计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2024,33(2):166−175 [doi: 10.15888/j.cnki.csa.009422]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于 Time-aware LSTM 双向自动编码器的患者疾病分型①

赵　奎[1,2], 李　琦[1,2], 高延军[3], 马慧敏[4]

[1](中国科学院 沈阳计算技术研究所, 沈阳 110168)
[2](中国科学院大学, 北京 100049)
[3](中国医科大学附属第四医院, 沈阳 110032)
[4](东软集团股份有限公司 医疗解决方案事业本部, 沈阳 110003)
通信作者: 李　琦, E-mail: li77990311@163.com

摘　要: 医学领域中, 患有相同疾病的患者之间也存在差异性, 看似简单的疾病也可能表现出不同程度的复杂性, 这给患者的识别、治疗和预后都带来巨大挑战. 本文使用以纵向非结构化时序存储的电子病历来解决患者异质性, 通过抓住就诊时间间隔不规律的特点增强对于隐藏信息的获取, 经过前向和后向的双向学习捕捉当前就诊记录与过去和未来信息的联系, 加深对于原序列特征提取的层次, 使模型做出更为精准的决策. 本文提出的 BT-DST 模型使用 time-aware LSTM 单元构造双向自动编码器学习患者强大的单一表示, 然后将其用于患者聚类, 通过统计分析得到患者针对当前疾病的亚型分型, 可针对不同群体采用不同类型的治疗干预, 为不同类患者提供针对其健康状况的精准医疗.

关键词: 异质性; 纵向非结构化; 自动编码器; 聚类

引用格式: 赵奎,李琦,高延军,马慧敏.基于 Time-aware LSTM 双向自动编码器的患者疾病分型.计算机系统应用,2024,33(2):166–175. http://www.c-s-a.org.cn/1003-3254/9422.html

Patient Disease Typing Based on Time-aware LSTM Bidirectional Autoencoder

ZHAO Kui[1,2], LI Qi[1,2], GAO Yan-Jun[3], MA Hui-Min[4]

[1](Shenyang Institute of Computing Technology, Chinese Academy of Sciences, Shenyang 110168, China)
[2](University of Chinese Academy of Sciences, Beijing 100049, China)
[3](The Fourth Affiliated Hospital of China Medical University, Shenyang 110032, China)
[4](Medical Solutions Business Division, Neusoft Group Co. Ltd., Shenyang 110003, China)

Abstract: In the field of medicine, there are differences between patients with the same disease, and seemingly simple diseases may show different levels of complexity, which brings great challenges to patient identification, treatment, and prognosis. In this study, the electronic medical history stored in vertically unstructured time sequence is used to solve the heterogeneity of patients, enhance the acquisition of hidden information by seizing the characteristics of irregular medical treatment intervals, and capture the connection between current medical records and past and future information through forward and backward bidirectional learning, so as to deepen the level of feature extraction of original sequences and make the model make more accurate decisions. The BT-DST model proposed in this study uses a time-aware LSTM unit to construct a bidirectional autoencoder to learn a strong single representation of a patient, which is then used in patient clustering to obtain the subtype of the patient for the current disease through statistical analysis. In addition, different types of therapeutic interventions can be applied to different populations, which provides precise medicine for different types of patients according to their health conditions.

Key words: heterogeneity; vertical unstructured; autoencoder; clustering

① 基金项目: 辽宁省“百千万人才工程”(2021921015); 沈阳市中青年科技创新人才支持计划 (RC210393)
收稿时间: 2023-07-04; 修改时间: 2023-10-09; 采用时间: 2023-10-20; csa 在线出版时间: 2023-12-26
CNKI 网络首发时间: 2023-12-28

不同的患者通常存在异质性, 异质性解释了不同的潜在疾病机制. 为解决患者群异质性, 有针对地实现更好预后的首选策略就是对于临床特征更为相似的同质患者进行分离. 通过寻找疾病相似的发展途径, 可针对不同群体采用不同类型的治疗干预, 为不同类患者提供针对其健康状况的精准医疗, 患者分型有助于研究特定类型的复杂疾病[1]. 电子病历可以为了解人口健康管理提供重要的资源, 并帮助制定更好的医疗保健运营政策决策[2], 在电子病历中隐藏着被现有的临床描述所限制的多种子类型, 通过学习患者深层表示, 对其进行聚类来发掘被隐藏模式的巨大潜力, 对病情和治疗产生更深入了解. 从复杂的患者数据中分析异质患者获取其子类型极具挑战, 深度学习已被应用于获得更稳健的患者表征, 以挖掘疾病的亚型分类, Lopez-Martinez-Carrasco 等人提出一种基于跟踪的聚类, 通过以前分区的重叠聚类评估来寻找患者表型[3]; Landi 等人提出一种基于词嵌入、卷积神经网络和自编码器 (即 ConvAE) 的表示学习模型, 将患者轨迹转换为低维潜在向量, 通过聚类实现患者分层[4]; Chaudhary 等人提出了基于自编码器的多组学数据特征提取, 进行肝细胞癌的生存亚群聚类[5].

因医疗数据多为时序数据, 利用机器学习等技术可以从数据本身获取其疾病描述和其他临床概念, 且用于此方面的数据具有纵向时序特点, 所以捕捉各个记录间的时序关系就尤为重要, Lu 等人提出了基于 LSTM 临床预测模型以预测临床事件, 采用果蝇优化算法解决临床事件间的时间间隔不同问题, 提高了网络的训练效率和预测精度[6]; 赵奎等人使用将诊疗日向量和相邻两个诊疗日之间的时间间隔进行拼接, 提高了诊疗项目预测精度[7]; Bai 等人提出了一种新的可解释深度学习模型, 称为时间轴, 学习每种医疗代码的时间衰减因子通过分析时间轴的注意力权重和疾病进展函数, 预测未来就诊风险随时间的变化[8]; Zhang 等人提出了一种基于注意力的时间感知疾病进展模型 ATTAIN, 将注意力机制产生的权重与时间间隔通过衰减函数转换为衰减权重结合起来, 得到一个可解释的、临床合理的预测模型, 有助于了解感染性休克的进展行为[9].

循环神经网络 (RNN) 是时间上的展开, 常用于处理以时间序列数据作为输入的预测问题, 但其在梯度反向传播过程中存在梯度消失问题. 为解决 RNN 存在的问题, Hochreiter 等人提出了长短期神经网络 (long short term memory, LSTM)[10], 引入了“门”机制调节信息流, 学习序列中信息的记忆遗忘与否, 能够更好地捕捉较长距离的依赖关系, 然而标准的 LSTM 网络仅能用于处理规则的时间间隔, 而我们现实世界的序列数据, 尤其是医疗领域中患者就诊的时序数据通常为不规律时间间隔数据, 记录之间的时间差异从一天到几年不等, 这些时间间隔可能预示着某些疾病的发生, 若一位患者在短期内频繁入院, 其自身可能存在严重的健康问题, 而若两次就诊间间隔较长时间, 则前一次的就诊记录往往不能对当次诊断提供有用信息, 此时使用 LSTM 处理数据则不能抓住不规律间隔元素之间的依赖性.

Baytas 等人提出的时间感知长短期神经网络 (time-aware long short term memory, T-LSTM)[11]是 LSTM 的一种变体, 在 LSTM 记忆单元中加入时间不规则性以提高 LSTM 性能, 通过使用衰减函数将时间间隔转换为权重, 将单元记忆分解为短期记忆和长期记忆, 使用由衰减函数转换的时间间隔权重折算, 改变上一单元记忆对于当前输出的影响. 而 T-LSTM 同 LSTM 一样, 仅能编码从前到后的信息, 这总使得后序列信息的重要程度大于前序列, 而在处理患者入院就诊的时序数据时, 过去和未来的观测值可能包含了不同的信息, 患者诊断结果则是由其上下文记录共同决定, 对于理解疾病的发展轨迹和特征的演变, 捕捉病情变化、病程发展等动态信息至关重要, 能够更准确地预测患者的疾病状态或诊断结果. 医疗数据包含了患者的长期观测记录, 而这些观测数据之间存在着复杂的时间关系和演变模式, 通过考虑时间的因素, 模型能够更好地捕捉到医疗数据的动态变化和趋势, 提高模型对医疗数据的建模能力, 自动编码器提供了一种从原始数据直接学习映射的无监督方式[12], 原始序列由编码器重建, 其所得到的重建表示为输入序列的摘要[13] , 在患者疾病分型研究中, 自动编码器可以通过学习患者就诊数据的高级特征表示, 发现其中潜在的亚型或模式, 从而实现患者疾病的分型. 因此本文提出一种患者疾病分型模型, BT-DST 模型, 本模型使用双向时间感知长短期神经网络 (bi-directional time-aware long short term memory, Bi-T-LSTM) 构造的双向自动编码器 (Bi-T-LSTM autoencoder, Bi-T-LSTM AE), 将前向 T-LSTM 和后向 T-LSTM 组合, 更好捕捉双向的记录信息, 从两个方向学习患者时序记录的单一表示, 然后使用 K-means 算法进行聚类, 输出针对当前疾病的患者亚型分类. 患

者疾病分型研究涉及的时间序列数据, 包含患者的病历记录、治疗过程、病情变化等信息, 而 BT-DST 模型能够有效地捕捉时间间隔不规律的时序特征, 在学习表示时考虑时间因素, 在处理时序数据上具有较好的泛化性能. 基于 BT-DST 模型在处理时序数据上的优势和适应性将其用于患者疾病分型研究, 该模型结合了时间感知性和双向性, 能够有效地学习时间序列数据的复杂结构和时序依赖关系, 这种高性能的表示学习能力使得在患者疾病分型研究中能够更好地发现数据中的隐藏模式和关联信息, 提高对患者疾病亚型分类的准确性.

1 方法

1.1 Bi-T-LSTM

1.1.1 循环神经网络

RNN 是一种对序列数据进行建模的神经网络, 其结构包含输入层、隐含层和输出层. 以输入层的输入和上一时刻隐含层的输出作为当前隐含层的输入, 即 RNN 会将之前输出的信息进行记忆并应用于当前输出中. RNN 能处理任意长度的序列数据, 但是若对于长时间序列数据的输入, 后一时间节点对之前时间节点的感知能力将逐渐下降, 若当前输出与非常久远的时间序列有关, 则 RNN 难以学习, 造成梯度消失或梯度爆炸问题.

1.1.2 长短期神经网络

LSTM 将 RNN 隐藏层的简单节点改进成记忆单元, 利用记忆单元记住长期的历史信息, 利用门机制进行管理. LSTM 单元由遗忘门、输入门和输出门控制单元状态, 门结构不提供信息, 仅用于限制信息. 输入门控制信息输入, 遗忘门控制单元历史状态信息保留, 输出门控制信息输出. 但是 LSTM 结构隐含序列元素之间时间均匀分布的假设, 对于数据中可能出现的不规律时间间隔却并未考虑.

1.1.3 时间感知长短期神经网络

序列数据中时间间隔并不总是均匀分布, 不同的时间间隔可被视为是序列数据中所包含信息的一部分, 因此应在处理数据时加入此部分信息, 挖掘序列中连续元素之间不规律时间间隔的隐含信息. 两连续记录间的时间间隔可能是几天, 也可能是几年, 若时间间隔相隔过大, 则当前输出对于前一记录的依赖程度极低, 则前一单元的记忆对于当前单元的贡献应该被忽略. T-LSTM 不将长期记忆完全丢弃, 而是对于前一时间步的记忆进行子空间分解, 将短期记忆按照时间步 t 和 $t-1$ 记录之间的时间跨度大小进行调整, 使用连续记录之间的间隔时间来对短期记忆进行加权, 使得短期记忆产生相应折损, 分解网络的参数通过反向传播与其他网络参数同时学习.

首先, 式 (1) 通过网络获取前一单元的长期记忆 C_{t-1}, 使用tanh (双曲正切函数) 作为激活函数得到短期记忆 C^S_{t-1}, 其中 $\{W_d, b_d\}$ 为子空间分解的网络参数.

$$C^S_{t-1} = \tanh(W_d C_{t-1} + b_d) \tag{1}$$

式 (2) 使用非递增函数 $g(\cdot)$ 将间隔时间 Δ_t 转换为适当权重, 将其与短期记忆相乘得到折损后的短期记忆 $\hat{C}^S_{t-1}$.

$$\hat{C}^S_{t-1} = C^S_{t-1} \times g(\Delta_t) \tag{2}$$

式 (3) 为长期记忆 C_{t-1} 减去短期记忆 C^S_{t-1} 得到长期记忆的补子空间 C^T_{t-1}, 将其与折损后的短期记忆 $\hat{C}^S_{t-1}$ 相加, 组成调整后的前一单元记忆 C^*_{t-1}, 即式 (4) .

$$C^T_{t-1} = C_{t-1} - C^S_{t-1} \tag{3}$$

$$C^*_{t-1} = C^T_{t-1} + \hat{C}^S_{t-1} \tag{4}$$

后接 LSTM 标准门控体系结构, 具体公式如下. f_t 为遗忘门, 用于控制调整后的前一单元记忆 C^*_{t-1} 对当前时刻的影响. i_t 为输入门, 决定候选记忆 $\tilde{C}$ 中哪些信息需要添加至当前记忆.

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \tag{5}$$

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \tag{6}$$

$$\tilde{C} = \tanh(W_c x_t + U_c h_{t-1} + b_c) \tag{7}$$

$$C_t = f_t \times C^*_{t-1} + i_t \times \tilde{C} \tag{8}$$

o_t 为输出门, 负责决定当前时刻的单元记忆 C_t 输出大小.

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \tag{9}$$

$$h_t = o_t \times \tanh(C_t) \tag{10}$$

其中, x_t 表示当前输入, h_{t-1} 和 h_t 表示前一单元和当前单元隐藏层状态, $\{W_f, U_f, b_f\}$, $\{W_i, U_i, b_i\}$, $\{W_o, U_o, b_o\}$, $\{W_c, U_c, b_c\}$ 分别是遗忘门、输入门、输出门、候选记忆的网络参数.

$g(\cdot)$ 代表启发式衰减函数, 根据不同应用场景为 $g(\cdot)$ 选择不同类型的单调递增函数, 若数据集中两连续记录的时间间隔数值较小时, 使用式 (11), 时间间隔数值较

大时, 使用式 (12). T-LSTM 单元内部结构如图 1 所示.

$$g(\Delta_t)=\frac{1}{\Delta_t} \tag{11}$$

$$g(\Delta_t)=\frac{1}{\log(e+\Delta_t)} \tag{12}$$

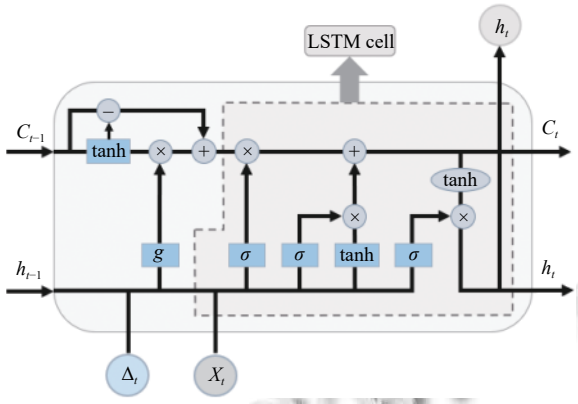


图 1 T-LSTM 单元内部结构

1.1.4 双向时间感知长短期神经网络

传统的 LSTM 仅能从前向后的传递消息, 对于当前记录之后的记录信息则无法利用, 通常对于一组时间序列数据进行处理时, 因其不依赖因果关系, 仅利用当前记录前向信息无法做出完善详尽的决策, 这在很多任务中都有局限性, 充分考虑时序数据的正反向信息规律, 进一步挖掘当前记录同过去及未来时刻记录的内在联系, 加深对于原序列特征提取的层次可以提高模型精度. 如图 2 所示, Bi-T-LSTM 能够将前向和后向 T-LSTM 的输出拼接, 相对于单向 T-LSTM 来说, 能够同时获取从前向后和从后向前两个方向的信息.

$$H=[h_L\oplus h_R] \tag{13}$$

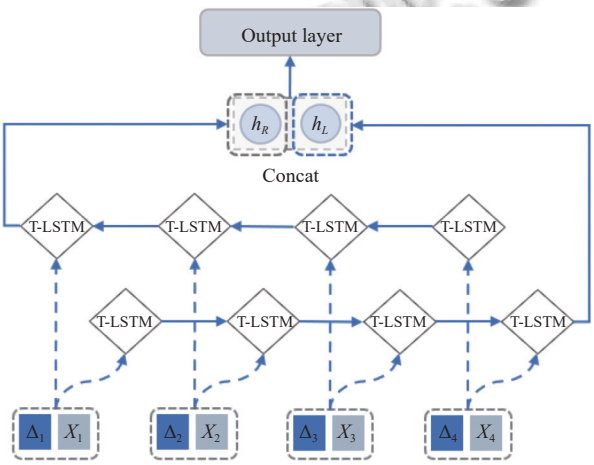


图 2 Bi-T-LSTM 模型结构

其中, h_L 表示正序 T-LSTM 隐藏层输出, h_R 表示逆序 T-LSTM 隐藏层输出, 将两输出拼接得到 Bi-T-LSTM 的输出 H, 对于时间序列数据处理时能够获取更多的信息, 有利于后续任务, 因此 Bi-T-LSTM 在解决时序数据问题时能够获得比单向 T-LSTM 更好的效果.

1.2 患者疾病分型模型

在医学上存在共同医学特征的集合称为一个亚型, 本研究通过处理患者电子病历数据, 在某种疾病范畴内, 根据其特征、病理机制、临床表现等方面的差异, 将患者进行聚类, 区分不同的亚型. 对于真实世界的疾病数据集没有关于患者队列内的任何先验信息, 疾病本身的复杂性和异质性使得某些疾病可能存在多个隐含的亚型, 而且这些亚型在现有的临床分类系统中尚未被充分认识或明确定义, 因此将患者亚型划分为无监督聚类问题. 从大规模患者就诊记录的电子病历中提取疾病的亚型可以指导下一代个性化医疗. 然而患者电子病历数据集的一个重要性质就是其序列中连续记录的时间间隔是不规律的, 其中患者记录的频率和数量是非结构化的, 其纵向数据缺少信息, 无法得知患者未就诊的时间间隔其疾病的发展情况, 且仅利用当前就诊记录前向信息无法做出完善详尽的决策, 需同时考虑双向记录获取过去及未来对于疾病当前发展状况的内在联系, 使用现有的临床数据进行聚类分析可能无法完全捕捉到其存在的亚型, 因此, 使用一种能够克服这种时间间隔不规律, 并充分考虑就诊记录时序信息的模型结构学习患者疾病特征, 将由此学习到的患者表示用于聚类能够更为精确地得出疾病的亚型分类.

本文提出了一种基于 Bi-T-LSTM 构造的双向自动编码器 (Bi-T-LSTM autoencoder, Bi-T-LSTM AE) 和 K-means 算法组成的疾病分型模型, 简称 BT-DST. 使用此无监督框架处理时序电子病历数据, 通过自动编码器模型重构患者数据, 在就诊记录存在时间不规则的情况下, 双向考虑就诊时序, 训练自动编码器重构数据, 直至重构误差最小. 然后将患者序列重新输入至已保存训练权重的编码器部分, 由编码器学习到的输出即为患者记录的有效表示, 此表示将患者信息转换为低维潜在向量, 保持了原始序列随时间变化的时间动态, 能够从双向抓住患者因就诊间隔不同而暗含的疾病状况, 然后使用 K-means 算法对于上述编码器部分的输出所得患者数据的判别表示进行聚类, 将拥有相似特征的患者分为一组, 被聚类为一组的患者为一个亚型, 这样的特征

提取能力可以帮助模型发现患者疾病数据中的关键模式和特征, 从而提高分型任务的性能. 当为一位新患者判断亚型分类时, 将新患者列数据输入至已保存训练权重的编码器结构中, 输出患者的单一表示, 将新加入患者与各个聚类子类型的质心距离进行比较, 将其加入距离最近的质心所在组即为该患者的亚型分类.

图 3 为 BT-DST 模型整体框架, 其由 Bi-T-LSTM AE 部分以及聚类部分构成, Bi-T-LSTM AE 使用两个 T-LSTM 编码器和两个 T-LSTM 解码器组成, 其中编码器分别为前向编码器和后向编码器, 前向编码器和后向编码器均由多个 T-LSTM 单元组成, T-LSTM 单元的个数由当前输入的单个患者的序列记录数决定, 解码器与编码器结构相同, 通过联合学习使双向重构误差最小化, 以 T-LSTM 单元构建自动编码器能够将不规律的时间间隔信息加入数据依赖学习中, 捕获长期和短期记忆间的关系, 更好的学习到输入序列的单一表现形式, 用此单一表现形式进行无监督聚类得到更准确的分类效果.

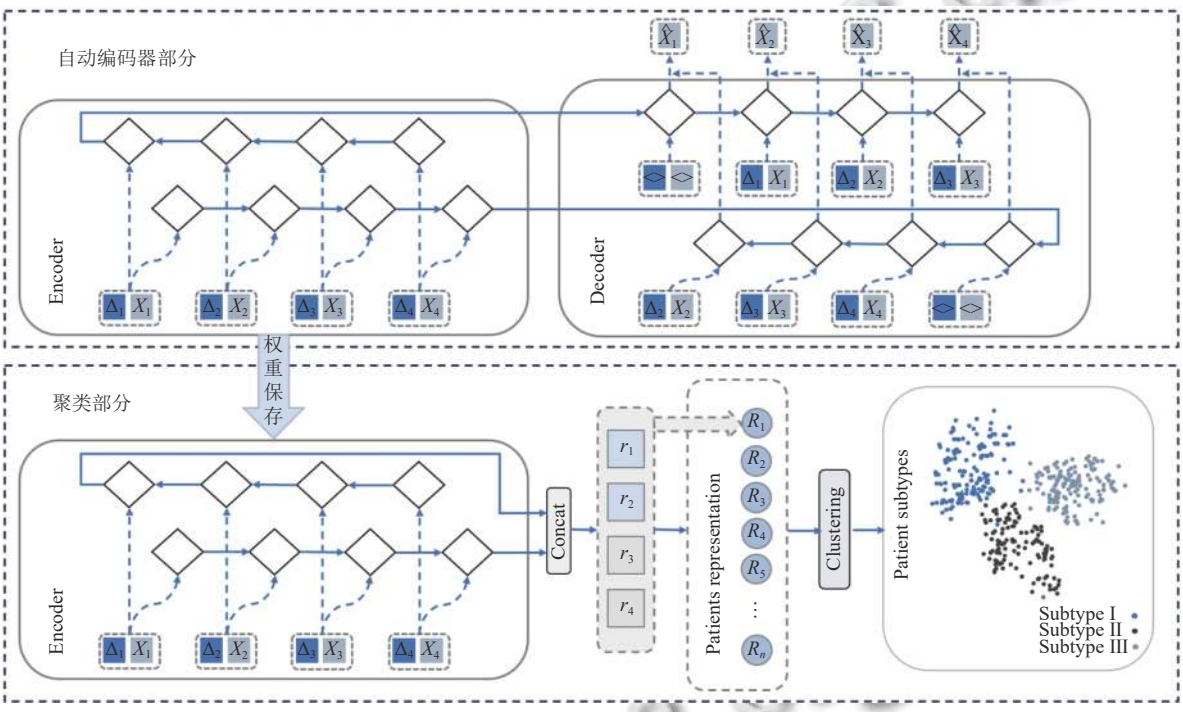


图 3 BT-DST 模型架构

前向 T-LSTM 以$[X_1,X_2,X_3,X_4]$序列作为输入, 后向 T-LSTM 以$[X_4,X_3,X_2,X_1]$序列作为输入, 编码器双向末端 T-LSTM 单元输出的隐藏状态和单元记忆均分别作为解码器双向首个 T-LSTM 单元的初始隐藏状态和上一单元记忆. 解码器的双向首个 T-LSTM 单元数据输入和间隔时间输入均被设置为零. 前向解码器第 1 个输出为原始序列最后一个元素X_4的重构$\hat{X}_4$, 输出序列为$\left[\hat{X}_4,\hat{X}_3,\hat{X}_2,\hat{X}_1\right]$, 后向解码器第一个输出为原始序列首个元素$X_1$的重构$\hat{X}_1$, 输出序列为$\left[\hat{X}_1,\hat{X}_2,\hat{X}_3,\hat{X}_4\right]$.

$$E_r=\sum_{i=1}^{L}\left\|X_i-\hat{X}_i\right\|_2^2+\sum_{j=L}^{1}\left\|X_j-\hat{X}_j\right\|_2^2 \tag{14}$$

其中, L为输入序列长度, X_i为正向序列的第i个元素, X_j为反向序列的第j个元素, 训练目标重构误差E_r最小化时, 将编码器部分的权重保存. 重新将序列输入至已保存权重的编码器中, 分别获取双向编码器末端的隐藏状态, 隐藏状态携带了前向和后向的所有信息, 因此解码器可根据其重构原始序列, 将两隐藏状态进行拼接即得到双向序列经编码器学习后的表示. Bi-T-LSTM AE 的隐藏层维度均设为 2, 序列输入后经编码器将高维数据减小至二维, 将前向和后向输出进行拼接, 得到四维表示向量, 使用 K-means 算法对此向量进行聚类, 在进行聚类可视化时仅选取表示向量的前两维, 使其能够在二维空间中绘制表示.

图 3 中将 n 位患者的单一表示$(R_1,R_2,\cdots,R_n)$由 K-means 聚类算法进行聚类分组, K-means 算法将拥有相似特征的患者分为一组, 在医学上存在共同医学特

征的集合称为一个亚型, 因此被聚类为一组的患者即为一个亚型, 图 3 中所示存在 3 种亚型.

2 实验

本实验由于无法得知真实世界的疾病数据集聚类情况的基本事实, 因此分别对已知数据分布及分类情况的人工生成的电子病历数据集和真实世界的电子病历数据集进行实验. 首先在人工生成数据集使用所提出 Bi-T-LSTM 模型在有监督情况下进行数据分类, 检验其对于挖掘序列电子病历数据蕴含信息的能力. 在无监督情况下, 使用 BT-DST 对人工生成的患者数据进行学习获得的单一表示, 检验其是否能更精准的进行聚类得到不同的亚型分型. 其次在真实世界的数据集 SEER 上进行无监督聚类, 验证本文提出的患者疾病分型模型是否能够对于胃癌患者的亚型分类产生积极影响. 本实验中所有权重的学习均以数据驱动方式同时进行, 将患者时序数据按照相同序列长度分批, 使用固定迭代次数, 以小批量 Adam optimizer 优化器进行优化.

2.1 评价指标

针对有监督分类实验采用 3 种在医疗领域研究中被广泛使用的分类性能评估指标对于不同模型的性能进行评估, 分别为准确率 (*Accuracy*)、*F*1-*score* 以及 AUC (area under curve), 而对于无监督聚类实验则使用聚类算法的常用评价指标 Rand 指数 (Rand index, *RI*). 令 *TP* 表示预测为正类的正样本; *TN* 为预测为负类的负样本; *FP* 为预测为正类的负样本; *FN* 为预测为负类的正样本.

各个评估指标的定义以及计算公式如下.

Accuracy (准确率) 是分类任务中最常用的指标, 它表示的是分类正确的预测数在总预测数中占据的比例, 其计算公式为:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{15}$$

然而当数据不平衡时, *Accuracy* 评估方法的缺陷尤为显著. 因此, 我们需要引入 *Precision* (精确率), *Recall* (召回率) 和 *F*1-*score* 评估指标, *Precision* 是所有被预测为正的样本中实际为正的样本的比例, 计算公式如下:

$$Precision = \frac{TP}{TP+FP} \tag{16}$$

Recall 是实际为正的样本中被预测为正样本的比例, 计算公式如下:

$$Recall = \frac{TP}{TP+FN} \tag{17}$$

而 *F*1-*score* 则是 *Precision* 和 *Recall* 二者的调和平均, 其计算公式为:

$$F1\text{-}score = \frac{2\times Precision\times Recall}{Precision+Recall} \tag{18}$$

受试者工作特征曲线 (receiver operating characteristic, ROC) 是通过不断移动分类器的“截断点”来生成曲线上一系列关键点的. 在此引入两个公式 *TPR* (真阳性率) 和 *FPR* (假阳性率).

$$TPR = \frac{TP}{(TP+FN)} \tag{19}$$

$$FPR = \frac{FP}{(FP+TN)} \tag{20}$$

ROC 曲线是以假阳率为横轴, 以真阳率为纵轴的二维平面曲线, 其能在数据集出现类不平衡现象时不受干扰, ROC 曲线越接近左上角, 该分类器性能越好. 由于 ROC 很难反映出模型之间的差异, 因此使用 ROC 曲线下面积进行对比, 即 AUC.

Rand 指数用于衡量聚类结果与数据的外部标准类之间的一致程度, 评估聚类方法所学习表示的辨别能力.

$$RI = \frac{TP+TN}{TP+FP+FN+TN} \tag{21}$$

RI 取值范围为[0, 1], 值越大意味着聚类结果与真实情况越吻合, 划分的一致程度越高, 当 Rand 指数值取 1 时, 划分完全一致.

2.2 人工生成数据集

2.2.1 数据处理

此人工生成的电子病历数据集中包含 10 万多名患者病历数据, 其中包括患者入院时间, 实验室测量结果, 用药日志, 病情诊断等多种特征, 虽然其数据为人工生成, 但参考其生成方法[14]可知, 此数据集中数据分布具有与真实世界电子病历数据相似的特征, 故可用其进行研究.

有监督分类实验选取其中与糖尿病相关的患者 767 例, 将数据中缺失值大于 90% 的特征列删除, 剩余存在缺失值的特征列, 连续变量使用均值填充, 分类变量使用“Not recorded”填充进行独热编码后得到数据维度为 529. 将数据集按照 7:3 的比例划分为训练集和验证集, 其中训练集 536 例, 验证集 231 例.

无监督聚类实验中使用具有与患者电子病历数据相同结构的 4 种不同均值和相同协方差的正态分布聚

类数据进行实验, 患者序列数据被随机丢弃, 以形成存在不规律时间间隔的数据分布, 其分批序列长度为 4, 6, 18, 22, 30, 输入数据维度为 5.

2.2.2 有监督实验结果

实验比较 Bi-T-LSTM 与仅考虑正向序列信息的单层 T-LSTM 模型, 常用于处理电子病历等时序数据的未加入不规律时间间隔信息的传统单层 LSTM 模型, 以及不考虑时序信息的传统 logistic 回归分类器的性能差别, 验证 Bi-T-LSTM 在同时考虑当前时间步之前和之后的上下文信息以及序列记录随时间变化的时间动态后, 对于有监督分类实验性能是否得到明显提升, 能否捕捉更丰富上下文信息, 从而更全面地理解序列数据. 本实验分别使用 Bi-T-LSTM、T-LSTM、LSTM、logistic 回归模型对于数据进行分类, 其中神经网络模型的参数设置如表 1. 由于逻辑回归模型为线性分类模型, 不考虑时间信息, 因此将每个患者的就诊记录进行汇总, 作为逻辑回归模型的输入, 其学习率为 1E–3, 迭代次数为 100. 实验结果如表 2 所示.

表 1 神经网络模型参数设置

参数	learning_rate	training_epochs	hidden_dim	full_connect
值	1E–3	100	128	64

表 2 展示了 4 种不同模型对于此数据集进行分类的准确率和 AUC 值. 从表 2 结果, 逻辑回归分类器的 *Accuracy*, AUC, *F*1-*score* 均为最低可知, 另外 3 种神经网络模型能够有效地建模序列信息, 捕捉到序列数据中的时间依赖关系, 且电子病历数据中的每个观测结果都可能与前面的观测结果相关联, 具有上下文信息, 传统的逻辑回归分类器只能考虑当前观测结果的特征, 无法有效地利用上下文信息, 因此在对数据进行分类时, 神经网络模型相较于不考虑时间信息的传统逻辑回归分类器的分类效果有明显优越性; T-LSTM 比之 LSTM 在准确率上提升了 7.25%, AUC 提升了 6.6%, *F*1-*score* 提升了 8.93%, T-LSTM 通过引入时间注意力机制, 在输入中加入了不规律时间间隔成分后能够比 LSTM 获取更多与时间间隔相关信息, 使得模型能够自适应地学习不同时间间隔的重要性权重, 从而更好地学习患者病历数据中的动态变化, 提高模型精度; 由 Bi-T-LSTM 比 T-LSTM 的准确率提升了 0.95%, AUC 提升了 1.15%, *F*1-*score* 提升了 0.84%, 可知 Bi-T-LSTM 模型从双向分析数据能够更好地抓住记录的双向信息, 捕捉时序数据中的上下文和依赖性, 该模型在进行双向特征拼接时充分利用了电子病历数据的前向和后向信息, 从而更全面地学习患者病情的变化, 这种双向建模的能力使得模型更为适应序列数据的建模任务, 更好地理解电子病历数据中的复杂关系, 使之具有更优秀的分类效果.

表 2 人工数据集下不同模型分类结果

Method	*Accuracy*	AUC	*F*1-*score*
Logistic regression	0.5553	0.5762	0.6024
LSTM	0.8226	0.8423	0.8475
T-LSTM	0.8951	0.9083	0.9368
Bi-T-LSTM	**0.9046**	**0.9198**	**0.9452**

2.2.3 无监督实验结果

为验证 BT-DST 在患者疾病分型任务中学习表示的表达能力. 为此, 使用一个已知基本事实的合成数据集, 对其进行亚型聚类, 并对聚类结果进行评估. 与有监督实验相同, 同样分别使用单层单向 T-LSTM、LSTM 构造模型中自动编码器部分, 后加 K-means 算法, 分别组成 T-LSTM AE-KM 模型和 LSTM AE-KM 模型, 以及使用在医疗领域常用于表示学习的 Deep Patient 神经网络模型[15]进行数据表征学习能力的对比.

本实验中 T-LSTM AE-KM 模型和 LSTM AE-KM 模型输入均为正向数据序列及时间间隔, 编码器末端单元输出的隐藏状态和单元记忆分别作为解码器首个单元的初始隐藏状态和上一单元记忆. 解码器的首个单元数据输入和间隔时间输入均被设置为零. 各模型参数设置如表 3 所示, 训练自动编码器使得重构误差最小化, 然后仅保留编码器部分的输出用于输入至 K-means 算法中进行亚型聚类. BT-DST 模型中 Bi-T-LSTM AE 隐藏层维度均设为 2, 双向向量拼接得到 4 维表示向量, T-LSTM AE 和 LSTM AE 隐藏层维度均设置为 4, 同样得到 4 维表示向量用于聚类. Deep Patient 采用 3 层自动编码器, 每层神经元数均为 50, 将每个患者的就诊记录进行汇总作为输入, 输出层维度为 4. 各模型训练迭代次数均为 2 000 次.

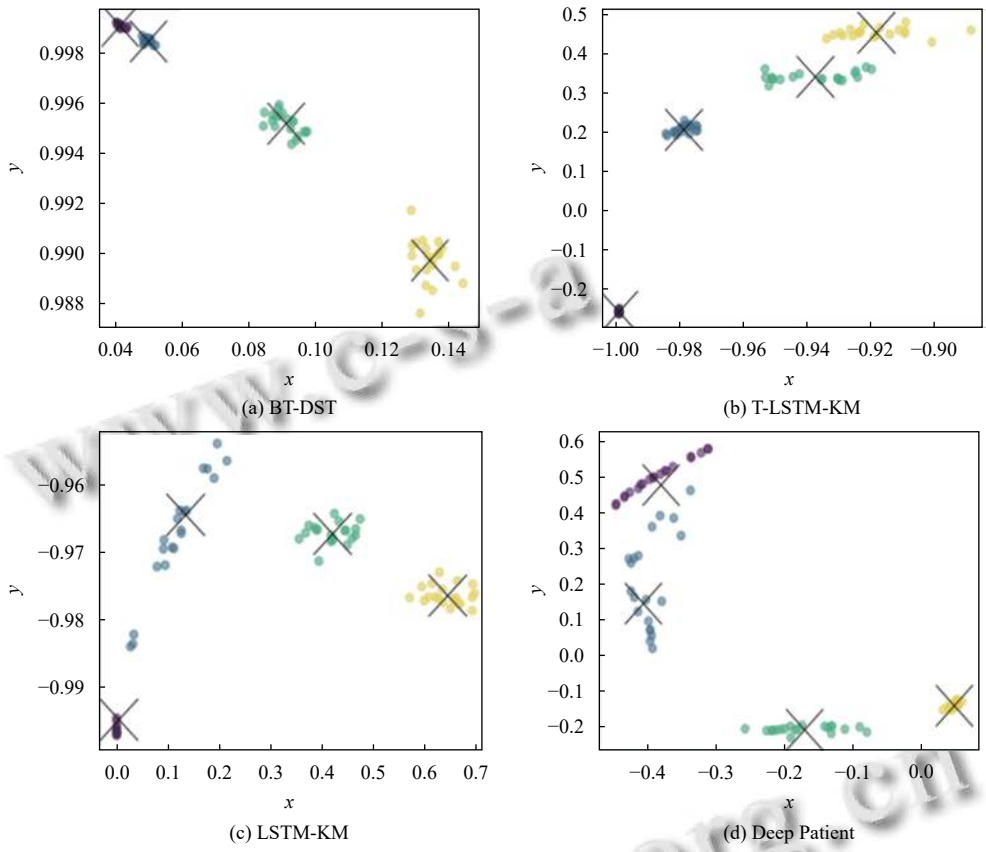
由于我们知道合成数据的基本事实, 因此计算聚类的 Rand 指数 (*RI*), 以观察所学习到表征的判别能力. Rand 指数的较大值表明学习到的表征聚类接近基本事实. 启发式衰减函数选用式 (12), K-means 聚类的 *K* 值设为 4, 每组均进行 10 次重复实验, 使用平均 Rand 指数评估各个模型对于数据的聚类效果. 各模型对比 *RI* 值如表 4 所示, 将 4 种模型所得二维表示进行聚类可视化如图 4 所示.

表 3 模型参数设置

Method	learning_rate	hidden_dim
Deep Patient	1E–3	50/50/50
LSTM AE-KM	1E–3	4
T-LSTM AE-KM	1E–3	4
BT-DST	1E–3	2

表 4 人工数据集下不同模型聚类结果

Method	Mean *RI*
Deep Patient	0.795 6
LSTM AE-KM 1-layer	0.886 3
T-LSTM AE-KM 1-layer	0.956 5
BT-DST	**0.964 6**



(a) BT-DST (b) T-LSTM-KM

(c) LSTM-KM (d) Deep Patient

图 4 不同模型聚类可视化对比

由表 4 结果可知, 不考虑序列时间特征的 Deep Patient 的平均 Rand 指数相较于其他 3 种用于处理时间序列数据的模型表现最差, 由于电子病历数据中的时序特征反映了患者在不同时间点的医疗情况和变化趋势, 因此通过建模这种时间依赖性, 可以更好地捕捉到疾病的演变过程、治疗效果以及患者的健康状态变化等. 表 4 中 BT-DST 的平均 Rand 指数比标准 LSTM AE-KM 高 0.078 3, 比 T-LSTM AE-KM 高 0.008 1, 可知 BT-DST 聚类性能优于其他两种表示学习方法, 其聚类结果与真实标签的一致性最高, 能够有效地将相似的样本归为一类, 此类拥有不规律时间间隔的序列数据对于此不规律性较为敏感, BT-DST 具有良好的时间建模能力, 在学习数据的同时, 捕捉了数据内在的结构和模式, 有助于更好地理解时间序列数据的动态演变过程, 从而提高了患者疾病分型的准确性, 不但能够更为全面的学习到在数据中存在非结构化时间间隔时数据的隐含信息, 同时能够捕捉输入序列所携带的前后信息, 能够利用时间上的依赖关系, 提取时间上的特征, 更全面地理解患者疾病数据. 观察图 4 可知, BT-DST 对于输入序列的学习表示能够在二维空间中产生更为紧凑的组, 更能够捕捉到数据中的内在结构和类别之间的差异, 反映了数据的更好可分性, 能够很好地捕捉到数据的内在异质性, 并将其反映在聚类结果中, 使其更为准确, 这对于理解数据的结构和属性分布、揭示数据之间的关联关系具有重要意义.

2.3 真实数据集

2.3.1 数据处理

本研究将利用美国国家癌症研究所 (National

Cancer Institute, NCI) 建立的权威监测、流行病学及肿瘤患者预后随访数据库 (surveillance, epidemiology and end results, SEER), SEER 公开数据集的数据存储格式与上述人工生成电子病历相同, 均为非结构化纵向时序数据. 选取序列长度不小于 3 的胃癌患者记录为研究队列, 统计特征缺失值百分比, 删除缺失率大于 90% 的特征列, 共保留 226 维特征. 对于数值型特征使用平均值进行插值, 对于分类特征使用“Not recorded”填充, 然后进行独热编码, 最终得到 809 例患者时序数据, 共 12 833 条数据, 数据维度为 953, 其中 655 例胃癌临床 N1 期患者和 154 例胃癌临床 N0 期患者对照[16].

2.3.2 实验结果

在本实验中, 由于学习到的表示与原始输入相比维数较低, 由高维学习到低维空间的映射需要更多的复杂性, 以便捕获高维输入序列的更多细节, 单层自编码器需要更多的迭代次数来最小化重构误差. 在本实验中, 由于上述原因, 均使用两层 T-LSTM 和两层 LSTM 构造模型中自动编码器部分, 其中第 1 层的输出是第 2 层的输入, 后加 K-means 算法, 分别组成 T-LSTM AE-KM 模型和 LSTM AE-KM 模型与 BT-DST 进行对比实验. 本实验中各模型参数设置如表 5 所示, BT-DST 的隐藏层维度设为 2, 经双向训练后拼接的表示向量维度为 4, T-LSTM AE-KM 和 LSTM AE-KM 的两层隐藏层维度分别为 32 和 4, 各模型训练迭代次数均为 3 000 次. 然后将编码器部分学习到的单一表示用于 K-means 聚类.

表 5 模型参数设置

Method	learning_rate	hidden1_dim	hidden2_dim
LSTM AE-KM 2-layer	0.001	32	4
T-LSTM AE-KM 2-layer	0.001	32	4
BT-DST	0.001	2	—

分别对数据分析患者聚类的子类型, 由于无法得知聚类的基本事实, 因此通过统计分析了解聚类结果间接知悉聚类性能.

尝试使用多个 K 值进行 K-means 聚类, 观察结果发现常见两种主要聚类, 因此设置 K-means 聚类参数为 K=2. 为研究聚类产生的子类型分类性能, 对于聚类结果进行解释, 将分类特征使用卡方检验, 连续特征使用 f 检验进行聚类比较, 评估特征与聚类结果之间的相关性, 当聚类结果的特征具有小于 0.05 的显著性水平 (即 p-value 小于 0.05) 时, 意味着特征与聚类结果之间的相关性是显著的, 这表明特征在区分不同的聚类簇时起到了重要的作用, 将对应的 p-value 小于 0.05 的特征认定为存在显著群体效应的相关特征, 表 6 列出了使用 3 种模型分别进行聚类所得结果, 表中仅列出 p-value 小于 0.05 的显著特征及其分别在各自类群中的均值. 由表 6 可知, LSTM AE-KM 聚类结果中无显著相关性特征, 大多数患者被归为一个类群 (null 表示空值); BT-DST 较之 T-LSTM AE-KM 多识别出 positive lymph node 以及 lymph nodes removed 两种特征. p-value 小于 0.05 的特征越多, 表示这些特征在不同聚类簇之间具有较大的差别, 能够更好地区分不同簇之间的差异和相似性, 这些特征在聚类分析中对区分不同聚类簇的能力更强, 聚类结果更可靠. 比较 BT-DST 所得两聚类均值, 子类型 1 的 positive lymph node 值显著高于子类型 2 的患者, N0 期和 N1 期的胃癌患者在淋巴结转移的程度上有明显区别, N0 期患者没有淋巴结转移, 癌细胞局限在原发肿瘤部位, 而 N1 期患者存在淋巴结转移, 癌细胞已经扩散到胃周围的淋巴结, 胃癌转移至淋巴结可能会影响癌症的治疗策略和预后, 这种区别对于指导治疗方案的选择和预后评估非常重要. 因此可知处于 N1 期的患者主要集中于子类型 1, 处于 N0 期的患者主要集中于子类型 2 中. BT-DST 产生了更多具有小 p-value 的显著特征, 意味着该方法在真实世界患者数据集上, 对于存在未知基本事实的数据进行患者亚型聚类分析, 能够通过加入非结构化时间间隔和通过双向获取电子病历信息提供更为精确的患分型结果.

表 6 真实数据集下不同模型聚类结果分析

Method	Feature	p-value	Mean	
			Cluster1	Cluster2
BT-DST	Tumor length	0.001 6	6.843 75	4.824 742
	Positive lymph node	0.001 9	**5.567 8**	**1.094 8**
	Grade of differentiation	0.002 3	2.265 6	2
	Histological type	0.010 8	0.668 0	2.077 5
	Lymph nodes removed	0.030 5	**0.21**	**0.08**
	Location of cancer	0.035 6	4.663 8	3.287 5
	Lauren classification	0.042 5	1.765 85	1.368 75
T-LSTM AE-KM	Tumor length	0.020 5	6.235 44	5.326 56
	Grade of differentiation	0.025 9	2.260 2	2.153 6
	Histological type	0.034 5	0.985 6	2.166 3
	Location of cancer	0.044 5	4.236 5	4.054 2
	Lauren classification	0.048 7	1.656 25	1.426 56
LSTM AE-KM	null	null	null	null

3 总结

疾病的发展和演变通常受到时间的影响, 不同时

间点上的观测值可能对患者疾病分型有不同的重要性. 本文使用一种能够将时间序列数据中存在的不规律时间间隔信息嵌入处理的 T-LSTM 单元构建 Bi-T-LSTM 模型, 使用 Bi-T-LSTM 结构分别在人工生成的数据以及真实世界的数据上分别进行有监督分类和无监督聚类实验, 捕捉时间上的依赖关系, 从而更好地表达患者疾病的动态变化. 实验结果表明使用 Bi-T-LSTM 结构构建的 BT-DST 模型通过学习患者电子病历数据的压缩表示来捕捉数据的重要特征进行聚类时, 不但能够通过加入到输入中的时间间隔信息来挖掘数据中因两连续记录间隔时间长短产生的影响, 同时利用过去和未来的观测值, 联合进行更为全面的数据学习, 增强对患者疾病发展的建模能力. BT-DST 所输出的患者单一表示提取出对疾病分型具有较高区分能力的特征, 更深刻的抓住了患者信息, 为研究疾病子类型的聚类提供了极大的积极影响, 从而产生更为精确的分型结果. 本文所提出的 BT-DST 模型被设计用于处理患者电子病历数据, 模型所使用的 T-LSTM 单元决定其在处理具有不规律时间间隔特征的时序数据时具有突出表现, 擅长于捕捉时间依赖性和动态变化的建模任务. 因 BT-DST 模型具有的强大的特征提取能力, 将 BT-DST 模型应用于如病程预测, 病因分析等其他业务或领域时, 具有一定的适应性和泛化性, 需确保该领域数据与本论文所用数据具有相似时间特性, 领域数据的时间序列信息和连续记录之间的间隔对模型的性能产生重大影响, 针对模型进行特定的调整后, 在其应用场景中进行研究, 以确保其在不同领域的有效性.

参考文献

1 Che ZP, Kale D, Li WZ, *et al*. Deep computational phenotyping. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Sydney: ACM, 2015. 507–516.

2 Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: Towards better research applications and clinical care. Nature Reviews Genetics, 2012, 13(6): 395–405. [doi: 10.1038/nrg3208]

3 Lopez-Martinez-Carrasco A, Juarez JM, Campos M, *et al*. A methodology based on trace-based clustering for patient phenotyping. Knowledge-based Systems, 2021, 232: 107469. [doi: 10.1016/j.knosys.2021.107469]

4 Landi I, Glicksberg BS, Lee HC, *et al*. Deep representation learning of electronic health records to unlock patient stratification at scale. npj Digital Medicine, 2020, 3: 96. [doi: 10.1038/s41746-020-0301-z]

5 Chaudhary K, Poirion OB, Lu LQ, *et al*. Deep learning-based multi-omics integration robustly predicts survival in liver cancer. Clinical Cancer Research, 2018, 24(6): 1248–1259. [doi: 10.1158/1078-0432.CCR-17-0853]

6 Lu WJ, Ma L, Chen H, *et al*. A clinical prediction model in health time series data based on long short-term memory network optimized by fruit fly optimization algorithm. IEEE Access, 2020, 8: 136014–136023. [doi: 10.1109/ACCESS.2020.3011721]

7 赵奎, 闫玉芳, 曹吉龙, 等. 融合规范化判断的双向循环神经网络诊疗预测模型. 小型微型计算机系统, 2022, 43(6): 1278–1284.

8 Bai T, Zhang SS, Egleston BL, *et al*. Interpretable representation learning for healthcare via capturing disease progression through time. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. London: ACM, 2018. 43–51.

9 Zhang Y, Yang X, Ivy J, *et al*. ATTAIN: Attention-based time-aware LSTM networks for disease progression modeling. Proceedings of the 28th International Joint Conference on Artificial Intelligence. Macao: AAAI Press, 2019. 4369–4375.

10 Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 1997, 9(8): 1735–1780. [doi: 10.1162/neco.1997.9.8.1735]

11 Baytas IM, Xiao C, Zhang X, *et al*. Patient subtyping via time-aware LSTM networks. Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Halifax: ACM, 2017. 65–74.

12 Bengio Y, Courville A, Vincent P. Representation learning: A review and new perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(8): 1798–1828. [doi: 10.1109/TPAMI.2013.50]

13 Cho K, van Merriënboer B, Gulcehre C, *et al*. Learning phrase representations using RNN encoder-decoder for statistical machine translation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha: Association for Computational Linguistics, 2014. 1724–1734.

14 Kartoun U. A methodology to generate virtual patient repositories. arXiv:1608.00570, 2016.

15 Miotto R, Li L, Kidd BA, *et al*. Deep patient: An unsupervised representation to predict the future of patients from the electronic health records. Scientific Reports, 2016, 6: 26094. [doi: 10.1038/srep26094]

16 国家卫生健康委员会. 胃癌诊疗规范 (2018 年版). 中华消化病与影像杂志 (电子版), 2019, 9(3): 118–144.

(校对责编: 孙君艳)