基于深度学习的异构时序事件患者数据表示学习框架

刘卢琛，沈剑豪，张铭，王子昌，李浩然，刘泽群
北京大学信息科学技术学院，北京 100871

摘要
患者数据的表示学习可以将患者历史信息综合表达为一个向量，用于预测未来可能发生的疾病。患者的历史记录可以被建模为多来源数据构成的采样频率差异很大、包含非线性时序关系的异构时序事件。提出了一个新的异构事件长短期记忆表示学习框架，用于学习患者异构时序事件的联合表征。异构事件长短期记忆模型加入了一个可以控制事件访问频率的门，以对不同事件的不规则采样频率建模，同时抓住事件中的复杂时序依赖关系。真实临床数据的实验表明，该方法可以在一系列先进模型的基础上，提升死亡预测和异常实验结果预测的准确度。

关键词
电子病历；患者数据表示学习；异构时序事件；深度学习

中图分类号：TP183 文献标识码：A doi: 10.11959/j.issn.2096-0271.2019003

Deep learning based patient representation learning framework of heterogeneous temporal events data

LIU Luchen, SHEN Jianhao, ZHANG Ming, WANG Zichang, LI Haoran, LIU Zequn
School of Electronics Engineering and Computer and Computer Science, Peking University, Beijing 100871, China

Abstract
Patient representation embeds patients' longitude records from multiple sources into continuous low-dimension vectors, which can be used to predict whether a disease will happen in the future. However, the problem is very challenging since patients' history records contain multiple heterogeneous temporal events. The visiting patterns of different types of events vary significantly, and there exist complex nonlinear relationships between different events. A novel model for learning the joint representation of heterogeneous temporal events was proposed. The model adds a new gate to control the visiting rates of different events which effectively models the irregular patterns of different events and their nonlinear correlations. Experiment results with real-world clinical data on the tasks of predicting death and abnormal lab tests prove the effectiveness of the proposed approach over competitive baselines.

Key words
electronic health record, patient representation learning, heterogeneous temporal events, deep learning

1 引言

电子病历的大量积累，为机器学习和数据挖掘的研究者们提供了很好的数据基础，以此服务于辅助诊断和智慧医疗，提供智能化的辅助诊疗服务。智慧医疗中一个重要的核心研究问题是患者数据的表示学习，它可以综合提炼、挖掘丰富历史病例数据中的信息，为疾病诊断、重要症状、指标异常等各种临床结果的预测提供支撑。

本文提出了一个基于深度学习的患者电子病历数据表示学习框架，用于临床终点（clinical endpoint）的预测任务。临床终点是指反映病人感觉、功能、生存的特征或目标变量。研究证明，深度学习在各类应用场景下有着比传统机器学习方法更优越的性能，比如图片分类[1]、语音识别[2]以及自然语言处理[3]等。深度学习的主要思路是自动化地从基本数据中抽取特征，得到对样本的有效语义表征。在电子病历方面，也希望通过深度学习的方法对患者的历史记录进行有效的表征学习。

然而，患者历史数据表示学习问题非常有挑战性，因为患者的历史记录包含了异构的时序事件，比如化验结果、生理指标、药物注射、临床事件等（如**图1**所示）。不同事件的采样记录频率千差万别，比如患者可能每个早上都要化验血糖，每个小时都要测量体温和血氧。而且不同事件之间有复杂的时序依赖关系，例如一个诊断结果需要根据患者之前的一些症状以及化验结果的某种趋势才能够得出。可以看到，对于这种包含上千种事件、采样频率差异巨大、隐含着丰富时序依赖关系的异构时序事件，学习其向量表征确实是非常复杂的。

很多文献探讨了对序列数据的表示学习方法，特别是在语音和自然语言处理领域[2-3]。现有经典的长短期记忆（long short term memory，LSTM）模型可以用于同构的序列数据，然而对于异构的序列数据，直接使用LSTM模型并不方便。也有一些工作（如多过程高斯模型[4]）对不同序列之间的相关度进行建模，然而它的计算复杂度非常大，在处理上千种时序数据时，计算代价不可承受。因此，笔者希望能够找到一种既可以对不同事件、不同采样频率建模，也能抓到不同事件之间的相关关系，并且可以比较方便地扩展应用到高维度的医疗病例数据中的方法。

本文提出了一个叫作异构事件长短期记忆（heterogeneous event long short term memory，HE-LSTM）网络的算法模型，用于学习异构时序事件的联合表征。本文算法是在相位LSTM（phased LSTM）[5]等稀疏更新的循环神经网络基础上发展起

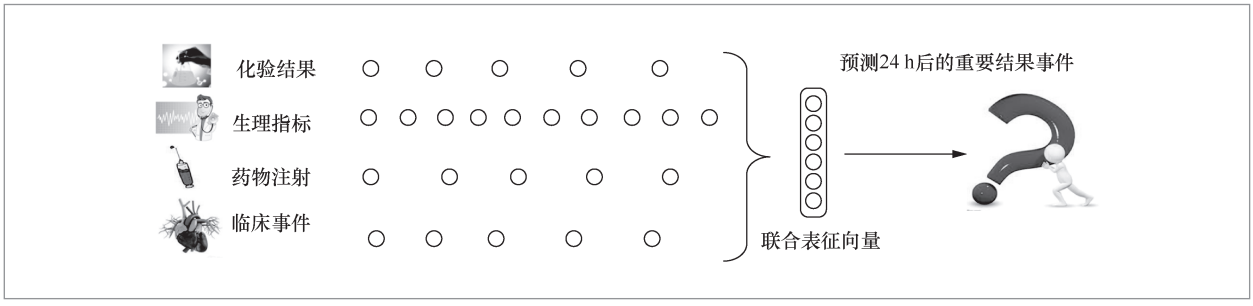


图1 异构时序事件的表示学习建模框架

来的。相位LSTM处理不规则的事件序列数据，使用一个相位门整合任意采样频率传感器收集的数据，但是它不能直接用于含有上千种事件类型的电子病历异构时序事件数据。

本文针对种类数量巨大并且采样频率差异很大的异构事件建模，每个事件及其属性被嵌入一个向量表征中，然后输入HE-LSTM模型。HE-LSTM模型依靠事件门分工合作，异步追踪记录不同采样频率下某些相关事件簇的时序信息，于是最终学习到的联合表征可以借助隐藏层延迟更新的模型结构自动整合，以此捕捉到各类事件相关的时序依赖关系。

本文在真实医疗数据上进行了一系列实验，在死亡预测和异常化验检测的任务上，都证明了本文的模型相比其他有竞争力的对照模型，预测效果提升显著。本文的模型也可以用于各类多源多频率的时序数据，比如移动传感器、慕课学生行为记录、手机应用中的用户行为记录等。

本文的主要贡献如下。

- 针对电子病历中多源异步采样的异构时序事件数据，提出了患者数据表示学习的问题建模框架。
- 提出了异构LSTM模型用于学习异构时序事件的表征向量，该模型可以适用于多规模采样频率的不同类型事件数据，并对其复杂时序依赖关系建模。
- 在真实临床数据中对死亡预测和异常化验结果预测的实验证明了模型的实用效果。

2 相关工作

2.1 电子病历数据的性质

电子病历数据的分析于2016年前后逐渐成为学术界研究的热点。电子病历可以抽象成一系列临床事件的集合，其主要性质可以从3个角度理解：信息广度（高维度、信息源异构）、时序性质（采样频率不规则、采样频率分布广）以及信息深度（医疗事件属性、先验知识丰富），从这3个基本的角度可以派生出这些医疗数据之间的多类型、多尺度、多层次的时序依赖关系。

2.1.1 信息广度

电子病历数据中包含不同来源的上万种事件（医疗特征）[6]，这就导致了信息的维度很高。比如诊断编码这一信息，在加州大学欧文（尔湾）分校(UCI)的机器学习数据库的糖尿病复诊数据中，有900个不同的诊断编码；在医疗急救服务中心的一个1万条患者数据样本中，有153个粗粒度的诊断编码。并且不同来源的数据之间有很大的差异，有多变量时间序列、不规则的事件序列、带属性的临床事件、文本等。

2.1.2 时序性质

临床事件的采样不规则性使得其采样的时间序列数据与一般等间隔采样的时间序列数据差别很大，并且事件发生的密度和患者的临床状态成一定的相关性。从时间序列的角度看，不规则采样的临床事件经常有缺失值，只包含前后顺序的序列无法反映采样频率的分布。不同类型的事件有不同规模的采样频率，如诊断事件的采样频率大约在月数量级，用药事件的采样频率大约在天数量级，生理信号的采样频率大约在小时数量级，脑电波信号的采样频率大约在毫秒数量级。

2.1.3 信息深度

有些临床事件包含了比一个数值或者

一个医疗编码更复杂的信息。比如在用药事件中，既有药物类型，又有注射剂量和给药时间信息；在化验事件中，既有化验项目，又有化验结果指数、异常标记的信息；有时患者的既往病史中还包含和医疗知识相关的先验信息和人口统计描述。

2.1.4 时序依赖关系

电子病历数据中的时序依赖关系与文本词序列中的顺序依赖关系有很多不同之处。在自然语言处理研究中，词汇的短期顺序很重要，比如，“in the front”由于语法规则，后面大多情况会跟“of”。但是长期顺序（比如段落之间）往往不是特别重要，只要是关于同一主题的即可。但是在电子病历数据中，医疗事件的短期顺序往往和病历记录系统录入相关，顺序性不重要，比如在5 min之内先做化验或先进行注射都是可以的，它们的顺序可以认为是任意指定的。然而长期顺序却非常重要，当前要做的化验和用药往往是根据之前的患者症状或指标以及疾病的阶段性发展规律判断得出的。并且，临床事件的时间戳表达的是真实时间，而不仅仅是前后关系。在病例数据中，有很多事件发生于相同时间，这更印证了医疗事件的短期无序性和长期有序性。

2.2 传统统计特征抽取与机器学习方法

为了方便建模，传统关于临床终点预测的工作一般只使用部分电子病历记录数据，从而避免直接处理异构时序事件。

一些工作在专家指导下，只使用了一个全体医疗事件的子集作为患者的特征[7]。例如，AlaaA M使用21个时序生理信号（包括11个生理指标以及10个化验结果）预测重症加强护理病房（intensive care unit，ICU）转诊[8]。也有工作选取了50个时间序列，并把这个电子病历信息的子集用多任务高斯过程建模，将多任务高斯过程的超参数作为对患者的表示向量，在这个向量空间中，通过计算患者的相似度或输入传统分类器进行预测。值得注意的是，人工选取部分特征会被动地带来专家偏向性（expert bias），只能反映出电子病历数据的部分信息。这类工作很难完全利用到电子病历整体的信息。

另一大类工作是在处理病例数据时，忽略了临床事件中的属性信息，只使用临床事件的离散变量（如ICD 编码的序列）预测临床终点[9]。例如一些工作为不同类型的临床事件分别训练语义嵌入向量，再综合预测后续的药物滥用事件（ADE）[10]。Retain[11]使用两个循环神经网络产生患者每次就诊过程中各个ICD编码的注意力权重，对原始ICD编码的嵌入向量进行加权，再预测心力衰竭的发生概率。也有工作使用卷积神经网络对不规则的医疗编码序列进行建模，预测未来的发病风险[12]。

这类工作没有利用除了医疗编码之外的、包含丰富细粒度属性的临床事件信息，如包含注射剂量属性的药物使用事件等。本文提出的电子病历分析框架可以同时建模全体生理指标时间序列、医疗编码以及包含丰富属性的临床事件信息，以充分挖掘利用电子病历中蕴含的知识。

2.3 深度学习方法

利用深度学习技术的患者表示学习主要沿用了传统事件序列（event sequence）和多变量时间序列（multivariate time series）的建模方法，独立建模病例中的各类数据。如谷歌团队提出了对这种全体未加工数据的快速可交互资源存储格式（fast healthcare interoperability

resources, FHIR)，并以此格式为基础，进行深度学习建模，从而预测多种重要的结果性事件[2]。在电子病历记录的基本格式层面，本文方法与其不谋而合，不仅可以比较宽松地适应各个医院的病例记录形式和解决预测问题，也不会因为数据规则化而丢失重要信息。

然而，在电子病历建模层面，传统方法会根据数据类型，对离散的数据（如ICD诊断编码）和连续时间序列变量（如血压时间序列）单独建模。比如时间注意力前馈模型[2]（feed forward model with time-aware attention）直接对离散的事件序列（event sequence）的表征向量建模，同时时间序列自举嵌入模型[2]（boosted embedded time-series model）对每个时间序列（time series）自举十大类时序谓词（如在某个时间点T之后值大于V等），最后筛选出10万种谓词，作为特征输入神经网络。在对记录时间的处理上，通常会用区间给固定时间间隔的事件做一个统一的时间戳，如加权循环神经网络[2]（weighted RNN）模型把离散事件序列数据分成10个大类，固定12 h为时间间隔，把序列划分为事件组的序列，每一类序列分别学习向量表示，在预测结果层面进行整合输出。

本文的思路不是以病例的数据形式做最初的划分依据，而是以医疗过程中的事件为基本单位，保留精确的事件发生时间，进行建模。这样可以更大限度地保留和反映各个事件之间的关系以及不同事件发生的频率结构。本文使用异构时序事件的表示学习框架全面建模电子病历数据时，把所有的病例信息作为带有属性的临床事件，并按照时间排序，同时保留事件的精确发生时间。这样虽然可以全盘融合异构时序事件，但由异构事件组成的序列非常长，在MIMIC-III数据集中，绝大多数患者样本序列长度过万。于是抽取异构时序事件的复杂时序依赖的问题的难点就转化成了如何保持超长序列的长期依赖关系。

然而，由于梯度消失问题，传统的循环神经网络只能抓到10跳以内的依赖关系[13]。通过设计反向传播中的无损反馈流结构，长短期记忆网络可以解决部分问题，但它的长期依赖关系大约只能维持在50跳以内[14]。

时钟循环神经网络（clock-work-RNN）[15]把隐藏层划分为多个模块，每个模块分别在不同的频率下处理输入信息，使得长期依赖容易被抽取，这种固定频率的稀疏化更新比传统LSTM更好地辅助了对长期依赖关系的建模。有一类自动决策延迟更新的神经网络——相位LSTM[5]，也对长期依赖的保持起到了比较好的作用。相位LSTM是一个先进的循环网络，它给LSTM增加了一个计时的机制，稀疏地处理非常长的输入序列。这种机制可以使相位LSTM的训练误差在反向传播中被比较好地保持，从而达到非常快的收敛速度。跳跃循环神经网络（skip RNN）也可以根据输入自动地跳过不重要的输入数据，实现稀疏化采样，更好地抽取长序列的长期依赖关系[16]。

然而，这类模型都是在同构的序列上把握长期依赖关系的，并不能直接在异构的时序事件中抓取更复杂的时序依赖关系，从而服务于对患者数据的表征学习。

3 模型设计

3.1 问题定义

3.1.1 异构事件序列

给定电子病历数据中的病人p，p的特征由序列长度为N的动态特征

$\{X_t\}_{1\leq t\leq N}$组成。$\{X_t\}_{1\leq t\leq N}$可以看作一个异质医疗事件序列，X_t为一个三元组X_t=(type,value,time)，type为事件的种类，value为事件的属性，time为事件被记录下来的时间。$\{X_t\}_{1\leq t\leq N}$中的事件按照时间先后顺序排列。事件种类向量记为e=X_t.type。病人p对应一个二元标签0或1，表示在X_N.time+24h发生的临床终点事件，例如病情平稳或患者死亡。

3.1.2 临床终点预测任务

本文目标是基于患者的历史临床数据（异构时序事件），动态地预测两个重要的临床结果。第一个任务是死亡预测，预测患者是否还在接受治疗，或是已经死亡。第二个任务是钾离子浓度异常预测，预测的目标是确认血钾化验是否为异常值。

3.2 事件嵌入

对于时间点t的事件X_t=(type,value,time)，分别对type和value进行编码。type向量为一个独热（one-hot）向量，令$C_{\text{type}}\in R^{N\times M}$为type的编码矩阵，其中$N$为编码后的维数，$M$为type种类数，则type的编码为：

$$p=C_{\text{type}}\times\text{type} \tag{1}$$

事件属性value由两部分组成：value=[value$_C$, value_n]。这二者也都是独热向量，value$_C$为事件的离散型变量属性，value_n是数值属性。同样地，用$V_C\in R^{N\times C}$，$V_N\in R^{N\times U}$分别表示它们的编码矩阵，其中，U是离散型属性的总数，C是连续型属性的总数。属性的编码与事件类型p相加得到动态事件的总体编码：

$$x_d=p+V_c\times\text{value}_c+\tanh(V_n\times\text{value}_n) \tag{2}$$

其中，V_c、V_n和C_{type}都是待学习的参数。

3.3 异构事件LSTM

图2为基础LSTM的神经元基本结构，**图3**为异构事件LSTM（HE-LSTM）的神经元基本结构。两者主要区别在于，基础LSTM神经元中有3个门函数，分别为输入门（input gate）、输出门（output gate）、遗忘门（forget gate）。

$$\begin{aligned}
i_t&=\sigma(W_{ix}x_t+W_{ih}h_{t-1}+w_{ic}c_{t-1}+b_i)\\
f_t&=\sigma(W_{fx}x_t+W_{fh}h_{t-1}+w_{fc}c_{t-1}+b_f)\\
c_t&=f_tc_{t-1}+i_t\tanh(W_{cx}x_t+W_{ch}h_{t-1}+b_c)\\
o_t&=\sigma(W_{ox}x_t+W_{oh}h_{t-1}+w_{oc}c_{t-1}+b_o)\\
h_t&=o_t\tanh(c_t)
\end{aligned} \tag{3}$$

其中，i_t、f_t、o_t分别表示t时刻的输入门、输出门和遗忘门函数。c_t是激活向量，x_t和h_t分别为t时刻的输入向量和t时刻的隐藏层输出向量。$\sigma(\cdot)$表示sigmoid激活函数，tanh(·)表示双曲正切激活函数。W_{ix}、W_{ih}、W_{fx}、W_{fh}、W_{cx}、W_{ch}、W_{ox}、W_{oh}是神经网络的矩阵参数，w_{ic}、w_{fc}、w_{oc}、b_o、b_i、b_f、b_c均为神经网络的向量参数。其中，所有参数的下标是为了区分各自所在的网络结构的位置而设置的，其命名规则是取自该参数连接的输入层和输出层的变量名，例如W_{ix}连接了输入x_t和输出i_t，故下标命名为ix，后文涉及的参数均按上述规则命名。

HE-LSTM模型的核心思路是分工合作。神经元分别以不同的周期、异步地追踪记录不同事件簇的信息。在经典LSTM的基础上，本文为每个隐藏层神经元设计了一个事件门（event gate）j_l（下标l表示第l个输入，j_l取值为0或1），如果它打开，那就正常按照LSTM更新，如果它关闭，就保持原来的值不变，如式（4）中的c_l所示。其中，$\hat{c}_l$是经典LSTM计算出的更新值。

$$c_l = j_l\tilde{c}_l + (1-j_l)c_{l-1}$$
$$\tilde{c}_l = f_l c_{l-1} + i_l \tanh(W_{cx}x_l + W_{ch}h_{l-1} + b_c) \quad (4)$$

事件门$j_{s,t}$由事件类型S和事件记录的时间t来决定开关状态。式(5)是事件门$j_{s,t}$的表达式。其由两部分构成，一部分是事件过滤器(event filter)e_s(由事件类型S决定)，另一部分则是相位门(phased gate)k_t(由时间t决定)。

$$j_{s,t}=e_s k_t \quad (5)$$

事件过滤器只允许特定的某些种类事件输入神经元，相位门使得该神经元只有特定周期下才是开放的。这样就保证了每个神经元只会抓取特定几类事件的特征，并对其进行采样，解决了时间的复杂多样性和医疗事件序列过长而导致的训练效果变差问题。事件门的开关由相位门中某个特定的周期控制，只有当门开放时才能更新式(3)和式(4)的各个参数。

事件过滤器的表达式如式(6)所示。其中，$\sigma(\cdot)$表示sigmoid函数，tanh(·)表示双曲正切函数，W_{ms}、W_{em}为训练中学习的矩阵参数，b_m、b_e为训练中学习的向量参数。事件过滤器可以让每个神经元关注各自不同的一组事件种类，从而更好地学习出混合事件序列中的信息。

$$e_s = \sigma(W_{em}\tanh(W_{ms}s + b_m) + b_e) \quad (6)$$

相位门k是一个周期性变化的函数，如式(7)所示。给定了周期τ初始相位s，则k_t是ϕ_t的函数，也随t以τ为周期变化，r_{on}是一个超参数，控制相位门开放状态占全周期的比例。在$0<\phi_t<r_{on}$时，相位门打开；在其他时刻，由于又是一个非常接近于0的超参数，k_t接近于0，相位门关闭。只有在相位门打开时，所有参数才可以更新，这样就可以对输入进行周期性的采样，从而解决输入序列过长的问题。

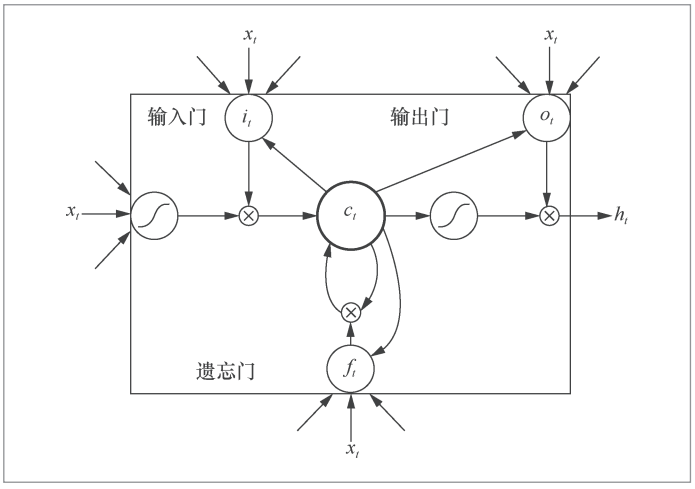


图 2 LSTM 的神经元基本结构

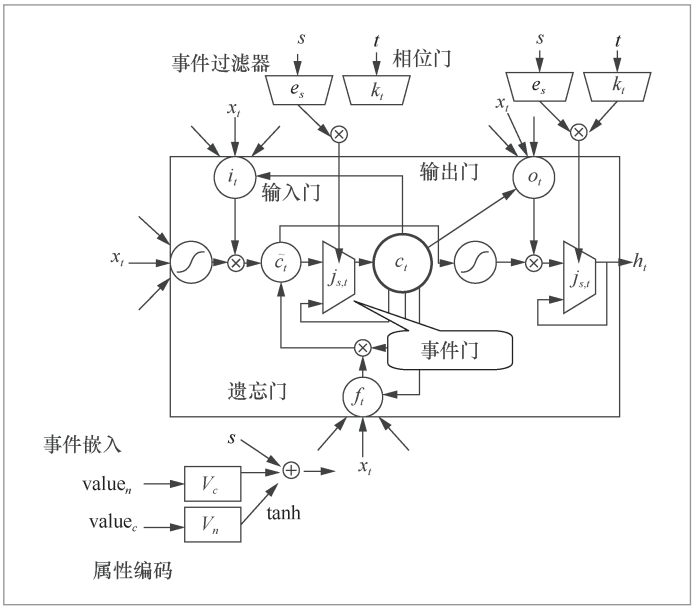


图 3 异构事件 LSTM（HELSTM）的神经元基本结构

$$\phi_t = \frac{(t-s)\bmod\tau}{\tau}, k_t = \begin{cases} \dfrac{2\phi_t}{r_{on}}, \phi_t < \dfrac{1}{2}r_{on} \\ 2-\dfrac{2\phi_t}{r_{on}}, \dfrac{1}{2}r_{on} < \phi_t < r_{on} \\ \alpha\phi_t, \text{其他} \end{cases} \quad (7)$$

综上，对于某个神经元，只有符合对应事件门的类型条件，并且在其采样周期中的事件信息才会被更新到神经元中，因此

可以认为，这个神经元表示了某一类事件在某种采样周期下的状态。

式（8）是损失函数，采用了交叉熵的形式。其中，y_t为模型t时刻的预测结果，$\hat{y}_t$表示真实指标，N为总训练样本数。式（9）中，h_t为t时刻隐藏层的输出，w_p和p_b为要在训练中学习的向量参数。

$$\mathrm{Loss}(\hat{y}_t, y_t) = \frac{1}{N}\sum_{1\leqslant t\leqslant N}(\hat{y}_t \times \ln y_t + (1-\hat{y}_t)\times \ln(1-y_t)) \quad (8)$$

$$y_t = \mathrm{sigmoid}(w_p h_t + p_b) \quad (9)$$

4 实验结果

4.1 数据描述和实验设置

本文的实验在死亡预测数据集和异常化验结果预测（钾离子异常预测）数据集上进行，该数据集由美国的一个医疗健康中心（Beth Israel Deaconess Medical Center）的ICU患者病历数据（MIMIC-III[16]）生成。

数据集抽取了MIMIC-III中的24 301个病人样本，共涵盖3 418种总数达20 290 879个异质时序事件，平均时间跨度为87 h 58 min。各类型事件的统计信息见**表1**。

实验中，数据集被划分为训练集、验证集和测试集，比例为7:1:2。通过验证集选择超参数，使用“提前结束技术（early stop）”决定训练轮数。所有实验都由Theano[17]实现，采用Adam优化算法[18]进行优化，学习率设为0.001。

表1 用于临床结果预测的电子病历异构事件来源、举例和数量统计

事件来源	举例	事件类型数量
化验指标	血细胞压积、白细胞数	525
生理指标	心率、呼吸率	385
药物注射	0.9%生理盐水、5%葡糖糖溶液	60
临床症状	异位类型、运动反应	2382
治疗程序	心电图、有创呼吸机	36

4.2 比较方法

本文将HE-LSTM与独立序列模型（独立LSTM、共享参数的独立LSTM）、延迟更新循环神经网络模型（如clock-work RNN、phased LSTM）以及医疗领域的异构序列模型（LSTM+事件嵌入、Retain）这3类方法进行比较。前文介绍过clock-work RNN和phased LSTM，这里不再赘述。

（1）独立LSTM

该模型使用多个LSTM对每个同构的事件建模，然后平均所有的结果，用逻辑回归来预测。这种方法无法对上千种的事件直接建模，因此本文只选出了可以承受的25种重要事件，与很多重要工作的筛选方式一致[8]。

（2）共享参数的独立LSTM

共享参数的独立LSTM和独立LSTM一致，但是其所有模型参数共享，从而能够支持建模所有的医疗事件。

（3）LSTM+事件嵌入

该模型使用本文提到的事件类型嵌入以及属性编码方法得到事件向量，作为传统LSTM的输入进行预测。

（4）Retain

Retain模仿内科医生逆序对病例记录建模，用两个循环神经网络输出注意力权重，并将其作为病例事件的加权权重，提供有可解释性的预测。

4.3 评测指标

由于数据集具有正负样本不平衡的问题，本文采用AUC（the area under ROC

curve)和AP(average precision[19])作为评价指标。AUC是ROC曲线与x轴所围成的面积，AP是PRC(precision-recall curve)曲线与x轴所围成的面积，二者对正负样本不平衡的数据都具有顽健性。

为了验证事件门的收敛速度，使用测试集的交叉熵(cross entropy)衡量模型在测试集的拟合程度，具体计算方法按照式(8)进行。

表2 死亡预测和异常化验结果预测的实验结果

方法	死亡预测		钾离子异常预测	
	AUC	AP	AUC	AP
独立LSTM	0.877 1	0.557 3	0.719 6	0.296 9
共享参数的独立LSTM	0.806 4	0.530 1	0.530 8	0.109 8
phased LSTM	0.847 4	0.490 0	0.772 2	0.357 5
clock-work RNN	0.840 0	0.718 1	0.651 6	0.220 8
Retain	0.896 7	0.580 8	0.732 5	0.309 6
LSTM + 事件嵌入	0.946 6	0.744 5	0.723 1	0.302 1
HE-LSTM	0.951 6	0.768 7	0.798 7	0.391 4

4.4 量化实验结果

表2展示了不同方法在死亡预测和异常化验结果预测数据集上各自的AUC和AP。从**表2**的数据中可得出以下结论。

首先，可以对不同事件相关性建模的模型，整体好于对各个数据源采样的时序数据单独建模的模型，其中HE-LSTM取得了最好的表现效果。例如，在患者死亡预测任务中，相对于表现最好的独立序列模型(独立LSTM)，Retain、LSTM+事件嵌入和HE-LSTM对AP的提升分别为0.023 5、0.187 2和0.211 4。在其他数据集和评测指标上也有一致的结果。并且本文的模型在这类异构事件序列模型上，取得了最好的效果。例如，HE-LSTM在异常化验结果预测上，相比Retain和LSTM+事件嵌入，提升了0.081 8和0.089 3的AP。可以得出如下结论：异构时序事件的相关依赖关系对临床终点预测十分有效；相比对事件独立建模，学习不同事件的联合分布可以有效地捕捉不同事件的时序依赖关系。

其次，相比密集更新的RNN，可以适应采样频率系数更新的模型更优。例如clock-work RNN相比独立建模的序列模型，在死亡预测实验上AP分别提升了0.160 8和0.188。同时，在异常化验预测结果任务上，相对于表现最好的独立序列模型(独立LSTM)，phased LSTM提升了0.052 6的AUC和0.060 6的AP。可以得出如下结论：多频率采样的模式对预测临床终点也是有帮助的，稀疏更新的模型可以充分利用这个性质，让不同单元聚焦更重要的输入，而非等同对待长序列中所有的输入。

再次，HE-LSTM在不同数据集和不同的评价指标上都取得了最高的表现。HE-LSTM不仅优于稀疏更新的RNN模型，也优于异构序列模型。仅利用多频率采样特性的稀疏更新模型和直接汇合所有不同类型事件的异构序列模型，都不是临床终点预测最好的选择。例如，HE-LSTM相比于最好的稀疏更新模型(clock-work RNN)，在死亡预测任务上提升了0.111 6的AUC和0.050 6的AP。同时，在异常化验结果预测任务上，相对于异构事件序列模型Retain，HE-LSTM提升的AUC和AP分别是0.066 2和0.081 8。可以得出如下结论：HE-LSTM由于追踪了不同事件的时序依赖关系，并自动适应不同事件的不同采样频率，取得了最优秀的预测表现。

4.4.1 不同设置下的事件门销蚀实验

为了评测不同组成部分对事件门的影响，本文将事件门分别替换成它的两个

因子，事件过滤器和相位门，其余部分和HE-LSTM一致。两个数据集的实验结果见**表3**，包括AUC、AP以及测试集交叉熵，还有在第一轮训练完成后的这3个指标。

事件过滤器的主要作用是通过对异构事件的关联建模，提升模型效果。HE-LSTM和只用事件过滤器模块的模型都可以达到比较好的预测表现。在死亡预测数据集上，HE-LSTM和只用事件过滤器的模型相比于只有相位门的模型分别提升了0.004 5和0.004 7的AUC，提升了0.022和0.021 2的AP，交叉熵降低了0.007 2和0.006 8。

相位门模块帮助HE-LSTM达到一个比较快的收敛速度，因为它具备适应不同频率采样规模的系数更新效果，可以比较好地保留回传梯度。HE-LSTM和仅使用相位门的模型在所有数据集的第一轮训练结束后的评测指标上都优于仅使用事件过滤器模块的模型。对于化验异常预测任务来说，只使用相位门的模型和HE-LSTM在第一轮训练结束后相比单独使用事件过滤器的模型，AUC要高0.030 3和0.052 8，AP要高0.034 9和0.056 2，交叉熵大约低0.007 5和0.015。

通过上述比较，可以得出如下结论：事件过滤器和相位门在对尺度采样的异构时序事件建模时，可以协同合作，达到了准确预测表现和高效的训练速度，从而提升临床终点预测的准确度。

表3 不同设置下的事件门销蚀实验

事件门的不同设置		只用事件过滤器的模型	只用相位门的模型	HE-LSTM
死亡预测	AUC（第一轮）	0.910 5	0.930 1	0.937 0
	AUC	0.951 8	0.947 1	0.951 6
	AP（第一轮）	0.604 8	0.685 6	0.709 4
	AP	0.767 9	0.746 7	0.768 7
	交叉熵（第一轮）	0.183 5	0.156 1	0.147 9
	交叉熵	0.130 1	0.136 9	0.129 7
钾离子异常预测	AUC（第一轮）	0.674 7	0.705 0	0.727 5
	AUC	0.755 9	0.794 5	0.798 7
	AP（第一轮）	0.240 3	0.275 2	0.296 5
	AP	0.341 0	0.387 5	0.391
	交叉熵（第一轮）	0.344 8	0.337 3	0.329 8
	交叉熵	0.317 8	0.301 9	0.300 3

4.4.2 不同输入序列长度的模型效果比较

为了验证本文提出的模型以及其他序列模型捕捉异构事件时序依赖的能力，本文给模型输入不同长度的序列数据，长度范围为20~1 000个事件。从**图4**可以得出以下结论。

首先，时序信息对预测临床终点是有效的。大多数模型会随着输入序列的长度增加而效果提升，特别是在输入长度小于200个的时候。

其次，HE-LSTM相比于其他模型，更擅长捕捉异构时序事件中的时序依赖关系。当输入长度较小时，各个模型的表现差不多，这种情况下时序依赖关系基本是同一段时间内出现的事件共现关系。因此各个事件单独的表示向量组合和异构事件联合表征反应的信息量基本一致。然而，当序列长度超过200并越来越长时，模型的表现稳步提升（AP从0.755提升到0.769，AUC从0.948提升到0.952）。而其他模型由于不能比较好地捕捉超长序列下的时序依赖关系，故而基本没有特别大的提升。

4.4.3 不同的初始化周期的效果比较

为了探索在事件门中事件过滤器的作用，本文比较了HE-LSTM和去掉事件过滤器（即只用相位门）的模型的作用。在给模型提供不同的初始化周期的情况下，训练死亡

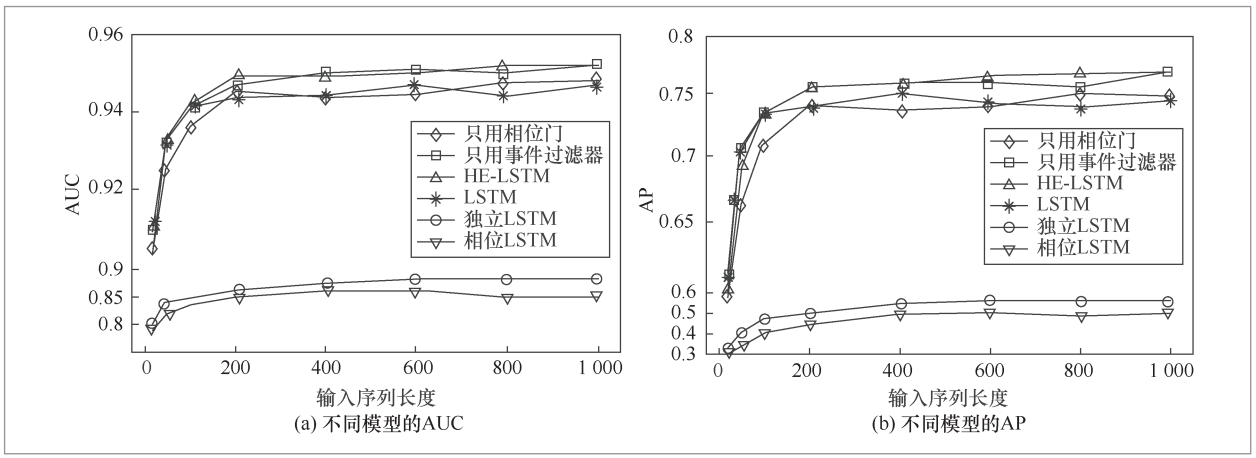


图 4 不同输入序列长度的模型效果比较示意

预测任务，这些周期是指数化后的4种均匀分布：exp(U(1,2))，exp(U(2,3)),exp(U(3,4)),exp(U(4,5))。

图5反映了两种模型在不同事件嵌入的维度及模型中门的类型设置下的表现。如“32相位门”表示事件嵌入向量的维度是32，并且模型中只包含相位门；“64事件门”表示事件嵌入维度为64，是完整的HE-LSTM模型。可以看到，完整的HE-LSTM对不同的初始化周期具有顽健性。比如在上述4种初始化设置下，完整的HE-LSTM相比于没有事件过滤器的模型，在死亡预测任务上平均提升了0.025、0.028、0.018和0.042的AP。因此可以得出如下的结论：事件门由于得到了事件过滤器的帮助，可以更容易地适应于异构时序事件的多尺度采样频率。

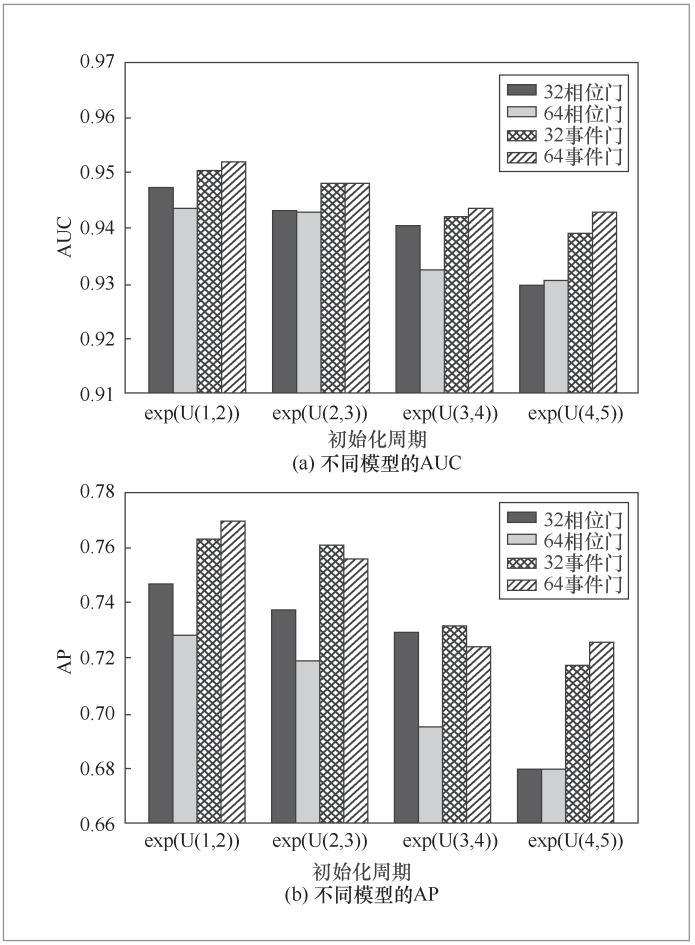


图 5 不同的初始化周期的效果比较

5 结束语

本文提出了一个异构时序事件的表示学习框架HE-LSTM模型，可以自动适应多源异构数据的多尺度采样频率，通过异步追踪不同事件的时序信息，该模型得到的患者表示向量可以捕捉到不同时间之间的时序依赖关系。并且本文的模型在真实数据集上也体现出了相对于其他典型方法的优越性。

本文的表示学习框架和方法有很多可以扩展推广的空间，例如在医疗大数据领域，可以借助医学知识图谱，整合通用先验知识和患者人口统计信息；可以加入脑电波、心电图、挖掘等多尺度的层次化时序依赖关系；服务于多任务元学习的综合动态疾病诊断。

此外，该框架可以向其他领域迁移，特别是应用于多源异步采样的传感器数据或行为记录数据等。例如在推荐系统研究中的用户行为的表征学习或者智能教育研究中的学生学习行为的表征学习等，都可以用该框架进行建模。

参考文献:

[1] HE K, ZHANG X, REN S, et al. Deep residuallearning for image recognition[C]//The IEEE Conference on Computer Vision and Pattern Recognition, June 26-July 1, 2016, Las Vegas, USA. [S. l. :s. n.], 2016: 770-778.

[2] RAJKOMAR A,OREN E, CHEN K, et al. Scalable and accurate deep learning with electronic health records[J]. Digital Medicine, 2018, 1(1): 18.

[3] BLUNSOM P, CHO K, DYER C, et al. From characters to understanding natural language (c2nlu): robust end-to-end deep learning for nlp (dagstuhl seminar17042)[R]. 2017.

[4] GHASSEMI M, PIMENTEL M A, NAUMANN T, et al. A multivariate time series modeling approach to severity of illness assessment and forecasting in icu with sparse, heterogeneous clinical data[C]//The AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, USA. Palo Alto: AAAI Press, 2015: 446-453.

[5] NEIL D, PFEIFFER M, LIU S C. Phased LSTM: accelerating recurrent network training for long or event-based sequences[C]//Neural Information Processing Systems 2016, December 5-10, 2016,Barcelona, Spain. [S. l. :s. n.], 2016: 3882-3890.

[6] JOHNSON A E, POLLARD T J, SHEN L, et al. Mimic-iii, a freely accessible critical care database[J]. Scientific Data, 2016(3): 160035.

[7] BARAJAS K L C,AKELLA R. Dynamically-modeling patient's health state from electronic medical records: a time series approach[C]//The 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 10-13, 2015, Sydney, Australia. New York: ACM Press, 2015: 69-78.

[8] ALAAA M, HU S, SCHAARM V D. Learning from clinical judgments: semi-markov-modulated marked hawkes processes for risk prognosis[J]. Computer Science, 2017, arXiv:1705. 05267.

[9] LIU C, WANG F, HU J, et al. Temporal phenotyping from longitudinal electronic health records: a graph based framework[C]//The 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 10-13, 2015, Sydney, Australia. New York: ACM Press, 2015: 705-714.

[10] HENRIKSSON A, ZHAO J, BOSTROM H, et al. Modeling electronic health records in ensembles of semantic spaces for adverse drug event detection[C]// 2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), November 9-12, 2015, Washington, DC, USA. Piscataway: IEEE Press, 2015.

[11] CHOI E, BAHADORI MT, SUN J, et al. Retain: an interpretable predictive modelfor healthcare using reverse time attention mechanism[C]//Neural Information Processing Systems 2016,December 5-10, 2016,Barcelona, Spain. [S.l.:s.n.], 2016: 3504-3512.

[12] NGUYEN P, TRAN T, WICKRAMASINGHE N,

et al. Deepr: a convolutional net for medical records[J]. IEEE Journal of Biomedical and Health Informatics, 2017, 21(1): 22-30.

[13] HOCHREITER S, BENGIO Y V, FRASCONI P, et al. Gradient flow in recurrent nets: the difficulty of learning long-term dependencies[M]. Wiley: Wiley-IEEE Press, 2001.

[14] HOCHREITER S, SCHMIDHUBERJ. Long short-termmemory[J]. Neural Computation, 1997, 9(8):1735-1780.

[15] KOUTNIK J, GREFF K V, GOMEZ F, et al. A clockwork RNN[C]//The 31st International Conference on International Conference on Machine Learning, June 21-26, 2014, Beijing, China. [S. l.]: JMLR. org, 2014: 1863-1871.

[16] CAMPOS V, JOU B, GIR-I-NIETO X, et al. Skip RNN: learning to skip state updates in recurrent neural networks[C]// International Conference on Learning Representations, April 30-May 3, 2018, Vancouver, Canada, 2018.

[17] BERGSTRA J, BREULEUX O, BASTIEN F, et al. Theano: aCPU and GPU math compiler in Python[C]//The 9th Python for Scientific Computing Conference, June 28-July 3, 2010, Austin, USA. [S. l. :s. n.], 2010: 1-7.

[18] KINGMA D, BA J. Adam: a method for stochastic optimization[J]. Computer Science, 2014, arXiv:1412. 6980.

[19] TURPIN A, SCHOLERF. User performance versus precision measures for simple search tasks[C]//The 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 6-11, 2006, Seattle, USA. New York: ACM Press, 2006: 11-18.

作者简介

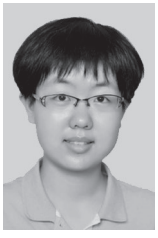
刘卢琛(1991-),男,北京大学信息科学技术学院博士生,主要研究方向为深度学习、医疗大数据等。

沈剑豪(1995-),男,北京大学信息科学技术学院博士生,主要研究方向为机器学习、自然语言处理等。

张铭（1966- ），女，北京大学信息科学技术学院教授、博士生导师，教育部高等学校大学计算机课程教学指导委员会委员，中国计算机学会（CCF）教育工作委员会副主任，ACM教育专家委员会唯一的中国理事，中国ACM教育专家委员会主席。主要研究方向为数据挖掘、机器学习、知识图谱等。

王子昌（1996- ），男，北京大学信息科学技术学院硕士生，主要研究方向为深度学习、医疗大数据、知识图谱等。

李浩然（1993- ），男，北京大学信息科学技术学院硕士生，主要研究方向为知识图谱和医疗大数据。



刘泽群（1997- ），女，北京大学信息科学技术学院本科生，研究兴趣为深度学习、医疗数据挖掘等。

收稿日期： 2018-11-05

通信作者： 张铭，mzhang_cs@pku.edu.cn

基金项目： 国家自然科学基金资助项目（No.61772039，No.91646202，No.61472006）；北京市科学技术委员会重点研发基金资助项目（No. Z181100008918005）

Foundation Items: The National Natural Science Foundation of China (No.61772039, No.91646202, No.61472006), Beijing Municipal Commission of Science and Technology Grant (No. Z181100008918005)