klompas, M. and Shah, N.R. (2015) Uses of Electronic Health Records for Public Health Surveillance to Advance Public Health. *Annual Review of Public Health*, **36**, 345-359. https://doi.org/10.1146/annurev-publhealth-031914-122747

[3] The Office of the National Coordinator for Health Information Technology (2016) Adoption of Electronic Health Record Systems among U.S. Non-Federal Acute Care Hospitals: 2008-2015.

[4] 刘艺聪, 张婷婷, 张紫君, 等. 标准化护理语言在电子健康记录中的应用现状[J]. 中国实用护理杂志, 2019, 35(9): 717-721.

[5] Li, S.R. (2014) A Retrospective Analysis of the Prevalence and Risk Factors of CKD in Type 2 Diabetes Mellitus Patients Based on Minhang District Electronic Health Record (EHR) Platform. Fudan University, Shanghai.

[6] Ferrão, J.C., Janela, F., Oliveira, M.D., *et al*. (2013) Using Structured EHR Data and SVM to Support ICD-9-CM Coding. *IEEE International Conference on Healthcare Informatics*, Philadelphia, 9-11 September 2013, 511-516. https://doi.org/10.1109/ICHI.2013.79

[7] Karlsson, I. and Bostrom, H. (2015) Handling Sparsity with Random Forests When Predicting Adverse Drug Events from Electronic Health Records. *IEEE International Conference on Healthcare Informatics*, Verona, 15-17 September 2014, 17-22. https://doi.org/10.1109/ICHI.2014.10

[8] 唐雯, 高峻逸, 马辛宇, 等. 循环神经网络模型在腹膜透析临床预后预测中的初步应用[J]. 北京大学学报(医学版), 2019(3): 602-608.

[9] Shickel, B., Tighe, P.J., Bihorac, A., *et al*. (2017) Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. *IEEE Journal of Biomedical & Health Informatics*, **22**, 1589-1604. https://doi.org/10.1109/JBHI.2017.2767063

[10] Ganchev, K., Graça, J. Gillenwater, J., *et al*. (2010) Posterior Regularization for Structured Latent Variable Models. *Journal of Machine Learning Research*, **11**, 2001-2049.