深度神经网络在不规则弥漫大 B 细胞淋巴瘤时间序列数据分类预测中的应用*

李　琼[1,2]　张岩波[1,2,3]　余红梅[1,2,3]　周　洁[4]　赵艳琳[1,2]　李雪玲[1,2]　王俊霞[1,2]　张高源[1,2]
乔　宇[1,2]　赵志强[5△]　罗艳虹[1,2,3△]

【提　要】 **目的**　探讨深度神经网络在不规则时间序列数据中的分类效果，并对山西某医院 2014-2020 年 362 例弥漫大 B 细胞淋巴瘤（diffuse large B-cell lymphoma，DLBCL）患者进行复发预测。**方法**　回顾性地收集了确诊且治疗后达到完全缓解的 362 例 DLBCL 患者的病例资料，并预测其两年内的复发。先利用 LASSO 回归进行变量的筛选，再构建基于 GRU-ODE-Bayes（gated recurrent unirt-ordinary differential equation-Bayes）的不规则时间序列深度神经网络模型，并与传统模型及其他深度神经网络模型进行比较。**结果**　在本文的所有模型中，传统模型的分类性能不及深度神经网络模型。其中 GRU-ODE-Bayes 模型最优，其 AUC 为 0.85，灵敏度为 0.84，特异度为 0.71，*G*-means 为 0.77。**结论**　关于不规则 DLBCL 时间序列数据，与本文其他模型相比，GRU-ODE-Bayes 模型可以更精准地预测 DLBCL 患者的复发情况，可为患者个性化治疗和医生决策提供参考。

【关键词】 弥漫大 B 细胞淋巴瘤　不规则时间序列数据　复发预测　深度神经网络
【中图分类号】 R195.1　　【文献标识码】 A　　**DOI** 10.11783/j.issn.1002-3674.2024.02.006

Application of Deep Neural Networks into Classification in Irregular Time Series Data of Patients with Diffuse Large B-cell Lymphoma

Li Qiong，Zhang Yanbo，Yu Hongmei，et al（*Department of Health Statistics*，*School of Public Health*，*Shanxi Medical University*（030001），*Taiyuan*）

【**Abstract**】 **Objective**　To investigate the classification effect of deep neural networks in irregular time series data，and to predict the recurrence risk of 362 patients with diffuse large B-cell lymphoma（DLBCL）in a hospital in Shanxi from 2014 to 2020. **Methods**　A total of 362 diagnosed DLBCL patients who achieved complete remission after initial chemotherapy were collected retrospectively，and the recurrence risk was predicted within the next two years. First，LASSO regression was used to screen the variables. Then a deep neural network model of irregular time series data based on GRU-ODE-Bayes was constructed and compared with some traditional models and other deep neural network models. **Results**　Among all the models under study，the traditional models do not perform as well as the deep neural network models in classification. The GRU-ODE-Bayes model was the best，with AUC of 0.85，sensitivity of 0.84，specificity of 0.71，and *G*-means of 0.77. **Conclusion**　Compared with other models，the GRU-ODE-Bayes model can predict the recurrence of DLBCL patients more accurately. It could benefit the individualized treatment for patients and decision-making for physicians.

【**Key words**】 Diffuse large B-cell lymphoma；Irregular time series data；Recurrence risk prediction；Deep neural networks

弥漫大 B 细胞淋巴瘤（diffuse large B-cell lymphoma，DLBCL）是一种血液系统的恶性肿瘤，是非霍奇金淋巴瘤（non-Hodgkinlymphoma，NHL）中最常见的亚型，占 NHL 的 45.8%[1]。在中国，每年罹患 DLBCL 者约 8.4 万人，死亡人数约 4.7 万人，是发病率增长速度最快的恶性肿瘤之一[2]。目前，通过一线疗法治疗后患者反应不尽相同，约 70%的患者能够达到完全缓解，剩下 30%的患者会在达到完全缓解后复发，进而发展为难治性疾病[3]。因此，亟需根据患者的特征及疾病相关信息预测患者预后情况，以便在早期制定个性化的治疗方案。

现有的弥漫大 B 细胞淋巴瘤患者复发的预测研究[4-7]，大多是基于静态数据，没有充分利用动态的电子病例数据。不同于传统的流行病队列研究和临床试验数据，电子病历数据受临床评估结果、患者健康状况及临床治疗费用等影响，导致收集的数据具有稀疏性和不规则性[8]。对于这种不规则的时间序列数据，采取经典的时间序列分析方法，如结合了自回归滑动平均（auto-regressive moving average，ARMA）模型与 logistic 模型的方法[9]是存在问题的[10]。针对这种不规则时间序列数据主要有两种建模方式：边际模型[11]和广义线性混合效应模型[12]，但这些传统方法在复杂数据建模过程中需要满足很多要求，特别是训练集很小时预测效果很不好，并且不能做分类预测，大多数用来做趋势变化分析[13]和效果分析[14-15]。近年来，基于深度学习的方法备受青睐，卷积神经网

* 基金项目：山西省科技厅应用基础研究计划面上项目（202103021224245）；国家自然科学基金青年科学基金（81502897；82273742）；山西医科大学博士启动基金（BS2017029）
1.山西医科大学公共卫生学院卫生统计教研室（030001）
2.重大疾病风险评估山西省重点实验室
3.煤炭环境致病与防治教育部重点实验室
4.山西省肿瘤医院核医学 PET/CT 中心
5.山西省肿瘤医院血液科
△通信作者：罗艳虹，E-mail：lifearena@ 163.com；赵志强，E-mail：zqzhao69@ 163.com

络、循环神经网络及自动编码机等方法对时间序列数据有较好的适应性且具有能够主动学习特征，表现出良好的分类性能，但仅能处理均匀分布的规则时间序列数据。因此，Brouwer 等[16]提出了一种新的深度神经网络学习方法——GRU-ODE-Bayes（gated recurrent unit-ordinary differential equation-Bayes），该方法结合了基于神经常微分方程的门控循环单元（gated recurrent unit - ordinary differential equation，GRU - ODE）和处理零星观测的贝叶斯更新网络（GRU-Bayes），适应于分布不均匀、观察时间间隔不等的时间序列数据。本研究探讨了对于不规则采样的时间序列数据，GRU-ODE-Bayes 与其他传统模型及深度神经网络模型[17-19]的分类效果的比较，并构建 DLBCL 患者复发预测模型。

资料与方法

1.资料来源

根据《中国弥漫大 B 细胞淋巴瘤诊断与治疗指南 2013 版》[20]，回顾性地收集了山西某医院 2014-2020 年确诊的 603 例 DLBCL 患者的病例资料，包括人口统计资料与实验室相关检查资料，如性别、年龄、疾病分期等 56 个变量。纳入标准：①山西某医院 2014-2020 年确诊；②在治疗后达到完全缓解（complete remission，CR）；③有两次及两次以上就诊经历。排除随访中临床资料不全者。在 603 例 DLBCL 患者中，治疗后达 CR 有 362 例（60.0%），部分缓解有 181 例（30.0%），稳定有 41 例（6.8%），进展有 19 例（3.2%）。复发是指在达到完全缓解后复发。经纳入和排除，最终得到符合研究标准的患者有 362 例。根据完全缓解后两年内是否复发，分为复发组（82 例）和未复发组（280 例）。

每个患者获取两种类型数据，分别为静态数据（首次住院的年龄、性别等）和时间相关数据（随访中的 WBC、LDH 等）。研究将 1 个患者视为 1 条时间序列，就诊次数是时间序列的长度，共 362 条时间序列，最长为 11，最短为 2。因患者就诊间隔时间不相等，每条时间序列是非等间距的。故不规则 DLBCL 时间序列是长度不一且非等间距的时间序列。

2.方法及原理

（1）GRU-ODE-Bayes 模型

研究 N 个随机观测的 D 维时间序列，例如，来自 N 个病人的 D 个纵向变量被观测。每个时间序列 $i \in \{1,\cdots, N\}$ 在 K_i 时间点上测量，该时间点由观测次数 $t \in \mathbb{R}K_i$ 的向量指定。这些观测值由观测 $y_i \in \mathbb{R}^{K_i \times D}$ 和观测掩码 $m_i \in \{0, 1\}^{K_i \times D}$ 组成的矩阵指定（表示在每个时间点测量哪些变量）。模型提出一个非学习过程，简单地将 y_i 与 i 连接起来，从（i，y_i）中构造出一条通道 X，对应于这些 i 的 X 通道就是观测强度。

假设观测值 y_i 来自于一个 D 维随机过程 $Y(t)$ 的实现，该过程的动力学由一个未知的随机微分方程（stochastic differential equation，SDE）驱动：

$$dY(t)=\mu(Y(t))dt+\sigma(Y(t))dW(t)$$

其中 $dW(t)$ 为 Wiener 过程，$Y(t)$ 的分布是根据 Fokker-Planck 方程演化，将其概率密度函数的均值和协方差参数称为 $\mu_Y(t)$ 和 $\Sigma_Y(t)$。通过对随机向量 $Y(t)$ 在一些观测噪声 $\in$ 的 t_i 时刻进行采样，可以从零星的测量数据 y_i 中模拟未知的时间函数 $\mu_Y(t)$ 和 $\Sigma_Y(t)$。

GRU-ODE 是用于演化连续时间内的隐藏状态 $h(t)$ 的模块。首先将提出的门控循环单元 GRU 写成一个差分方程。设 r_t，z_t 和 g_t 分别为 GRU 的复位门，更新门和更新向量：

$$r_t=\sigma(W_r x_t+U_r h_{t-1}+b_r)$$
$$z_t=\sigma(W_z x_t+U_z h_{t-1}+b_z)$$
$$g_t=tanh(W_h x_t+U_r(r_t\odot h_{t-1}+b_h)$$

其中，$\odot$ 是元素乘积，那么 GRU 的隐藏状态 h 标准更新为：

$$h_t=z_t\odot h_{t-1}+(1-z_t)\odot g_t$$

最终得到差分方程：

$$\Delta h_t=h_t-h_{t-1}=z_t\odot h_{t-1}+(1-z_t)\odot g_t-h_{t-1}=(1-z_t)\odot(g_t-h_{t-1})$$

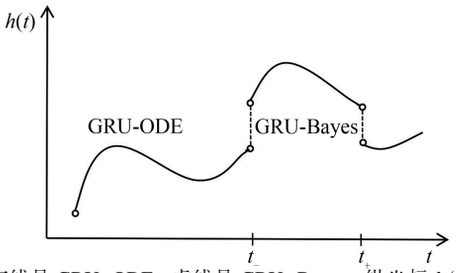
这个差分方程可以自然地推导出常微分方程 ODE：

$$\frac{dh(t)}{dt}=(1-z(t))\odot(g(t)-h(t))$$

GRU-Bayes 是处理零星观测数据以更新隐向量的模块。为了向 GRU-Bayes 内部的 GRU 单元提供一个非完全观测向量，首先使用 f_{prep} 对其进行预处理。对于给定的时间序列，可以将基于观测 y 的隐藏状态从 $h(t_-)$ 变换为 $h(t_+)$：

$$h(t_+)=GRU(h(t_-), f_{prep}(y[k], m[k], h(t_-)))$$

GRU-ODE-Bayes 使用 GRU-ODE 在两个观测时间之间演化隐藏状态，使用 GRU-Bayes 以离散方式处理观测值并更新隐藏变量 h，并且两者是交替进行的（图 1）。



*：实线是 GRU-ODE，虚线是 GRU-Bayes，纵坐标 $h(t)$ 表示隐藏过程，横坐标 t 表示时间。

图 1 GRU-ODE-Bayes 模型示意

GRU-ODE-Bayes 模型结合了 GRU-ODE 和 GRU-Bayes 两种模块，可以更自然地处理零星数据，更精准地模拟观测特征之间的动态和相关性，这使得

预测性能高于其他方法。当样本量稀少时，嵌入 GRU-ODE 中的连续性先验发挥了至关重要的作用。

(2)其他对比模型

①传统模型

logistic 回归是一种分类学习方法，无需事先假设数据分布，直接对分类可能性进行建模；决策树是基于树结构进行决策的，采用自顶向下的递归方法，其分类依据为信息增益，通过信息增益最大字段对样本数据分割；随机森林在以决策树为基学习器构建 Bagging 集成的基础上，进一步在决策树的训练过程中引入了随机属性选择；AdaBoost 针对同一个训练集训练不同的弱分类器，将这些弱分类器集合起来，构成一个强分类器。在传统模型建模之前使用均值法填补缺失值。

②深度神经网络模型

Neural CDE（neural controlled differential equations）模型通过控制常微分方程使其直接适用于部分观测的不规则采样多元时间序列数据；ODE-RNN（ordinary differential equations-recurrent neural networks）模型将标准的循环神经网络推广为具有由 ODE 定义的连续时间隐藏动力学模型，可以自然地处理观测之间的时间间隔，使其可以用于稀疏和/或不规则的数据；GRU-Δt（gated recurrent unit-Δt）模型使用观测值之间的时间差作为输入的 GRU；GRU-D（gated recurrent unit-decay）模型通过对 GRU 的输入和网络状态施加掩蔽和时间间隔来捕获观测和它们之间的依赖关系，并使用反向传播联合训练所有模型组件，适用于有缺失数据的时间序列分类问题。

3.模型预测结果评价指标

将数据集划分为 80% 的训练集和 20% 的测试集，重复抽样并构建模型 100 次，最终各评价指标的结果取 100 次的均值。采用 ROC 曲线下面积（area under the ROC curve，AUC）、灵敏度、特异度、*G*-means 值和 Brier-Score 作为评价指标，其中 AUC 反映模型的区分度，其值越大代表模型分类正确的可能性越大，即区分复发和未复发患者的能力越大；Brier-Score 反映模型的校准度，是模型预测复发和实际复发之间的均方误差，其值越小说明模型的预测误差越小；*G*-means 综合考虑了少数类和多数类的分类性能，只有当两者都较高时，其值才会较高。

结　果

根据 LASSO 回归分析结果、临床医生意见及查阅相关文献，最终筛选出 19 个静态变量和 4 个时间相关变量，具体信息见表 1。

使用 GRU-ODE-Bayes 和其他模型对变量筛选后的 DLBCL 数据分别建模训练，得到各模型预测的 AUC、灵敏度、特异度、*G*-means 值及 Brier-Score。各模型的分类预测结果见表 2。

表 1　362 例 DLBCL 患者的 23 个变量及赋值

	变量	赋值	例数	构成比(%)
静态变量	确诊年龄	0="<60 周岁"	207	57.2
		1="≥60 周岁"	155	42.8
	性别	0=男	188	51.9
		1=女	174	48.1
	疾病分期	1=Ⅰ级	42	11.6
		2=Ⅱ级	126	34.8
		3=Ⅲ级	71	19.6
		4=Ⅳ级	123	34.0
	IPI 得分	0="<3 分"	297	82.0
		1="≥3 分"	65	18.0
	Ki-67	0="≤80%"	177	48.9
		1=">80%"	185	51.1
	BCL6	0=阴性	209	57.7
		1=阳性	153	42.3
	CD3	0=阴性	306	84.5
		1=阳性	56	15.5
	CD10	0=阴性	300	82.9
		1=阳性	62	17.1
	CD20	0=阴性	335	92.5
		1=阳性	27	7.5
	MPO	0=阴性	301	83.1
		1=阳性	61	16.9
	PAX5	0=阴性	297	82.0
		1=阳性	65	18.0
	Vim	0=阴性	337	93.1
		1=阳性	25	6.9
	是否鼻转移	0=否	352	97.2
		1=是	10	2.8
	肿瘤长径	1="<3cm"	194	53.6
		2="3~6cm"	102	28.2
		3="6~9cm"	39	10.8
		4="9~12cm"	15	4.1
		5="≥12cm"	12	3.3
	侵犯数量	0=0 个	162	44.8
		1=1 个	63	17.4
		2=2 个	45	12.4
		3=3 个	29	8.0
		4=4 个	26	7.2
		5=5 个	14	3.9
		6=6 个	12	3.3
		7=7 个	11	3.0
	是否出现发热	0=否	336	92.8
		1=是	26	7.2
	是否出现消化道反应	0=否	268	74.0
		1=是	94	26.0
	是否出现上呼吸道反应	0=否	262	72.4
		1=是	100	27.6
	是否使用 R-CHOP	0=否	311	85.9
		1=是	51	14.1
时间相关变量	WBC	0=正常		
		1=偏高		
		2=偏低		
	LDH	0=正常		
		1=偏高		
		2=偏低		
	β_2-MG	0=正常		
		1=偏高		
		2=偏低		
	ESR	0=正常		
		1=偏高		
		2=偏低		

*：362 例患者中，R-CHOP 有 51 例，非 R-CHOP 有 311 例（其中 R-CTOP 有 66 例，CHOP-E 有 63 例，CTOP 有 62 例，CHOP 有 46 例，CTOP-E 有 25 例，R-CDOP 有 9 例，R-CEOP 有 6 例，R-CHOPE 有 6 例，CDOP 有 4 例，其他有 24 例）。

表 2 模型性能评价(均值±标准差)

模型(测试)	AUC	灵敏度	特异度	*G*-means	Brier-Score
logistic 回归	0.66±0.07	0.55±0.14	0.67±0.07	0.60±0.08	0.36±0.06
决策树	0.60±0.02	0.58±0.05	0.60±0.01	0.59±0.02	0.40±0.01
Adaboost	0.62±0.01	0.64±0.03	0.65±0.01	0.64±0.02	0.36±0.01
随机森林	0.72±0.02	0.79±0.04	0.51±0.07	0.64±0.04	0.27±0.03
Neural CDE	0.84±0.06	0.79±0.13	0.71±0.09	0.74±0.06	0.27±0.06
ODE-RNN	0.81±0.04	0.79±0.12	0.70±0.07	0.74±0.05	0.28±0.04
GRU-Δt	0.80±0.05	0.78±0.14	0.69±0.09	0.73±0.07	0.29±0.05
GRU-D	0.80±0.04	0.74±0.10	0.72±0.05*	0.73±0.05	0.27±0.03
GRU-ODE-Bayes	0.85±0.03*	0.84±0.08*	0.71±0.07	0.77±0.04*	0.26±0.05*

*：表示在所有模型中性能最优。

结果如表 2 所示，相比其他 8 种模型，GRU-ODE-Bayes 具有良好的分类预测效果(AUC=0.85，灵敏度=0.84，特异度=0.71，*G*-Means=0.77，Brier-Score=0.26)，证明其可以处理不规则的时间序列数据，并且适合于小样本，可为 DLBCL 患者预后分类提供一定帮助。

讨　论

许多的时间序列分析方法都假定变量是按固定的时间间隔测量的，但真实世界的临床数据可能是零星的。因为这些数据违反了传统机器学习方法的主要假设，所以建模变得很有挑战。为了解决不规则采样问题，一种流行的方法是将观察结果重新放入固定连续事件中，然而，这种表示方式会导致在时间和特征维度上都缺失观察，使得直接使用神经网络架构变得棘手。最近，Brouwer 等[16]提出的 GRU-ODE-Bayes 是一种结合了 GRU-ODE 和 GRU-Bayes 两种新技术的模型，它允许将零星的观测数据输入到连续的 ODE 动态中，不需要预处理数据缺失问题，描述了数据的概率分布的演化，编码了潜在过程的连续性先验使其适合于小样本数据。这对现实世界的临床数据是十分重要的，因为许多数据集的规模仍然相对较小而且可能存在大量缺失。

本研究比较了 logistic 回归、决策树、随机森林、Adaboost、Neural CDE、ODE-RNN、GRU-Δt、GRU-D、GRU-ODE-Bayes9 种模型的分类预测能力，研究结果表明 GRU-ODE-Bayes 适用于不规则采样多元时间序列数据的分类任务，尤其适用于小样本，并且能够更精准地预测患者治疗达到完全缓解后两年内是否复发。在其他研究中，Lin 等[21]通过预测 ICU 再入院，比较了 logistic 回归、随机森林等传统学习方法和卷积神经网络、长短期记忆网络这些深度学习方法的预测性能，表明了深度学习模型具有更高的准确性和敏感性，这是由于传统方法无法正确捕捉时间序列中的图表事件特征；Chu 等[22]利用电子健康记录预测心力衰竭再入院，结果表明所有时间序列模型(长短期记忆网络、门控循环单元等)都明显优于非时间序列模型(logistic 回归、支持向量机等)，这是因为非序列模型缺乏处理嵌入在患者治疗轨迹中的序列相关性的能力；还有研究[23]将时间序列数据和事件时间数据联合建模，分析发生癌症的风险。

本研究的不足之处：第一，研究只考虑了二分类结局，没有考虑生存时间；第二，构建的 DLBCL 患者复发风险预测模型的性能还有待提升，这可能与 DLBCL 数据不平衡有关。我们下一步的研究将尝试处理时间序列数据的不平衡问题，并且考虑对复发事件和生存时间联合建模。

参 考 文 献

[1] Intragumtornchai T, Bunworasate U, Wudhikarn K, et al. Non-Hodgkin lymphoma in South East Asia: An analysis of the histopathology, clinical features, and survival from Thailand. Hematological Oncology, 2018, 36(1): 28-36.

[2] 张晓娟, 杜伟, 郭树霞. 弥漫性大 B 细胞淋巴瘤组织 MYC 和 Bcl-2 及 Bcl-6 检测的预后价值. 中华肿瘤防治杂志, 2015, 22(15): 1193-1197.

[3] Parvez A, Tau N, Hussey D, et al. 18F-FDG PET/CT metabolic tumor parameters and radiomics features in aggressive non-Hodgkin′s lymphoma as predictors of treatment outcome and survival. Annals of Nuclear Medicine, 2018, 32(6): 410-416.

[4] 王蕾, 赵志强, 余红梅,等. 基于重采样和集成学习的弥漫大 B 细胞淋巴瘤患者复发风险预测模型. 中国卫生统计, 2019, 36(4): 588-592.

[5] 冯帆, 吴可, 张旭,等. 中期 PET-CT 检查在复发性弥漫大 B 细胞淋巴瘤预后分析中的作用. 中国肿瘤临床, 2018, 45(16): 844-849.

[6] 张会平, 徐瑞荣. 弥漫大 B 细胞淋巴瘤患者化疗后复发相关因素分析. 中国肿瘤临床与康复, 2019, 26(1): 76-78.

[7] 张静, 顾岩, 吴雪,等. 利妥昔单抗联合 CHOP/EPOCH 方案治疗弥漫大 B 细胞淋巴瘤患者的难治复发相关因素分析. 中国实验血液学杂志, 2020, 28(6): 1912-1918.

[8] Sun Y, McCulloch CE, Marr KA, et al. Recurrent Events Analysis With Data Collected at Informative Clinical Visits in Electronic Health Records. Journal of the American Statistical Association, 2021, 116(534): 594-604.

[9] 杰弗里·M·伍德里奇.计量经济学导论. 北京: 中国人民大学出版社.

[10] 刘艳, 李扬, 刘罡,等. 纵向有序数据的临床疗效评价方法应用研究. 中国卫生统计, 2017, 34(1): 74-77+81.

(下转第 199 页)