doi 10.12122/j.issn.1673-4254.2023.01.10 J South Med Univ, 2023, 43(1): 76-84

扩展Cox模型在非线性生存资料分析中的预测能力比较

陈雨轩[1],韦红霞[1],潘建红[2],安胜利[1]
[1]南方医科大学公共卫生学院生物统计学系,广东 广州 510515;[2]国家药品监督管理局药品审评中心,北京 100022

摘要:目的 系统性地比较两类扩展Cox模型的预测能力,观察它们应用于非线性生存数据中的预测能力优劣。**方法** 通过蒙特卡罗模拟和实证研究从预测能力方面研究比较限制性立方样条Cox模型(Cox_ RCS),深度生存神经网络Cox模型(Cox_DNN)这两种方法的优劣;并以传统Cox模型(Cox)和随机生存森林(RSF)作为参照。其中预测的区分度评价指标采用一致性指数(C-index),该指标越大,模型预测能力越好;预测的校准度评价指标采用积分布莱尔评分(IBS),该指标越小,模型预测能力越好。**结果** 在数据满足比例风险的情况下,无论样本量和删失率大小,Cox_RCS的预测能力都是最好的。在数据不满足比例风险的情况下,Cox_DNN的预测能力在大样本(本文中≥500)、低删失(本文中<40%)时是最优的,其余情况Cox_RCS的预测能力优于其他模型。在实例数据中,Cox_RCS的表现是最优。**结论** 在含有非线性关系的低维生存数据中,Cox_RCS和Cox_DNN在预测能力上各有优劣。因此可根据实际数据条件选择合适的分析方法,传统的生存分析方法在特定条件下并不差于机器学习以及深度学习方法。
关键词:生存分析;非线性关联;Cox模型;限制性立方样条;深度神经网络

Comparison of prediction ability of two extended Cox models in nonlinear survival data analysis

CHEN Yuxuan[1], WEI Hongxia[1], PAN Jianhong[2], AN Shengli[1]
[1]Department of Biostatistics, School of Public Health, Southern Medical University, Guangzhou 510515, China; [2]Center for Drug Evaluation, National Medical Products Administration, Beijing 100022, China

Abstract: Objective To compare the predictive ability of two extended Cox models in nonlinear survival data analysis. **Methods** Through Monte Carlo simulation and empirical study and with the conventional Cox Proportional Hazards model and Random Survival Forests as the reference models, we compared restricted cubic spline Cox model (Cox_RCS) and DeepSurv neural network Cox model (Cox_DNN) for their prediction ability in nonlinear survival data analysis. Concordance index was used to evaluate the differentiation of the prediction results (a larger concordance index indicates a better prediction ability of the model). Integrated Brier Score was used to evaluate the calibration degree of the prediction (a smaller index indicates a better prediction ability). **Results** For data that met requirement of the proportion risk, the Cox_RCS model had the best prediction ability regardless of the sample size or deletion rate. For data that failed to meet the proportion risk, the prediction ability of Cox_DNN was optimal for a large sample size (≥500) with a low deletion (<40%); the prediction ability of Cox_RCS was superior to those of other models in all other scenarios. For example data, the Cox_RCS model showed the best performance. **Conclusion** In analysis of nonlinear low maintenance data, Cox_RCS and Cox_DNN have their respective advantages and disadvantages in prediction. The conventional survival analysis methods are not inferior to machine learning or deep learning methods under certain conditions.
Keywords: survival analysis; nonlinear correlation; Cox model; restricted cubic spline; deep neural network

在医学研究中,研究者们经常通过构建回归模型来分析变量间的关系。通常回归模型的使用前提假设是自变量和因变量间或因变量的某种函数间呈线性关联,但在实际应用中这个条件并不一定满足。常见的解决方法是将连续型变量分类,但分类数目和节点位置的选择往往带有主观性,并且分类会损失信息。因此,一个更好的解决方法是拟合自变量与因变量之间的非线性关系。

以往生存分析中对非线性数据的处理通常有两种,一个是直接拟合机器学习模型;另一个就是把连续型的变量转换成分类变量并以哑变量形式纳入传统生存分析模型。但是后者可能导致协变量不满足比例风险假定[1]。为了弥补传统生存分析模型的应用条件限制及缺陷,避免对变量关系间的形状进行参数约束,目前常用平滑工具来解决这类问题。平滑工具中限制性立方样条(RCS)是目前分析非线性关系的最常见的方法之一[2-4]。生存分析中,限制性立方样条结合Cox模型,不仅可以用于探索非线性关系,越来越多的学者也将其运用于构建预测模型[5-8]。

近年来机器学习以及深度学习神经网络的发展,使得更多的方法可以用于解决复杂生存数据构建预测模型。深度学习属于机器学习的一个子域,机器学习方法

收稿日期:2022-07-14
基金项目:广东省自然科学基金(2022A1515012152);广东省组织构建与检测重点实验室开放基金(zzgjzd2021003)
作者简介:陈雨轩,在读硕士研究生,E-mail: chenyx9806@163.com
通信作者:安胜利,博士,副教授,E-mail: ASL0418@126.com

中，随机生存森林方法是最常见的分析方法，当有新的机器学习方法提出时一般会跟随机生存森林做比较，所以本文也考虑纳入随机生存森林作为参照方法。机器学习生存分析方法，包括随机生存森林、条件推断森林等[9-13]；深度学习神经网络方法如DeepSurv，DeepHit，Neural net-extended time-dependent cox model等[14]，都可以解决非线性、交互等复杂情况，其在高维数据中有明显的优势，并且不需要满足传统Cox回归模型比例风险假定等前提条件，但是机器学习以及神经网络模型实施起来较为复杂，特别是深度学习神经网络需要用Python软件来实现，其所花费的时间比传统模型要多。直接使用机器学习或者深度学习方法来分析，花费人力物力的同时其拟合效果也不一定比传统生存分析方法好。所以在非线性关系存在的生存数据分析中，各方法的优劣还有待研究比较。

1 方法

方法部分主要介绍Cox比例风险回归模型（Cox）、限制性立方样条Cox回归（Cox_RCS）、深度生存神经网络（Cox_DNN）这3种方法。

1.1 Cox比例风险回归模型

Cox比例风险回归模型是1972年由英国统计学家D.R.Cox提出的一种半参数模型[15]，其数学公式为：

$$\lambda(t,X_i)=\lambda_0(t)\exp(X_i\beta) \quad (1)$$

其中 $\lambda(t,X_i)$ 为受试者在一组预测变量 X_i 条件下所预测的t时刻的事件发生风险，$\lambda_0(t)$ 是基线风险函数，β 为预测变量的回归系数。医学研究人员常使用Cox模型评估预后协变量在死亡或疾病复发等事件中的重要性，并随后告知患者其治疗选择；Cox比例风险回归模型使用时需要满足两个基本假定：(1)比例风险假定：$\lambda(t,X_i)/\lambda_0(t)$ 为固定值，即协变量对生存率的影响不随时间的改变而改变；(2)对数线性假定：对数风险比应与模型中的连续型协变量呈线性关系。模型中的连续型变量可能会对生存时间产生非线性风险，如果不考虑这个风险，可能会导致结果偏差。

1.2 限制性立方样条Cox回归

限制性立方样条是最常用来检验假设关系不是线性的一种方法，或者用来总结非线性关系的一种方法。限制性立方样条函数只是一个自变量的变换。因此，它们不仅可以用于普通最小二乘回归，还可以用于逻辑回归、生存分析等。

临床评估科学研究所进行的两项研究很好地说明了限制性立方样条的两种用途。一是识别非线性关系，二是构建预测模型。如Kendzerka等[5]使用限制性立方样条来检验连续型预测变量与心血管事件相关呼吸暂停患者住院风险之间的线性关系，当发现非线性的关联时，通过RCS构建预测模型实行风险预测。正如前文所述，连续型变量进行分类可能会丢失重要的信息，因此限制性立方样条为分析连续型预测变量对预测结果的影响提供了一个有用的工具，它在预测变量和结局之间的关系形式上允许很大的灵活性。

限制性立方样条Cox模型，1989年由Durrleman等[16]学者提出，用于生存分析中拟合灵活的非线性关系。在Cox比例风险回归模型中，考虑变量的非线性效应后的模型为：

$$\lambda(t,X_i)=\lambda_0(t)\exp(X_j\beta+RCS(x,k)) \quad (2)$$

$$RCS(x,k)=\sum_{i=1}^{k-1}\beta_i S_i(x) \quad (3)$$

其中：$S_2(x)=x$，$S_i(x)=(x-t_{i-1})_+^3-(x-t_{k-1})_+^3\frac{t_k-t_i}{t_k-t_{k-1}}+(x-t_k)_+^3\frac{(t_{k-1}-t_i)}{t_k-t_{k-1}}, i=1,2,\cdots,k-2$

1.3 深度生存神经网络Cox模型

DeepSurv是一种深度前馈神经网络[17]，是类似于Faraggi-Simon网络的多层感知器。不同的是加入了多个隐藏层，以及其他新的技术，例如权重衰减正则化、整流线性单位（ReLU）、批处理归一化（Bacth normalization）、dropout、随机梯度下降使用Nesterov动量、梯度修剪和学习率调度等。它通过网络θ的权重参数来预测患者的协变量对其风险率的影响。图1是DeepSurv的网络结构，X为输入到网络的患者的基线数据。DeepSurv模型不涉及筛选特征变量，网络的隐藏层使用了一个全连接的节点层（Fully-Connected Layer）和一个dropout层，交替堆叠。对最后一个dropout层的输出进行线性组合，最终得到风险率 $\hat{h}_\theta(x_i)$，其估计了Cox模型中的对数风险函数 $h(x)$，通过设置目标函数的平均偏负对数似然来训练网络，其重设计的损失函数如下：

$$l(\theta):=-\frac{1}{N_{E=1}}\sum_{i:E_i=1}(\hat{h}_\theta(x_i)-\log\sum_{j\in R(T_i)}e^{\hat{h}_\theta(x_j)})+\lambda\|\theta\|_2^2 \quad (4)$$

其中，$N_{E=1}$ 是观察到结局的病人例数，E 为生存结局，T 为生存时间，x 为基线特征变量，λ 是L2正则化参数。

在使用深度神经网络时，无论网络有多少层，每一层节点的输入都是上层网络输出的线性组合，需要引入非线性激活函数来拟合非线性关系来提高模型的表达能力，目前为止激活函数种类已超过20种，本文所用的激活函数有Sigmoid、Tanh、ReLU和LeakyReLU，它们的数学形式分别为 $f(x)=\frac{1}{1+e^{-x}}$，$f(x)=\frac{e^x-e^{-x}}{e^x+e^{-x}}$，$f(x)=\begin{cases}x,x>0\\0,\text{其他}\end{cases}$，$f(x)=\begin{cases}x,x>0\\ax,\text{其他}\end{cases}$，本文中 $a=0.01$。激活函数是向神经网络中引入非线性因素，通过激活函

数神经网络就可以拟合各种曲线。激活函数主要分为饱和激活函数(Saturated Neurons)和非饱和函数(One-sided Saturations)。Sigmoid和Tanh是饱和激活函数,而ReLU以及其变种LeakyReLU为非饱和激活函数。非饱和激活函数主要有如下优势:(1)可以解决梯度消失问题;(2)可以加速收敛。Sigmoid极容易导致梯度消失问题。假设神经元输入Sigmoid的值特别大或特别小,对应的梯度约等于0,即使从上一步传导来的梯度较大,该神经元权重(w)和偏置(bias)的梯度也会趋近于0,计算费时并且导致参数无法得到有效更新。ReLU激活函数的提出就是为了解决梯度消失问题;但ReLU会存在神经元"死亡"问题。而LeakyReLU的提出解决了ReLU的神经元"死亡"问题。

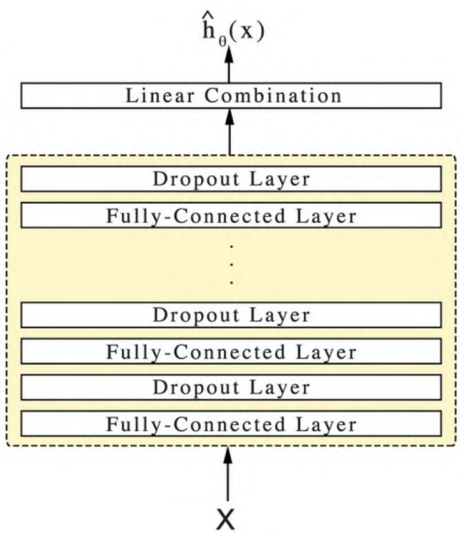

图1 DeepSurv网络结构图
Fig.1 Diagram of the structure of DeepSurv.

2 模拟与实例

2.1 模拟

本研究中的Cox回归使用R软件"survival"包实现。限制性立方样条Cox回归的拟合使用"rms"包完成。随机生存森林(RSF)实现使用"randomForestSRC"包完成。以上均基于R软件的4.0.3版本。DeepSurv(Cox_DNN)深度神经网络的拟合应用Python软件3.7.0版本完成,具体使用的是Pysurvival库中的NonLinearCoxPHModel函数。

2.1.1 生存数据模拟设置 在模拟研究中,生成生存时间常使用参数模型,常见的有参数PH模型、AFT模型。参数PH模型,其参数假定的分布包括指数分布、Weibull分布和Gompertz分布等[18, 19]。产生生存数据时,指数分布是最常见的分布假设,但实际数据往往并不满足指数分布,如医药领域常用Gompertz分布来描述数据[20]。因此本研究用Gompertz比例风险模型来产生满足比例风险假设的生存数据。同时本研究也将利用Weibull加速失效时间模型(AFT)产生不满足比例风险假设的生存数据[21, 22]。使用Gaussian非线性关联,其公式如下:

$$f(x)=a\times\exp\left(-\frac{(x-b)^2}{2\gamma^2}\right) \tag{5}$$

其中 a 表示Gaussian曲线尖峰高度,本文模拟时取值为1; b 表示Gaussian曲线中心位置,本文模拟时取值为0; γ 表示Gaussian曲线宽度,本文模拟时取值为0.5。

模拟一:

基于Gompertz比例风险模型生成生存时间,假设生存时间与特征变量之间的关系如下:

$$T_i=\frac{1}{\alpha}\log\left(1-\frac{\alpha\times\log(U)}{\lambda\exp\left(\beta^T X_i\right)}\right) \tag{6}$$

其中 α,λ 是分布参数。U为服从[0,1]的均匀分布的随机数。时间 T 的生存函数可表示为:

$$S_i(t)=\exp\left(\frac{\lambda}{\alpha}(1-e^{\alpha t_i})\right) \tag{7}$$

本研究使用Gaussian非线性Gompertz比例风险模型来产生满足比例风险假设的生存数据,公式如下:

$$T_i=\frac{1}{\alpha}\log\left(1-\frac{\alpha\times\log(U)}{\lambda\exp\left(\beta^T X_i+\beta^{\#}\exp\left(-\frac{\sum X_j^2}{2\gamma^2}\right)\right)}\right),\quad \gamma=0.5 \tag{8}$$

综上,通过以下过程产生Gaussian非线性Gompertz PH生存数据:(1)先确定样本量n,并根据生存分布的假设,设定好生存分布参数取值,Gompertz比例风险模型的分布参数 α,λ ,模拟时分别取0.5,1;(2)通过软件产生服从[0,1]均匀分布的随机数 U 和服从[-1,1]均匀分布的协变量 X ;(3)设定好回归系数 β 的取值,线性变量的系数 β 取值都为1,非线性变量的系数 $\beta^{\#}$ 在模拟时分别取值为1,3,6;根据公式5去模拟产生生存时间 T_i ;(4)根据事先指定的删失时间变量L的分布,以及设置的删失比例F%,本文设置的删失比例分别有10%,20%,40%,60%,80%;通过迭代确定不同删失分布的参数,生成删失时间的n个时间点 L_i ;(5)通过比较 T_i 、L_i 的大小,得到每个个体的截尾指示变量 $\delta_i=\mathrm{I}[T_i\leqslant L_i]$,其中 $\mathrm{I}[\cdot]$ 为示性函数。若 $T_i\leqslant L_i$,则 $\delta_i=1$,表示终点事件发生,否则 $\delta_i=0$,表示删失。生存时间 $t_i=\min(T_i,L_i)$ 。最终得到含随机删失的生存数据

(t_i, δ_i)。

模拟二：

基于加速失效时间模型(AFT)生成生存时间，假设对数生存时间与特征变量之间的关系如下：

$$Y_i = \log(T_i) = \beta^T X_i + \sigma\varepsilon_i \quad (9)$$

其中 σ 为尺度参数，ε_i 为随机误差。对于对数线性生存模型来说，时间 T 的生存函数可表示为 ε_i 的生存函数：

$$S_i(t) = S_{\varepsilon_i}\left(\frac{\log t - \mu - \beta^T X_i}{\sigma}\right) \quad (10)$$

若指定随机误差项 ε_i 的分布，则称为参数加速失效时间模型。

设置误差项服从极值分布，即 $f(\varepsilon) = \exp(\varepsilon - \exp(\varepsilon))$，尺度参数 $\sigma \neq 1$ 时，对应为Weibull回归模型。本研究使用Weibull加速失效模型来产生不满足比例风险的Gaussian非线性生存数据，公式如下：

$$Y_i = \log(T_i) = \beta^T X_i + \beta^{\#} \exp\left(-\frac{\sum X_j^2}{2\gamma^2}\right) + \sigma\varepsilon_i,$$

$$\gamma = 0.5 \quad (11)$$

与前述一样，通过以下过程产生Gaussian非线性Weibull AFT生存数据：(1)先确定样本量 n，以及Weibull AFT模型的分布参数 σ，取值为0.1；(2)通过软件产生服从[0,1]均匀分布的随机数 ε 和服从[-1,1]均匀分布的协变量 X；(3)设定好回归系数 β 的取值，$\beta^{\#}$ 在模拟时分别取值为1,3,6,其余 β 值都为1；根据公式10去模拟产生生存时间 T_i；第4和5步与前述Gaussian非线性Gompertz PH生存数据产生步骤一致。

综上，模拟数据主要分为两类，一类是满足PH假定的数据集(模拟一)，另一类是不满足PH假定的数据集(模拟二)。模拟总共设置10个变量，其中与生存结局有关的非线性变量1个(x_1)，线性变量1个(x_2)，与生存结局无关的线性变量8个，样本量分别设置为200,500,1000。

2.1.2 *模型参数设置* 模拟时，限制性立方样条节点数统一采用3节点。

对于神经网络参数的设置，由于基线特征变量的数量不多，因此在寻找神经网络最优参数时，各个参数范围分别为隐藏层神经元个数：3、5、7、10、15个；隐藏层层数：1到5层；激活函数类型：Sigmoid、Tanh、ReLU和LeakyReLU；学习率：1e-5,1e-4,1e-3,1e-2,1e-1。使用dropout和正则化技巧避免模型的过拟合，设定dropout比率为0.5，L2正则化参数为1e-4；默认使用Adaptive Momentum(adam)优化器。

在随机生存森林中，有两个重要的参数，分别是节点预选的变量个数(通常默认最佳的变量个数为全部变量个数的平方根)和随机生存森林中树的数量。通常来说，随机森林中树的数目决定了整个随机生存森林运行效果和时间。在拟合随机生存森林前，必须先确定生存树的个数，本文中当生存树的数量大于500时，错误率已趋于稳定，所以在本文的所有模拟研究中，随机生存森林模型的生存树棵数定为500，节点预选变量数为全部变量的平方根，分裂准则为log-rank。

在预测能力比较的模拟研究中，蒙特卡洛模拟的次数设置为1000次，产生数据总样本量设置为200、500、1000。删失率设置为10%、20%、40%、60%。为了使得四种方法的预测准确度具有可比性并避免过拟合现象，对每次模拟产生的数据集，我们随机选取其中的70%作为训练集拟合模型，剩余的30%作为测试集，得到模型的预测能力评价指标一致性指数(C-index)和积分布莱尔评分(Integrated Brier Score, IBS)，最后用所有模拟结果的C-index平均值和IBS平均值分别作为评估指标。

2.1.3 *模型评估* 在模型预测能力评估中，使用C-index和IBS这两个指标来评价：其中C-index为主要评价指标用于评价模型的预测区分度，IBS为次要评价指标用于评价模型的预测校准度。对于一个疾病预测模型，应先考虑区分度，如果模型区分度较差，不能区分不同风险人群，那么此模型就失去临床应用价值，再继续评价校准度也无意义了。

一致性指数(C-index)最早是由范德堡大学生物统计教授Frank E Harrell Jr 1996年提出，主要用于计算生存分析中的Cox模型预测值与真实之间的区分度，常用在评价患者预后模型的预测精度。C-index的取值范围为[0.5,1]，值越接近于1，表明模型预测准确度越高。C-index = 0.5，表明模型预测为随机预测，预测能力低。可以用以下公式计算[6,23]：

$$C - \text{ind ex} = \frac{\sum_{i,j} 1_{T_j<T_i} \cdot 1_{\eta_j<\eta_i} \cdot \delta_j}{\sum_{i,j} 1_{T_j<T_i} \cdot \delta_j} \quad (12)$$

其中，η_i 为第i个个体的风险得分，如果 $T_j < T_i$，则 $1_{T_j<T_i} = 1$，否则 $1_{T_j<T_i} = 0$；如果 $\eta_j < \eta_i$，则 $1_{\eta_j<\eta_i} = 1$，否则 $1_{\eta_j<\eta_i} = 0$。

布莱尔评分用于评估在给定时间t的预测生存函数的准确性，它表示观察到的生存状态和预测的生存概率之间的平均平方距离[24]，假设样本量为 N，$\forall i \in [1,N]$，(x_i,t_i)分别为第 i 个个体的预测变量和生存时间，$\hat{S}(t|X_i)$ 为预测的生存函数，此时布莱尔评分计算公式为：

$$BS = \frac{1}{N}\sum_{i=1}^{N}(I(t_i > t) - \hat{S}(t|X_i))^2 \quad (13)$$

当存在删失时，需要使用逆概率删失加权法对平方

距离进行加权来调整得分，调整后计算公式为：

$$BS(t)=\frac{1}{N}\sum_{i=1}^{N}\left\{\left[0-\hat{S}(t|X_i)\right]^2\frac{I(t_i\leqslant t,\delta_i=1)}{\hat{G}(t_i|X_i)}+\left[1-\hat{S}(t|X_i)\right]^2\frac{I(t_i>t)}{\hat{G}(t|X_i)}\right\} \quad (14)$$

其中 $\hat{G}(t|x)$ 是预测的条件生存函数Kaplan-Meier的估计值。积分后可得到IBS指标：

$$IBS(t_{\max})=\frac{1}{t_{\max}}\int_0^{t_{\max}}BS(t)\mathrm{d}t \quad (15)$$

IBS提供了在所有可用时间的模型性能的总体计算。

2.2 模拟结果

2.2.1 满足PH假定的情况　图2展示了变量满足比例风险假定时，不同删失率（设置为0.1，0.2，0.4和0.6）和样本量（设置为200，500，1000）情况下4种模型的预测区分度指标C-index和预测校准度指标IBS的变化情况，C-index值越高，模型的预测区分度就越高，模型的预测能力越好。图2的右纵坐标为生存数据产生时非线性变量的系数值（$\beta^{\#}$）及其余线性变量的系数取值均为1。其中Cox为Cox比例风险回归模型，Cox_DNN为深度生存神经网络Cox模型，Cox_RCS为限制性立方样条Cox模型，RSF为随机生存森林模型。在所有模拟条件下（样本量200-1000，删失率10%~60%），删失率和样本量对Cox_RCS模型预测区分度的稳定性影响不大，此时的Cox_RCS模型预测区分度始终最好的；Cox_DNN受删失率的影响较大，当删失率>40%时，Cox_DNN的预测区分度大幅降低；但当样本量较大（本文≥1000）删失率不高（<40%）时，Cox_DNN与Cox_RCS的C-index接近；Cox_DNN在样本量足够大、删失率越低时预测能力会越高。RSF的预测区分度略逊于其他模型。

IBS值越低，模型的预测校准度就越好。IBS的结果与C-index一致；删失率和样本量对Cox_RCS模型预测校准度的稳定性影响不大，此时的Cox_RCS模型预测校准度始终是最好的；RSF的预测校准度表现不佳，Cox_DNN受删失率的影响大，删失率大于40%时Cox_DNN的预测校准度大幅降低。

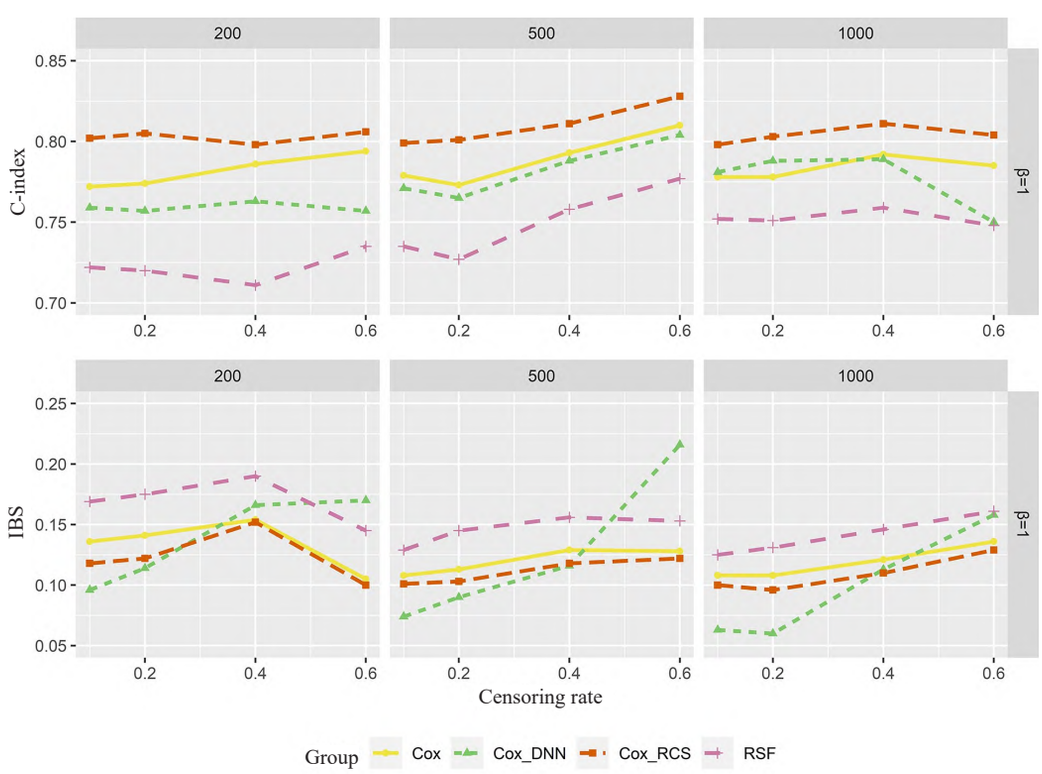

图2　4种方法在非线性变量系数 $\beta=1$ 且满足PH模拟数据集中的C-index(上)以及IBS(下)

Fig.2　Concordance index (top) and integrated Brier score (bottom) of Cox, Cox_DNN, Cox_RCS and RSF for PH datasets when the nonlinear variable coefficient $\beta=1$.

总之，在满足PH假定的数据中，在预测区分度和测校准度上，Cox_RCS模型预测能力表现最好；在样本量较高（≥1000）、删失率较低（<40%）时，Cox_DNN模型预测能力与Cox_RCS相当，RSF预测表现不佳。

2.2.2 不满足PH假定的情况 图3展示了变量不满足比例风险假定时，不同删失率（设置为0.1，0.2，0.4和0.6）和样本量（设置为200，500，1000）情况下4种模型的预测区分度指标C-index和预测校准度指标IBS的变化情况，C-index值越高，IBS值越低，则模型的预测能力就越好。图3的右纵坐标为生存数据产生时非线性变量的系数值（$\beta^{\#}$）及其余线性变量的系数取值均为1。

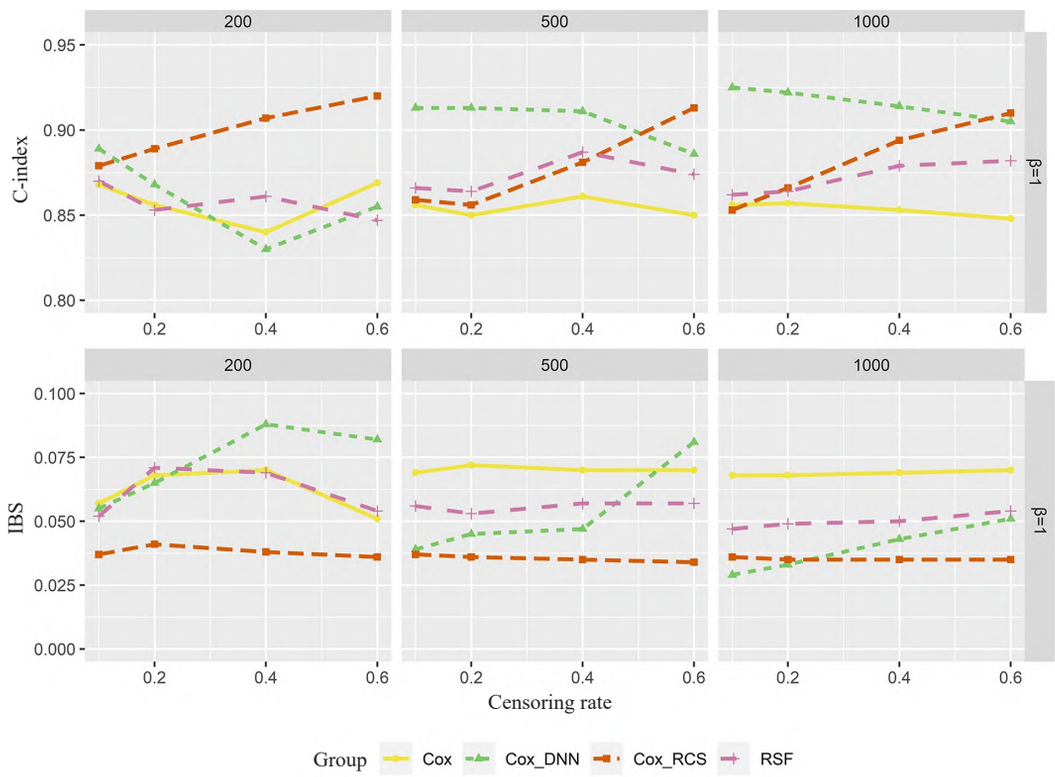

图3 4种方法在非线性变量系数 $\beta=1$ 且不满足PH模拟数据集中的C-index（上）和IBS（下）

Fig.3 Concordance index (top) and integrated Brier score (bottom) of Cox, Cox_DNN, Cox_RCS and RSF for non-PH datasets when the nonlinear variable coefficient $\beta=1$.

在预测区分度上，Cox模型的预测区分度低于其他3种方法，表明Cox模型在不满足比例风险的数据中拟合效果不佳；样本量为200，删失率在0.1~0.6时，Cox_RCS的C-index值在0.9左右波动，其余3个模型的值都在0.85左右波动，Cox_RCS的预测区分度高于其他3个模型；但是随着样本量增大，Cox_DNN的C-index值也在随之增大；当样本量大于等于500时，Cox_DNN的预测区分度最高；当样本量达到1000时，无论删失率变化如何，Cox_DNN的C-index值能保持在0.9以上；当样本量低于1000时，随着删失率增大，Cox_DNN的C-index值在逐渐降低。这也说明了在不满足比例风险假定的数据集中，模型拟合的不稳定性。

在预测校准度上，Cox_RCS模型的预测校准度高于其他3种方法，Cox模型的预测校准度最低；随着样本量的增加，Cox_DNN的预测校准度也在逐渐升高，删失率的增大会使Cox_DNN的预测校准度降低，但删失率对另外3个模型的预测校准度影响不大；RSF的预测校准度高于Cox模型，低于其他两种模型。

总之，在不满足比例风险的数据中，Cox_RCS在模型预测校准度上占优但在预测区分度上表现不稳定。Cox_DNN的预测校准度受删失率和样本量的影响大；样本量越大（本文中≥500）、删失率越低（本文中<40%），则Cox_DNN的预测校准度越高。Cox和RSF的预测能力都表现不佳。

2.3 实例

通过实例数据（WHAS500）比较四种方法（Cox，Cox_RCS, Cox_DNN和RSF）的优劣。为防止过拟合现象发生，将随机选取数据集的70%作为训练集训练模型，剩余的30%作为测试集，用C-index和IBS评价模型在测试集中的预测区分度和校准度，对该过程重复1000次，采用箱图展示预测指标C-index和IBS的结果（表1）。

伍斯特心脏病研究数据集（WHAS）包含了500名经历过急性心肌梗死患者，是右删失生存数据[25]，

表1 实例数据集中神经网络参数调节范围

Tab.1 Adjustment range of neural network parameters in the example dataset

Adjustment parameters	Range
Number of hidden units	[5,10,15,20,50]
Number of hidden layers	[1,2,3,4]
Activation function	[Sigmoid,Tanh,ReLU,LeakyReLU]
Learning rate	[0.00001,0.0001,0.001,0.01,0.1]

WHAS500数据集下载网址：https://github.com/rfcooper/whas/blob/master/whas500.csv。500名患者在观察时间内死亡215人，其余患者删失，删失率为57%。生存时间分布如图4所示，中位生存时间为1627 d(生存时间范围：1~2358 d)。表2列出了该数据集生存时间lenfol和删失变量fstat和14个基线变量的基本信息。

变量选择通过RSF最小深度法变量重要性筛选得出，删除分布严重不均衡的变量sho、av3和重要性排名最末的变量afb和cvd，最终选择纳入模型的变量有10个age, hr, sysbp, diasbp, bmi, los, gender, chf, miord, mitype。用Schoenfeld残差检查比例风险，结果显示sysbp，mitype这2个变量不满足比例风险假定。

非线性检测显示bmi, los这两个变量为非线性预测变量，经过对比确定bmi的节点个数为3，los的节点个数为4。RSF树的个数在1500之后模型趋于稳定，因此RSF树的个数设为1500，节点为预测变量个数的对数取整数3。表1展示神经网络参数调节范围。通过1000次对比确定Cox_DNN(DeepSurv)神经网络参数分别设置隐藏层神经元个数为20，隐藏层层数为3，激活函数为tanh，学习率为0.001，其余参数选择默认结果。

表2 WHAS数据集的基线指标信息分布

Tab.2 Characteristics of the covariates in Dataset WHAS [*Mean*±*SD* (min~max) or *n* (%)]

Variable	Description	Codes	Values
Id	Identification Number	-	1-500
Age	Age at Hospital Admission	Years	69.85±14.49
Hr	Initial Heart Rate	Beats per minute	87.02±23.59
Sysbp	Initial Systolic Blood Pressure	mmHg	144.70±32.29
Diasbp	Initial Diastolic Blood Pressure	mmHg	78.27±21.55
Bmi	Body Mass Index	kg/m^2	26.61±5.41
Los	Length of Hospital Stay	Days	6.12±4.71
Cvd	History of Cardiovascular Disease	0=No	125(25.0%)
		1=Yes	375(75.0%)
Gender	Gender	0=Male	300(60.0%)
		1=Female	200(40.0%)
Afb	Atrial Fibrillation	0=No	422(84.4%)
		1=Yes	78(15.6%)
Sho	Cardiogenic Shock	0=No	478(95.6%)
		1=Yes	22(4.4%)
Chf	Congestive Heart Complications	0=No	345(69.0%)
		1=Yes	155(31.0%)
Av3	Complete Heart Block	0=No	489(97.8%)
		1=Yes	11(2.2%)
Miord	MI Order	0=First	329(65.8%)
		1=Recurrent	171(34.2%)
Mitype	MI Type	0=non Q-wave,	347(69.4%)
		1=Q-wave	153(30.6%)
Fstat	Vital Status at Last Follow-up	0=Censor	285(57.0%)
		1=Dead	215(43.0%)
Lenfol	Total Length of Follow-up	Time(days)	882.44±705.67

从预测区分度上看，C-index值越高模型预测区分度越好，Cox_RCS的预测区分度整体高于其他3个模型，其次是Cox_DNN，但Cox_DNN有较多数值较小的离群值，表现出模型预测的不稳定，由前面模拟可知在

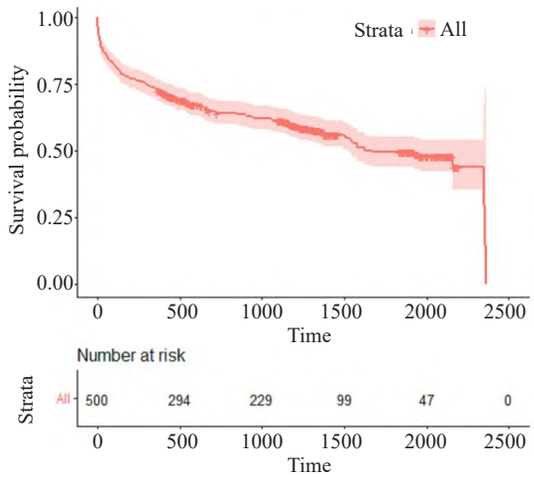

图4 数据WHAS500的生存曲线图

Fig.4 Survival curves of WHAS500.

删失率较高(>40%)时,Cox_DNN模型预测效果不佳,而WHAS500的删失率为57%,结果与模拟相符合;Cox和RSF的预测区分度相差不大(图5)。从预测校准度看,IBS值越低模型预测校准度越好,从图中可看出Cox_RCS的预测校准度整体高于其他3个模型,而Cox_DNN的预测校准度最低。综上所述,在含有非线性预测变量的数据中,使用Cox_RCS构建预测模型,其整体的预测准确度都高于其他3类模型。

3 讨论

尽管近年来陆续有学者提出深度学习神经网络方法以解决生存分析模型中非线性拟合问题,并且证明了在高维样本数据中深度神经网络方法优于传统的线性

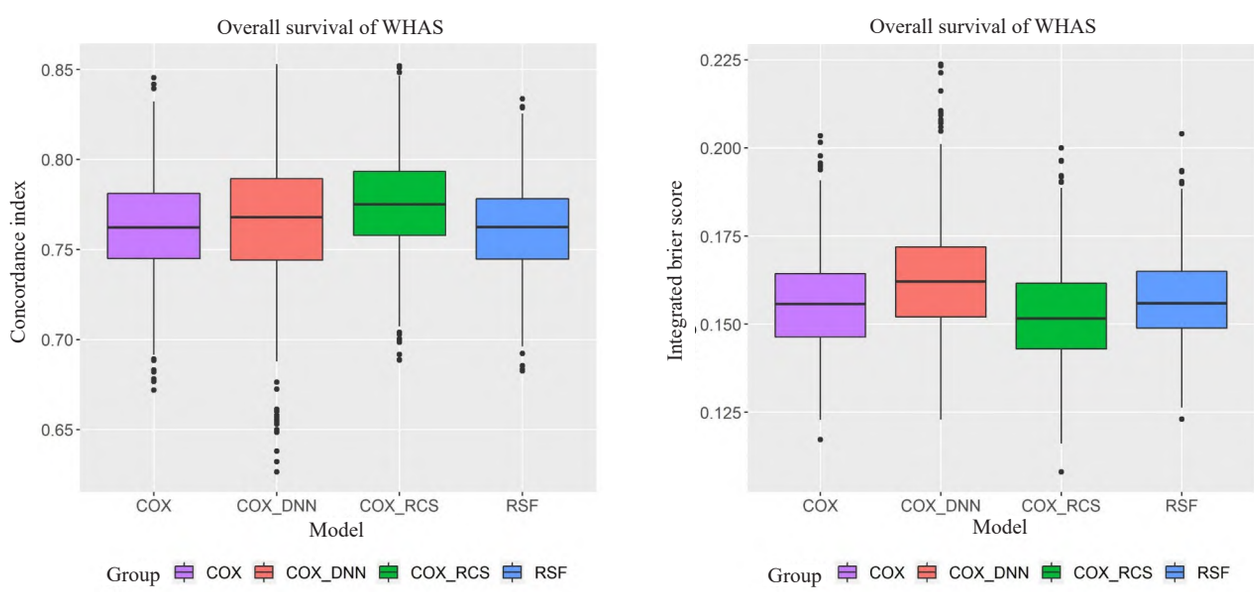

图5 WHAS数据集C-index(左)和IBS(右)的箱图

Fig.5 Concordance index (left) and integrated Brier scores (right) of WHAS dataset.

Cox回归,但对于扩展的Cox模型如限制性立方样条Cox模型与其他方法之间的比较研究还较少。并且目前还没有研究综合考虑样本量结合删失率时,各方法的优劣情况。

从模拟和实例研究来看,4个模型的预测区分度指标(C-index)和预测校准度指标(IBS),表现最好的是Cox_RCS。综合模拟与实例的结果,Cox_RCS在满足特定条件时的拟合效果优于其它3个模型。因此,对于低维且存在非线性变量的数据,协变量若满足比例风险假定,则推荐使用Cox_RCS。若不满足比例风险假定,高删失率(本文中≥40%)情况下可使用Cox_RCS,若为低删失率(本文中<40%)且大样本量(本文中≥500),推荐使用Cox_DNN,否则使用Cox_RCS。

当非线性变量存在时,传统的扩展Cox模型如Cox_RCS便能较好的进行预测,关于各个模型的运行时间和难易程度,Cox_RCS的实现难点是需要筛选出具体的非线性变量,并确认非线性变量是哪几个,而机器学习和神经网络方法都无需识别出具体的非线性变量,而是直接把变量纳入模型进行拟合。Cox_RCS的优点是可以直观了解协变量的表现形式。随机生存森林是机器学习树模型,由树节点不断分裂构成,机器学习方法构建模型前需要调参,如随机生存森林需要确定树分裂的个数以及节点数以达到树模型稳定。深度学习神经网络模型需要对更多的模型参数进行调优,通常

可以将调优超参数分为3类:网络参数、优化参数、正则化参数。网络参数包括神经网络的隐藏层层数,每一层的神经元个数,激活函数等;优化参数一般指学习率(learning rate)、批样本数量(batch size)、不同优化器的参数等;正则化参数的设置可以防止模型过拟合,一般包括权重衰减系数,丢弃法比率(dropout)等。因此深度学习神经网络更为复杂,所花费的时间也是最多的。

从结果解释性上看,Cox和Cox_RCS的结果解读起来较为直观易懂,因为其构建的是回归模型,模型拟合的形式较机器学习方法简单,我们能根据每个变量的回归系数直接解释其对生存结局的影响。而Cox_DNN和RSF等机器学习方法使用前需要不断探究最优解的参数,通过建立模型来预测最终结果。但我们无法从机器学习的方法中直接解读每个变量与生存结局的关系,其结果解释性较差。

传统扩展Cox模型的优点是模型较为简单,变量拟合的形式清楚明白,能输出具体每个变量的回归系数,并从中可以得到每个变量对生存时间的影响大小;缺点是无法在高维数据中拟合。随机生存森林的优点是可以高效识别并筛选变量,可以输出变量的重要性排名,适用于高维数据;其缺点是无法直观的得到变量与结局指标的具体表现形式,并且数据的拟合模型在预测上并没有优于其他模型。深度生存神经网络方法优点是可以很好处理大样本数据,并且可以识别图形进行建模;其缺点是数据删失程度对模型影响大,模型较为复杂,并且调参数需要花费大量时间。因此实际应用时需要结合数据,选择适合的方法。

参考文献:

[1] Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea[J]. Stat Med, 2006, 25(1): 127-41.

[2] Bhaskaran K, dos-Santos-Silva I, Leon DA, et al. Association of BMI with overall and cause-specific mortality: a population-based cohort study of 3·6 million adults in the UK[J]. Lancet Diabetes Endocrinol, 2018, 6(12): 944-53.

[3] Lee DH, Keum N, Hu FB, et al. Predicted lean body mass, fat mass, and all cause and cause specific mortality in men: prospective US cohort study[J]. BMJ, 2018, 362: k2575.

[4] 魏 源, 周锦辉, 张振伟, 等. 限制性立方样条在Cox比例风险回归模型中的应用[J]. 中华预防医学杂志, 2020, 54(10): 1169-73.

[5] Kendzerska T, Gershon AS, Hawker G, et al. Obstructive sleep apnea and risk of cardiovascular events and all-cause mortality: a decade-long historical cohort study[J]. PLoS Med, 2014, 11(2): e1001599.

[6] 刘 瑾, 杨燕玲, 严 可, 等. 列线图可预测首发缺血性脑卒中患者的复发[J]. 南方医科大学学报, 2022, 42(1): 130-6.

[7] 王会刚, 申聪香, 陈 芳, 等. 晚期鼻腔鼻窦腺样囊性癌21例临床特征分析[J]. 南方医科大学学报, 2017, 37(6): 847-52.

[8] 诸聪妍, 卢观婷, 祁婷婷, 等. 乙型肝炎相关慢加急性肝衰竭患者的长期预后及生存质量[J]. 南方医科大学学报, 2018, 38(6): 736-41.

[9] Hothorn T, Hornik K, Zeileis A. Unbiased recursive partitioning: a conditional inference framework[J]. J Comput Graph Stat, 2006, 15(3): 651-74.

[10] Wright MN, Dankowski T, Ziegler A. Unbiased split variable selection for random survival forests using maximally selected rank statistics[J]. Stat Med, 2017, 36(8): 1272-84.

[11] 刘颖欣, 康 佩, 许 军, 等. 条件推断森林在生存分析中的应用[J]. 南方医科大学学报, 2020, 40(4): 475-82.

[12] 黄福强, 康 佩, 刘颖欣, 等. 含治愈个体生存资料的亚组识别研究[J]. 中国卫生统计, 2020, 37(5): 672-7.

[13] 曾彭归航, 唐秀晓, 吴庭芩, 等. 潜在的胚胎干细胞自我更新与多能性的调控基因的鉴定: 基于随机森林算法[J]. 南方医科大学学报, 2021, 41(8): 1234-42.

[14] Adeoye J, Hui LL, Koohi-Moghadam M, et al. Comparison of time-to-event machine learning models in predicting oral cavity cancer prognosis[J]. Int J Med Inform, 2022, 157: 104635.

[15] Cox DR. Regression models and life-tables[J]. J Royal Stat Soc Ser B Methodol, 1972, 34(2): 187-202.

[16] Durrleman S, Simon R. Flexible regression models with cubic splines[J]. Stat Med, 1989, 8(5): 551-61.

[17] Katzman JL, Shaham U, Cloninger A, et al. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network[J]. BMC Med Res Methodol, 2018, 18(1): 24.

[18] Bender R, Augustin T, Blettner M. Generating survival times to simulate Cox proportional hazards models[J]. Stat Med, 2005, 24(11): 1713-23.

[19] 钱 俊. 生存分析中删失数据比例对Cox回归模型影响的研究[D]. 广州: 南方医科大学, 2009.

[20] Orbe J, Ferreira E, Núñez-Antón V. Comparing proportional hazards and accelerated failure time models for survival analysis[J]. Stat Med, 2002, 21(22): 3493-510.

[21] 康 佩, 许 军, 黄福强, 等. Adaptive Elastic Net结合加速失效时间模型在亚组识别中的应用[J]. 南方医科大学学报, 2019, 39(10): 1200-6.

[22] 韦红霞, 康 佩, 刘颖欣, 等. 基于Adaptive Elastic Net与加速失效时间模型的亚组识别方法的应用拓展[J]. 南方医科大学学报, 2021, 41(3): 391-8.

[23] Uno H, Cai T, Pencina MJ, et al. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data[J]. Stat Med, 2011, 30(10): 1105-17.

[24] Brier GW. Verification of forecasts expressed in terms of probability[J]. Mon Wea Rev, 1950, 78(1): 1-3.

[25] Hosmer DW, Lemeshow S, May S. Applied survival analysis: regression modeling of time to event data, Second edition[M]. Wiley, 2011.

(编辑:经 媛)