Statistics and Application 统计学与应用, 2022, 11(4), 878-884
Published Online August 2022 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2022.114091

Hans汉斯

基于深度学习的实际生存问题应用研究

张晓彤

北方工业大学，北京

收稿日期：2022年7月21日；录用日期：2022年8月1日；发布日期：2022年8月15日

摘　要

疾病是自古以来一直困扰着所有人类健康甚至是生命的重大难题，生存分析是一种可以模拟患者生存的方法，可以了解感兴趣事件和协变量之间的关系，比如某个癌症病人的死亡时间和他的年龄、性别等协变量的关系。近年来，生存分析的应用越来越广泛，不仅在医院方面，还在电子商务、广告、电信和金融服务等其他行业也获得了很大的发展，通过生存分析方法可以让这些公司更好地了解客户何时购买产品，何时会流失客户，何时会拖欠贷款等。本文使用一种基于深度学习的生存分析模型DeepHit模型处理真实的数据集并与其他模型进行对比，发现DeepHit模型效果良好。

关键词

生存分析，真实数据，深度学习，DeepHit模型

Application Research on Practical Survival Problems Based on Deep Learning

Xiaotong Zhang

North China University of Technology, Beijing

Received: Jul. 21st, 2022; accepted: Aug. 1st, 2022; published: Aug. 15th, 2022

Abstract

Disease is a major problem that has plagued all human health and even life since ancient times. Survival analysis is a method that can simulate the survival of patients, and can understand the relationship between interested events and covariates, such as the relationship between the death time of a cancer patient and his age, gender and other covariates. In recent years, the application of survival analysis has become more and more extensive. It has also achieved great development not only in hospitals, but also in other industries such as e-commerce, advertising, telecommunications and financial services. Through survival analysis, these companies can better understand

文章引用: 张晓彤. 基于深度学习的实际生存问题应用研究[J]. 统计学与应用, 2022, 11(4): 878-884.
DOI: 10.12677/sa.2022.114091

when customers buy products, when they will lose customers, and when they will default on loans. This paper uses a deep learning based survival analysis model, DeepHit model, to process the real data set and compare it with other models. It is found that DeepHit model has a good effect.

Keywords

Survival Analysis, Real Data, Deep Learning, DeepHit Model

1. 背景

传统的用于生存分析的方法有 Kaplan-Meier 算法、Cox 比例危险率模型、线性回归模型、位置－刻度回归模型，竞争风险模型等。KM 模型利用绘制生存曲线估算生存函数。优点是能够学习非常灵活的生存曲线，但缺点是不纳入患者协变量，在整体层面有用但在个人层面没用。Cox 比例危险率模型是一种半参数回归模型，可以纳入患者的协变量但其假设两个人的危险函数之比与时间无关，但由于危险函数时间成分不明确，使得在实际问题中效果较差。竞争风险模型[1]适用于多个终点的生存数据，是一种处理多种潜在结局生存数据的分析方法，通过计算每个结局的累积发生率函数(Cumulative Incidences Function, CIF)进行分析。

由于之前许多对于生存分析的方法例如 Cox 比率危险率模型是通过将生存时间视为随机过程的第一次达到时间来解决协变量与生存时间的关系问题，并假设随机过程的特定格式。本文采取了一种完全不同的生存分析方法 DeepHit [2]，它不对潜在随机过程进行假设，使用深度神经网络直接学习生存时间的分布。并使用累计发生率函数 CIF 作为指标函数，与 DSM [3] (用于以完全参数化的方式使用删失数据估计事件时间预测问题中的相对风险)，DeepSurv [4] (半参数模型假设基础风险恒定)等模型进行对比，发现本文提出的模型的性能较好，耗时也较短。

2. 数据及模型介绍

2.1. 数据来源

SUPPORT 数据集(表 1)来自一项以预测 9105 名重症住院患者在 180 天内的生存率的研究。在 9105 名患者中，6201 (68.1%)名患者被随访直至死亡，生存时间中位数为 58 天，平均生存时间 478.45 天。SUPPORT 数据含括了年龄、性别、种族等 30 个代表患者信息的协变量。

Table 1. SUPPORT data
表 1. SUPPORT 数据

sno	age	death	sex	hospdead	slos	d.time	dzgroup	dzclass	num.co	edu
1	62.84998	0	male	0	5	2029	Lung Cancer	Cancer	0	11
2	60.33899	1	female	1	4	4	Cirrhosis	COPD/CHF/Cirrhosis	2	12
3	52.74698	1	female	0	17	47	Cirrhosis	COPD/CHF/Cirrhosis	2	12
4	42.38498	1	female	0	3	133	Lung Cancer	Cancer	2	11

Continued

5	79.88495	0	female	0	16	2029	ARF/MOSF w/Sepsis	ARF/MOSF	1	
6	93.01599	1	male	1	4	4	Coma	Coma	1	14
7	62.37097	1	male	0	9	659	CHF	COPD/CHF/Cirrhosis	1	14
8	86.83899	1	male	0	7	142	CHF	COPD/CHF/Cirrhosis	3	
9	85.65594	1	male	0	12	63	Lung Cancer	Cancer	2	12
10	42.25897	1	female	0	8	370	Colon Cancer	Cancer	0	11

2.2. 数据预处理

生存类数据主要提供患者的几种信息：首先，生存数据提供观察到的影响患者生存时间的其它影响因素即协变量的具体信息例如年龄、性别等；其次，生存数据提供自收集这些协变量信息所度过的时间；并且生存数据提供发生的具体事件原因例如患者的死亡或者其他事件的标签；最后生存数据是最为真实的数据，通过对生存数据的分析可以帮助人们解决生活中具体遇到的实际问题。

由于一些原因，某些协变量含有缺失值，对于缺失值，我们使用 python 语言中的 Simple Imputer 函数的 mean 方法即使用该列的均值代替缺失值的方法。

将生存时间 T 视为离散且范围有限，将引发事件的原因视为 K 个可能的感兴趣的事件，并且由于并不能总是观察到事件的发生如患者失访即发生删失，我们将这种右删失记为 $\varnothing$ 可用 0 表示，此时可将引发事件的原因 K，假设引发最终事件如患者死亡有且只由一个原因导致发生。此时每个患者的信息可以由 X 表示，x 是协变量 X 的向量，s 是事件发生或删失的时间，k 是在 s 时发生的事件或删失。

2.3. 模型介绍

对于未删失样本，我们感兴趣的是概率 $P=\left(s=s^*,k=k^*\middle|x=x^*\right)$，即具有协变量 x^* 的患者在时间 s^* 经历事件 k^* 的概率。比如具有年龄、性别等协变量的一个患者在手术后 100 天后这个时刻，因癌症死亡的概率。由于真实概率未知，所以利用深度学习来学习 $\hat{P}$，即发生事件的时间和竞争风险的联合概率分布的估计。

DeepHit 模型通过训练神经网络学习估计事件和时间的联合分布。生存模型由一个共享网络和 K 个特定原因的子网络组成，并使用 softmax 层作为最终的输出层输出模型学习的 K 个竞争事件的联合分布和每个原因的边缘分布如图 1 所示。

共享网络由全连接层(Fully Connected Layers, FC)构成。全连接层在卷积神经网络 CNN 中起着相当于“分类器”的作用，全连接层可将输入的数据特征表示一一映射到样本标记空间的作用。全连接层前向计算时，是一个线性的加权过程，全连接层的输出可以看作是前一层的每一个神经元与权重系数 W 的乘积加上一个 bias 所得到。

例如当 X_1,X_2,X_3 作为全连接层的输入时，$\alpha_1,\alpha_2,\alpha_3$ 作为全连接层的输出，此时就有：

$$\alpha_1=W_{11}*X_1+W_{12}*X_2+W_{13}*X_3+b_1$$
$$\alpha_2=W_{21}*X_1+W_{22}*X_2+W_{23}*X_3+b_2$$
$$\alpha_3=W_{31}*X_1+W_{32}*X_2+W_{33}*X_3+b_3$$

在实际模型进行生存分析时，由于患者可能由于 K 种风险发生感兴趣的事件，而患者又具有多种的协变量 X，所以全连接层可以将患者的特征整合到一起，输出以某一个风险 K 下的可以表达特征的数值。

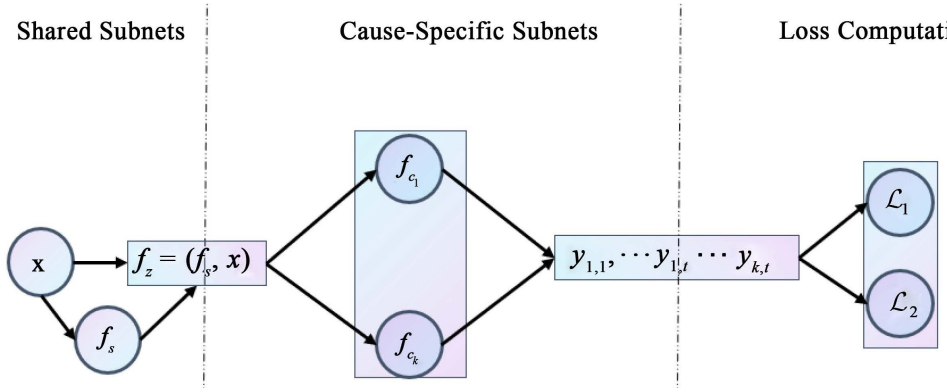

Figure 1. Model structure
图 1. 模型结构图

首先将患者的协变量 x 作为输入，带入共享网络层，产生一个具有 K 个竞争事件的潜在因素的向量，并与向量 x 共同组成共享子网络的输出 $z=\left(f_s(x),x\right)$，z 对应于特定原因 K 的第一次产生事件的时间的概率。

K 个特定原因子网络层将 $z=\left(f_s(x),x\right)$ 作为输入，学习协变量共有表示的向量 $f_s(x)$ 和潜在因素，输出特定原因 K 的第一次命中的时间的概率，这些输出的汇总是在首次命中事件和时间上的联合概率分布，病因特异性子网并行学习每个病因的首次命中时间的边缘分布。

累积发生函数 CIF 表示为具有协变量 x^* 的患者在时间 t^* 或之前发生特定事件 $K\in K^*$ 的概率，在具有竞争风险的生存分析中，找到 CIF 是关键一步。

$$F_{k^*}\left(t^*\middle|x^*\right)=P\left(s\le t^*,k=k^*\middle|x=x^*\right)$$
$$=\sum_{s^*=0}^{t^*}P\left(s=s^*,k=k^*\middle|x=x^*\right)$$

由于实际的 CIF 是未知的，所以替代为 $\hat{F}_{k^*}\left(s^*\middle|x^*\right)=\sum_{t=0}^{s^*}y_{k,t}^*$。

3. 实证分析

主要使用 pytorch-lighting 运行结果，它相较于 pytorch 可以更容易地识别和理解代码，简化了模型结构并且实现代码自动化，可以更加简洁的构建深度学习代码。

将处理过的 SUPPORT 数据集带入模型中进行训练，部分参数如表 2 所见，使用 train_test_split 并设置参数为 0.2 划分测试集和训练集，使用了 ReLU 激活函数，学习率设置为 0.0003 等。

Table 2. Model parameter
表 2. 模型参数

accumulate_grad_batches:	1	min_epochs:	1
activation:	relu	model_depth:	50
alpha:	1	num_layers_CS:	2
batch_size:	16	num_layers_shared:	2
beta:	0.5	num_nodes:	1

Continued

check_val_every_n_epoch:	1	num_processes:	1
dropout:	0.5	num_sanity_val_steps:	2
gpus:	1	num_workers:	0
gradient_clip_val:	0	out_dim:	128
hidden_dim_CS:	128	overfit_batches:	0
hidden_dim_shared:	128	precision:	32
limit_test_batches:	1	process_position:	0
limit_train_batches:	1	progress_bar_refresh_rate:	0
limit_val_batches:	1	row_log_interval:	50
log_save_interval:	100	track_grad_norm:	−1
lr:	0.0003	val_check_interval:	1
max_epochs:	100	weight_decay:	1.00E−05

在模型训练时，在 termin 终端设置 precision 为 16，batch_size 为 16，max_epochs 为 100 进行训练，模型训练结果显示随着训练的进行 CIF (图 2)呈上升趋势，并趋于稳定，在训练中平均 CIF 为 0.923 (0.892~0.953)。

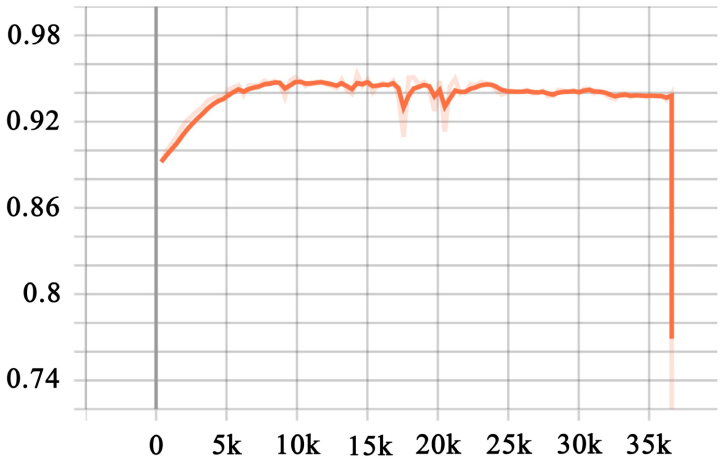

Figure 2. CIF
图 2. CIF 图

表示模型效果良好。由训练损失(图 3)和验证损失(图 4)曲线可以看出，随着训练的进行，验证损失和训练损失都呈下降趋势，虽然训练损失在下降至 5 左右波动，验证损失则下降至 50 趋向平稳，这表明训练和验证表现良好，且模型准确度也呈上升趋势且超过 90%，证明模型表现良好，适用于分析此数据集。

将模型结果与 DSM 模型进行对比：将数据在 DSM 模型，在训练中平均 CIF 为 0.832 (0.830~0.834)使用 DeepSurv 模型进行训练得到结果平均 CIF 为 0.805 (0.801~0.809)与 DeepHit 模型结果进行对比，发现 DeepHit 模型 CIF 评分高于其它模型，证明本文模型效果优于其它模型。

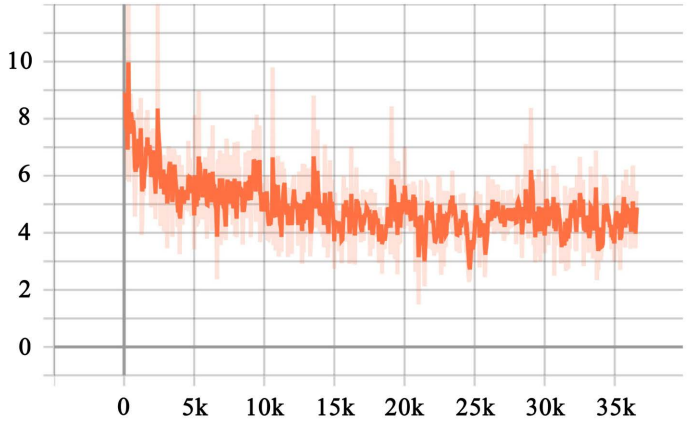

Figure 3. Train loss
图 3. 训练损失图

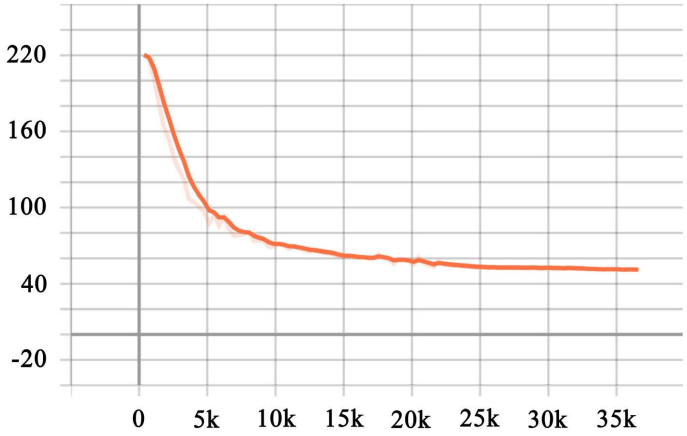

Figure 4. Val loss
图 4. 验证损失图

4. 结论

DeepHit 是一种基于深度学习的生存分析方法，本位通过使用 DeepHit 模型处理 SUPPORT 真实数据集，模型通过使用深度学习方法，利用共享子网络层和特定因子层直接学习了生存时间和生存事件的联合分布，并对其进行估计并推导出 CIF 的估计值。

通过训练发现模型结果 CIF 达到 0.9 以上，证明模型对于数据集来说效果良好，且训练损失下降到 5 左右波动，验证损失也下降到一定值后趋于平稳，说明模型拟合效果较好。通过将结果与 DSM 等模型进行对比发现在使用 DeepHit 训练数据时，模型 CIF 优于 DSM 等模型，这说明模型在预测方面有着更好的效果，可以更广泛地应用于生存分析问题中。

参考文献

[1] Fine, J.P. and Gray, R.J. (1999) A Proportional Hazards Model for the Subdistribution of a Competing Risk. *Journal of the American Statistical Association*, **94**, 496-509. https://doi.org/10.1080/01621459.1999.10474144

[2] Lee, C., Zame, W.R., Yoon, J., *et al.* (2018) DeepHit: A Deep Learning Approach to Survival Analysis with Competing Risks. *Proceedings of the AAAI Conference on Artificial Intelligence*, **32**. https://doi.org/10.1609/aaai.v32i1.11842

[3] Nagpal, C., Li, X. and Dubrawski, A. (2021) Deep Survival Machines: Fully Parametric Survival Regression and Representation Learning for Censored Data with Competing Risks. *IEEE Journal of Biomedical and Health Informatics*,

25, 3163-3175. https://doi.org/10.1109/JBHI.2021.3052441

[4] Katzman, J.L., Shaham, U., Cloninger, A., *et al.* (2018) DeepSurv: Personalized Treatment Recommender System Using a Cox Proportional Hazards Deep Neural Network. *BMC Medical Research Methodology*, **18**, Article No. 24. https://doi.org/10.1186/s12874-018-0482-1