基于梯度提升树的生存分析优化方法研究及应用

硕士学位论文答辩

答辩人：刘沛（201721060103）

答辩日期：2020 年 5 月 17 日

电子科技大学计算机科学与工程学院

目录

研究背景与意义

研究背景与意义

生存分析（a.k.a. time-to-event analysis）：

- **研究内容**：个体在不同观测期发生某个特定事件的概率
- **应用领域**：医疗健康、金融等

生存分析模型：

- **模型输入**：个体协变量、观测时间和观测状态
- **模型输出**：感兴趣的事件发生在各个时间点上的概率分布

研究背景与意义

在临床疾病预后研究中，生存分析方法用来：

- 分析和研究患者的随访数据
- 建立精准的、健壮的生存预后模型
- 辅助医生进行诊断和治疗，或发现与疾病相关的重要影响因子

例如在乳腺癌预后研究中

- 权威的乳腺癌预后模型：
  - 21Gene：使用基因表达量估计乳腺癌患者 *5/10* 年复发评分
  - PREDICT：使用临床特征预测乳腺癌患者生存曲线
- 现实意义：
  - 科学诊疗以减轻患者负担
  - 帮助临床研究者更好地认识疾病

所以，对生存分析方法的研究具有重要的价值和现实意义。

研究历史与现状

研究历史与现状

下面重点从以下三个方面简要介绍生存分析方法的研究历史与现状：

- 生存分析基础：生存数据、评估指标
- 统计线性模型：*Cox*、*FHT*
- 机器学习模型：*RSF*、*GBM*

研究历史与现状

生存分析基础 ——生存数据

使用集合 $\{(x_i, T_i, \delta_i) \mid i = 1, \ldots, n\}$ 表示

- n 表示数据中观测个体的数目
- 向量 $x_i \in \mathbb{R}^m$ 表示第 i 个个体的协变量
- $T_i \in \mathbb{R}^+$ 表示该个体的末次随访时间
- $\delta_i \in \{0, 1\}$ 表示在 T_i 时刻是否观测到该个体发生时间

其他重要概念：

- 右删失：个体 i 属于集合 $\{i \mid T_i < T_e, \delta_i = 0\}$
- *Ties*：多个个体在同一时刻发生事件
- 生存函数：生存时间 T 超过 t 的概率，即 $S(t) = Pr(T > t)$
- 风险函数：在生存时间大于 t 的条件下在 t 时刻发生事件的概率

研究历史与现状

统计线性模型 —— Cox 比例风险模型

生存分析中最为经典的模型，由 Cox 于 1972 年提出。它可以预测每个个体的风险比例。

Cox 比例风险模型：

- 写作 $f(x) = \theta^T x$，θ 为模型参数
- $e^{f(x)}$ 被称为风险比例
- 基于风险比例假设：$e^{f(x)} = \frac{h(t|x)}{h_0(t)}$，$h(t)$ 表示风险函数
- 优化 *Breslow* 偏似然估计函数来估计模型参数

Cox 比例风险模型非常直观，且易解释。但是，它本质上仍然是一个线性模型。

基于 Cox 模型，同样是统计线性模型的还有 CoxNet、Time-Dependent Cox、CoxBoost。

研究历史与现状

统计线性模型 —— FHT 首次命中时间模型

主要研究事件首次发生时间 FHT（First Hitting Time），其内容可见 Lee 和 Whitmore 于 2006 年发表的文章。

FHT 模型不再基于比例风险假设，而是对 FHT 的分布进行建模。

- 假设个体的风险函数是某个固定形式的带参数的随机过程
- 风险函数形式：$P(t) \sim W(t \mid s_0, \mu, \sigma^2 = 1),\ t \geq 0$
- 模型参数和个体协变量通过广义的线性链接函数建立联系
- 最大化极大似然估计函数来估计模型参数

FHT 模型从另外一个角度解决生存分析问题。它假设个体风险函数为一个维纳过程，本质上仍为线性模型。

基于 FHT 模型，同样是统计线性模型的还有 FHTBoost（Stikbakke, 2019）。

机器学习模型 —— RSF 模型

基于随机森林的随机生存森林模型（Random Survival Forest），由 Ishwaran 等人于 2008 年提出。

RSF 模型在随机森林原有框架下，使用

- *logrank* 指标作为树节点分裂策略
- *Kaplan-Meier* 估计得到叶子节点的生存函数
- *Nelson-Aalen* 估计得到叶子节点的累积风险函数

RSF 模型完全基于生存函数和风险函数的无参数估计方法。这种无参数的估计方法往往依赖样本量大小，而且容易出现过拟合。

机器学习模型 —— GBM 模型

梯度提升树模型是由多颗决策树（或 CART）组成的前向加法模型。常用的梯度提升树模型：

- *gradient boosting machine*（*Friedman*）：损失函数一阶梯度值作为拟合目标
- *XGBoost*（陈天奇）：损失函数中加入二阶近似

基于梯度提升树的生存分析模型，其研究工作最早可见 Ridgeway 和 Harald 分别于 2005 年和 2008 年发表的文章。该模型特点：

- 基于 *Cox* 比例风险假设，对风险比例进行建模
- 模型假设空间更大，不仅限于线性空间
- 使用 *Breslow* 偏似然估计函数优化模型

GBM 模型提升了经典的 Cox 模型的性能。但是，一方面它仍然遵循风险比例假设；另一方面它使用的优化目标函数不够精确。

研究历史与现状

机器学习模型 ——神经网络模型

基于深度神经网络的生存分析模型，常见的有

- *DeepSurv (2018)*：基于 *Cox* 模型假设，预测风险比例
- *DeepHit (2018)*：使用分层的复杂神经网络结构预测发生事件时间的分布
- *DRSA (2019)*：使用复杂的循环神经网络结构预测发生事件时间的分布

虽然 DeepHit 和 DRSA 模型均不再遵循任何模型假设，使用复杂的神经网络建模，但其缺点在于

- 需要相对更长的训练时间
- 无法直接对模型特征做出恰当解释

论文的主要工作及贡献

主要贡献与创新

针对现有生存分析方法存在的问题，论文主要研究并提出了基于梯度提升树的生存分析优化方法，最后将其应用于乳腺癌复发预后建模。

下面重点从以下三个方面介绍论文的主要工作及贡献：

- *HitBoost* 生存分析方法
- *BecCox* 生存分析方法
- 乳腺癌复发预后模型

主要贡献与创新

论文工作 —— HitBoost 生存分析方法

下面重点从以下三个方面介绍论文的主要工作及贡献：

- *HitBoost* 生存分析方法
- *BecCox* 生存分析方法
- 乳腺癌复发预后模型

主要贡献与创新

论文工作 —— BecCox 生存分析方法

下面重点从以下三个方面介绍论文的主要工作及贡献：

- *HitBoost* 生存分析方法
- *BecCox* 生存分析方法
- 乳腺癌复发预后模型

主要贡献与创新

论文工作 ——乳腺癌复发预后模型

下面重点从以下三个方面介绍论文的主要工作及贡献：

- *HitBoost* 生存分析方法
- *BecCox* 生存分析方法
- 乳腺癌复发预后模型

论文的主要工作成果

论文的主要工作成果

HitBoost: Survival Analysis via A Multi-output Gradient Boosting Decision Tree Method，2019 年 4 月，第一作者（导师通讯作者），已发表。

期刊：*IEEE Access* | IF：4.098 | JCR 二区

主要工作：

- 提出一种使用多输出梯度提升树预测事件发生时间概率分布的生存分析方法
- 在模型性能和模型解释性等方面改进了现有的生存分析方法
- 推导自定义损失函数梯度并实现该方法：HitBoost

论文的主要工作成果

Optimizing Survival Analysis of XGBoost for Ties to Predict Prognostic Status of Breast Cance，2020 年 5 月，第一作者（导师通讯作者），已录用。

期刊：*IEEE Trans. on Biomedical Engineering* | IF：4.491 | JCR 二区

主要工作：

- 使用梯度提升树预测风险比例
- 从目标函数的角度改进了基于 Cox 模型的生存分析方法
- 将提出的方法用于乳腺癌复发风险预测建模

论文的主要工作成果

Predicting Invasive Disease-Free Survival for Early-stage Breast Cancer Patients Using Follow-up Clinical Data，2018 年 11 月，第二作者（导师一作），已发表。

期刊：*IEEE Trans. on Biomedical Engineering* | IF：4.491 | JCR 二区

主要工作：

- 使用分层特征选择方法从乳腺癌临床数据的高维特征中筛选重要特征
- 构建，优化和验证用于预测早期乳腺癌患者复发的机器学习预后模型：*MP4Ei*

论文的主要工作成果

其他成果：

- **, 刘沛,**,**,**. 一种基于乳腺癌临床高维数据的分层重要特征选择方法 [P]. 中国, 发明专利,CN201810552686.3,2018 年 5 月 31 日（申请专利）
- **, 刘沛,**,***,**. 一种基于 Efron 近似优化的生存风险建模方法 [P]. 中国, 发明专利,CN201910315815.1,2019 年 4 月 19 日（申请专利）
- **, 刘沛,**,**,**,***. 一种用于生存风险分析的多输出梯度提升树建模方法 [P]. 中国, 发明专利,CN201910315829.3,2019 年 4 月 19 日（申请专利）

谢谢各位老师