doi:10.3969/j.issn.1001-4616.2021.04.001

基于随机生存森林的企业财务危机研究

肖叶宇[1],张　闪[2]

(1.吉林大学数学学院,吉林 长春 130012)
(2.南京财经大学应用数学学院,江苏 南京 210023)

[摘要]　以沪深两市 A 股制造业上市公司为样本,将随机生存森林模型引入企业财务危机研究中去. 通过计算两种度量下变量重要性排名,发现营业收入增长率和息税前利润对财务危机的影响最大. 随后将随机生存森林与 Cox、后向逐步 Cox 和 Lasso-Cox 模型进行对比,随机生存森林的预测性能要优于 3 种 Cox 模型. 同时结合随机生存森林下的生存函数和累积风险函数,对公司被特别处理的时间进行分析,结果显示模型有很好的预警功效,可以为各利益相关方的决策提供依据.

[关键词]　随机生存森林,生存分析,财务危机预警
[中图分类号]O213　[文献标志码]A　[文章编号]1001-4616(2021)04-0001-06

Research on Financial Crisis of Enterprises Based on Random Survival Forest

Xiao Yeyu[1],Zhang Shan[2]

(1.College of Mathematics,Jilin University,Changchun 130012,China)
(2.School of Applied Mathematics,Nanjing University of Finance and Economics,Nanjing 210023,China)

Abstract:Taking Shanghai and Shenzhen A-share manufacturing listed companies as a sample,the Random Survival Forest model is introduced into the research on corporate financial crisis. It is found that the growth rate of operating income and the profit before interest and tax have the greatest impact on the financial crisis by calculating the ranking of the importance of variables under the two measures. Subsequently,the Random Survival Forest was compared with Cox model,Cox model with backward stepwise variable selection and Lasso-Cox models. The prediction performance of the Random Survival Forest is better than the three Cox models. At the same time,combined with the survival function and cumulative hazard function under the random survival forest,the company is analyzed for the time when the company is ST. The results show that the model has a good early warning function,which can provide a basis for the decision-making of various stakeholders.

Key words:Random Survival Forest,survival analysis,financial crisis early warning

近年来随着我国证券市场的高速发展,上市公司的财务危机已成为投资者所关心的重要问题. 一旦上市公司因连续亏损或财务状况异常而被特别处理(ST),则该公司会被认为陷入财务危机,其投资者也会受到负面影响. 故而建立有效的企业财务预警模型,让上市公司及早发现危机并采取措施,对各利益相关方都有积极的意义.

目前我国的财务危机研究在统计方法上主要可分为两大类,一是将其看做二分类问题即公司是否被 ST,所用模型包括判别分析模型、Logistic 模型、支持向量机[1]等. 二是将生物统计里的生存分析理论引入财务危机问题中,由于生存分析模型可以处理删失数据,并给出研究对象各时间点的生存概率,所以将生存分析应用到财务危机预警中有很强的现实意义.

前人在财务危机研究中采用的生存分析模型多以 Cox 模型为主[2-3],但 Cox 模型依赖于比例风险假设,而且在确定变量的非线性影响、识别交互作用上都有局限性. 随机生存森林是在随机森林模型的基础上提出的一种可以分析右删失数据的集成树方法,它继承了随机森林的优点,故而可以克服这些局限,并

收稿日期:2021-03-12.
基金项目:国家自然科学基金项目(11601224).
通讯作者:张闪,博士,副教授,研究方向:生物数学. E-mail:shanzhang86@163.com

且在高维数据下显著优于其它生存分析方法.

本文首先采用随机生存森林模型研究财务危机,所研究样本为沪深两市 A 股制造业上市公司,得到随机生存森林模型下各财务指标的重要性,然后比较随机生存森林和 3 种 Cox 模型的预测精度,最后结合模型结果给出分析与建议.

1 随机生存森林与企业财务危机

生存分析是医学领域研究生存时间的主要方法,在生存分析中生存时间的主要刻画方式有:生存函数和风险函数. 生存函数定义为:

$$S(t)=P(T>t),0<t<\infty, \tag{1}$$

表示为生存时间 T 超过 t 的概率,其中 T 为非负随机变量.

风险函数定义为:

$$\lambda(t)=\lim_{h\to 0^+}P(t\leqslant T<t+h\mid T\geqslant t)/h, \tag{2}$$

表示 t 时刻还存活的个体发生结局事件的瞬时速率,即为个体在 t 时刻经历结局事件的风险. 累积风险函数定义为:$\Lambda(t)=\int_0^t\lambda(s)\mathrm{d}s$. 3 个函数可以相互转换[4-5].

本文将公司的首次上市作为生存时间的起始事件,以公司被 ST 作为结局事件,二者间隔的时间为该公司的生存时间. 若某公司在观测结束时还未被 ST 便被视为删失个体,其生存时间为右删失数据. 若某公司被 ST 则为失效个体,其生存时间为完全数据.

1.1 随机生存森林模型

首先使用 Bootstrap 重抽样在原始数据中抽取 B 个 Bootstrap 样本,其中每个 Bootstrap 样本的数据量与原始数据相同,由于是有放回抽样,故原始数据中约有 37%的数据不会出现在一个 Bootstrap 样本里[6],这部分数据称为该 Bootstrap 样本的袋外数据.

在每个 Bootstrap 样本上都建立二元递归生存树模型,一棵树生长的过程中,在每个内部节点需要分裂时都随机选择 p 个候选变量(假设原有变量 P 个,$P>p$). 使用最大化子节点之间生存差异的候选变量对节点进行分裂,本文选择 Log-rank 统计量来刻画生存差异,Log-rank 统计量的绝对值越大生存差异越大.

生存树生长为完整大小,不采取任何剪枝,每个节点必须包含最少 d_0 个拥有不同生存时间的 ST 公司,当不满足该条件无法生成子节点时终止生长. 那么这些最后的节点就被称为终节点,记为 H. 设 h 为树的一个终节点,$(T_{1,h},\delta_{1,h}),\cdots,(T_{n(h),h},\delta_{n(h),h})$ 为终节点 h 中公司的生存时间和删失信息,对公司 i 而言,$\delta_{i,h}=1$ 表示在 $T_{i,h}$ 被 ST,$\delta_{i,h}=0$ 表示在 $T_{i,h}$ 发生右删失. 设 $t_{1,h}<t_{2,h}<\cdots<t_{N(h),h}$ 为终节点 h 上 ST 公司的 $N(h)$ 个不同的生存时间,$d_{l,h}$ 和 $Y_{l,h}$ 分别为 $t_{l,h}$ 时刻 ST 公司数量和在风险公司个数. 终节点 h 的累积风险函数由 Nelson-Aalen 估计获得:

$$\hat{H}_h(t)=\sum_{t_{l,h}\leqslant t}\frac{d_{l,h}}{Y_{l,h}}. \tag{3}$$

注意到终节点 h 中的所有公司共用一个累积风险函数估计,那么一棵树的累积风险函数就可以写为:

$$H(t\mid x_i)=\hat{H}_h(t),\quad \text{如果 } x_i\in h, \tag{4}$$

式中,x_i 代表公司 i 的协变量,$h\in H$. 式(4)是一棵生存树的累积风险函数,而随机生存森林的累积风险函数需要计算生存树的平均,包括袋外数据估计和 Bootstrap 估计. 每棵生存树都是使用独立的 Bootstrap 样本建立的,共有 B 个样本. 则公司 i 的累积风险函数袋外数据估计为:

$$H_e^{**}(t\mid x_i)=\frac{\sum_{b=1}^{B}I_{i,b}H_b^*(t\mid x_i)}{\sum_{b=1}^{B}I_{i,b}}, \tag{5}$$

式中,$H_b^*(t\mid x)$ 为第 b 个 Bootstrap 样本所建立的生存树的累积风险函数. 如果公司 i 是 b 的袋外数据则 $I_{i,b}=1$,否则 $I_{i,b}=0$. 将袋外数据代入袋内数据生成的生存树中,追踪袋外数据所在的终节点及其累积风险函数,取这些累积风险函数的均值,这就等于式(5). 公司 i 的袋外数据估计仅使用了 i 作为袋外数据的

Bootstrap 样本,相反地,Bootstrap 估计使用了所有的生存树[7]:

$$H_e^*(t|x_i)=\frac{1}{B}\sum_{b=1}^{B}H_b^*(t|x_i). \tag{6}$$

1.2 模型评价指标

(1)一致性指数(C-index)

随机生存森林使用 C-index 计算预测误差. C-index 估计了在随机选择的一对个体中,拥有更差预测结果的个体最先发生失效的概率,而且不同于其他衡量指标,它考虑到了删失的情况. C-index 的计算步骤如下:

Step 1. 将数据集上的所有样本互相配对,共有 C_N^2 对,N 为样本数.

Step 2. 删掉样本配对中企业生存时间较短的公司是非 ST 公司的配对,删掉样本配对中企业生存时间相同且都是非 ST 公司的配对. 记剩下的有效配对数为 P.

Step 3. 在有效配对里,在 $T_i\neq T_j$ 时,若有更短的生存时间的企业有更差的预测结果则记为 1,若预测结果相同则记为 0.5;当 $T_i=T_j$ 且两个企业都是 ST 公司时,若预测结果相同记为 1,否则记为 0.5;当 $T_i=T_j$ 且只有一个是非 ST 公司时,若 ST 公司有更差的预测结果则记为 1,否则记为 0.5. 将上述有效配对的结果求和,记为 C.

Step 4. C-index 定义为:$C\text{-}index=\frac{C}{P}$.

计算 C-index 需要预测结果,这里预测结果定义为袋外集成累积风险函数. 设 $t_1^0,\cdots,t_m^0$ 为预先选定的时间点,对两个个体 i 和 j,如果 $\sum_{l=1}^{m}H_e^{**}(t_l^0|x_i)>\sum_{l=1}^{m}H_e^{**}(t_l^0|x_j)$,则称 i 有更差的预测结果[7-8].

袋外预测误差 PE 则被定义为:$PE=1-C$.

(2)Brier score 与 IBS

在生存分析中还有几种指标被用来评估模型:带有逆概率删失加权的 Brier score 和 Integrated Brier score(IBS). 带有逆概率删失加权的 Brier score 定义为:

$$\hat{BS}(t,\hat{S})=\frac{1}{M}\sum_{i\in D_M}\hat{W}_i(t)\{Y_i(t)-\hat{S}(t|x_i)\}^2, \tag{7}$$

式中,D_M 为测试集,M 为测试集中的公司个数,$\hat{S}(t|x_i)$ 为基于训练集得到的公司 i 的生存函数,$Y_i(t)=I(T_i>t)$ 是公司 i 在 t 的真实状态,T_i 为生存时间. $\hat{W}_i(t)$ 是公司 i 的逆概率删失加权,定义为:

$$\hat{W}_i(t)=\frac{(1-Y_i(t))\delta_i}{\hat{G}(T_i-|x_i)}+\frac{Y_i(t)}{\hat{G}(t|x_i)}, \tag{8}$$

式中,$\delta_i=1$ 表示 i 为 ST 公司,$\delta_i=0$ 表示为发生右删失. $\hat{G}$ 为删失时间的生存函数估计,本文使用边际 Kaplan-Meier 估计.

由上式可知 Brier score 是时间的函数,它的总体度量是 Integrated Brier score:

$$IBS(\hat{BS},\tau)=\frac{1}{\tau}\int_0^{\tau}\hat{BS}(u,\hat{S})\,\mathrm{d}u, \tag{9}$$

式中,$\tau>0$,τ 为样本中最大生存时间,用以总结测试集中的预测误差[9-10].

1.3 变量选择方法

变量重要性(VIMP)筛选变量:为计算变量 x 的 VIMP,将袋外数据代入袋内数据生成的生存树中. 当遇到使用变量 x 的分割时,随机分配一个子节点. 每个袋外个体的累积风险函数将被重新计算并取平均值. x 的 VIMP 等于使用随机 x 分配获得的袋外预测误差减去原始的袋外预测误差. VIMP 值越大表示该变量具有预测能力,而零或负值表示这是非预测变量[7].

通过计算变量的最小深度来筛选变量:变量的最小深度为从根节点到该变量最近的极大子树根节点的距离. 对于随机生存森林,计算每棵树下变量的最小深度,然后取平均. 最小深度越小则该变量对预测的影响越大[11].

2　实证分析

2.1　数据描述

(1)样本选取

本文研究观测期为公司首次上市到 2020 年 12 月 31 日,生存时间以年为单位. 研究样本来自沪深两市 A 股制造业上市公司,总计 1 606 家,其中 ST 公司 306 家,在观测期内未被 ST 的公司 1 300 家. 此外由于我国证券交易所多因连续两年亏损而对上市公司实施特别处理,故而为使模型具有预测性,若公司在 T 年被 ST 或删失,则选取该公司 T-3 年的指标来进行建模[12].

(2)财务指标初选

在财务指标的选取上,本文结合前人研究基础,从公司的偿债能力、比率结构、经营能力、盈利能力、现金流能力、风险水平、发展能力、每股指标、相对价值指标、股权集中度这 10 个方面出发,共选取 50 个指标作为初选财务指标,具体见表 1. 同时由于部分公司的指标存在数据缺失,本文使用随机生存森林中的自适应树插值进行数据补全[7].

表 1　财务指标初选

Table 1　Primary selection of financial indicators

指标分类	指标名称
偿债能力	流动比率,速动比率,利息保障倍数,资产负债率,产权比率,长期资本负债率
比率结构	流动资产比率,营运资金对净资产比率,流动负债比率,股东权益对固定资产比率
经营能力	应收账款周转率,存货周转率,营运资金周转率,流动资产周转率,固定资产周转率,总资产周转率
盈利能力	资产报酬率,固定资产净利润率,营业利润率,息税前利润,净资产收益率,息税前利润与资产总额比,总资产净利润率,流动资产净利润率,长期资本收益率
现金流能力	净利润现金净含量,营业收入现金含量,营业利润现金净含量,营运指数
风险水平	财务杠杆,经营杠杆,综合杠杆
发展能力	资本积累率,固定资产增长率,总资产增长率,营业收入增长率,每股净资产增长率
每股指标	每股收益,每股营业收入,每股营业利润,每股未分配利润,每股净资产,每股留存收益,每股现金净流量
相对价值指标	市盈率,市净率
股权集中度	前 5 位大股东持股比率,Z 指数,S 指数,Herfindahl_5 指数

2.2　变量重要性

通过五折交叉验证比较各参数选择下的预测误差,随机生存森林模型参数设置为:节点最小 ST 公司数为 10,节点分裂时候选变量数为 10,森林中有 1 000 棵树.

随后在此基础上计算变量最小深度和 VIMP,其中 VIMP 为 100 次试验后取平均. 于是得到这两种重要性度量下最有预测性的 10 个变量,如表 2 所示.

表 2　两种度量下变量重要性排名

Table 2　Variable importance ranking under two measures

变量名称	最小深度	变量名称	VIMP
营业收入增长率	2.447	营业收入增长率	0.066
息税前利润	3.045	息税前利润	0.063
应收账款周转率	4.333	应收账款周转率	0.026
每股未分配利润	4.909	每股未分配利润	0.025
固定资产增长率	5.067	每股营业利润	0.022
Herfindahl_5 指数	5.505	营业利润现金净含量	0.012
营业利润现金净含量	5.539	资本积累率	0.012
每股营业利润	5.580	固定资产增长率	0.010
总资产增长率	5.593	Herfindahl_5 指数	0.010
资本积累率	5.624	每股留存收益	0.009

由表 2 可知,两种度量下重要性排名前 10 的变量基本相同,只是从第 4 位后顺序有所改变. 值得注意的是不管是最小深度还是 VIMP,营业收入增长率和息税前利润都明显比其他变量更重要. 最重要的财务指标是营业收入增长率,它反映的是企业营业收入增长的比率,是评价企业发展能力和成长状况的重要指标,它的值越大表示对企业盈利有正面影响,则发生财务危机的可能性越低. 排名第 2 的是息税前利润,究其原因,

它是扣除所得税和财务费用前的利润,是企业真实的经营利润,故而是企业盈利能力的直观体现.

2.3 模型预测比较

本文采用 C-index、Brier score 和 IBS 3 种评价指标衡量随机生存森林(RSF)在企业财务危机中的预测性能,其中 C-index 越大模型预测性能越好,而 Brier score 和 IBS 越小预测性能越好. 本文将随机生存森林与 Cox 模型、后向逐步 Cox 模型和 Lasso-Cox 模型进行对比.

将原始数据集按 7:3 的比例随机划分为训练集与测试集,分别在训练集和测试集上计算 C-index,并重复 100 次试验得到结果,如图 1 所示.

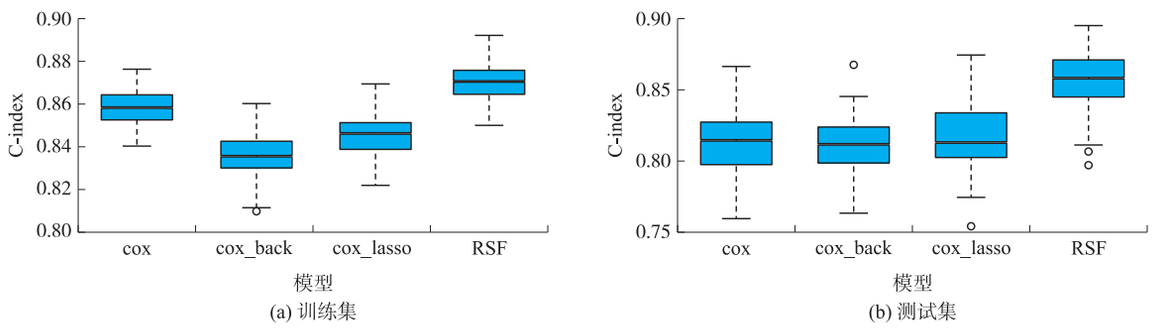

图 1 4 种模型在训练集(左)和测试集(右)上的 C-index

Fig. 1 C-index of the four models on the training set(left) and test set(right)

由图 1 可知,在训练集上随机生存森林的 C-index(中位数为 0.870)要优于 Cox(0.858)、后向逐步 Cox(0.836)和 Lasso-Cox(0.846). 在测试集上随机生存森林的预测性能优势更加明显,C-index 中位数为 0.859,而 Cox 模型为 0.815,后向逐步 Cox 为 0.812,Lasso-Cox 为 0.814.

为计算 4 个模型的 Brier score,对原始样本进行 100 次 Bootstrap 重抽样,对每个 Bootstrap 样本在袋内数据上训练模型,使用袋外数据计算 Brier score,最后取 100 次的平均.

由于原始样本中公司的最大生存时间约为 27 年,故而在计算 Brier score 时时间设为 0 到 27 年. 如图 2 所示,在绝大部分时间点上随机生存森林的 Brier score 都要小于 3 种 Cox 模型,因此随机生存森林的预测效果最好.

同时使用 Integrated Brier score 总结测试集中的预测误差,如表 3 所示,随机生存森林的 IBS 是最低的,即预测效果最好.

表 3 4 种模型的 Integrated Brier score

Table 3 Integrated Brier score of the four models

	Cox	Cox_backward	Cox_lasso	RSF
IBS	0.164	0.143	0.156	0.142

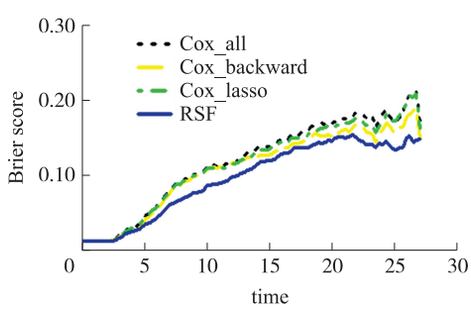

图 2 4 种模型的 Brier score

Fig. 2 Brier score of the four models

2.4 个体分析

为了说明随机生存森林在财务危机问题上的预测性,本文选取股票代码为 600866 的公司进行分析. 该公司于 1994 年上市,2016 年被 ST,生存时间约为 21.7 年. 将该个体放入已训练好的随机生存森林模型中,计算其生存函数和累积风险函数,如图 3 所示.

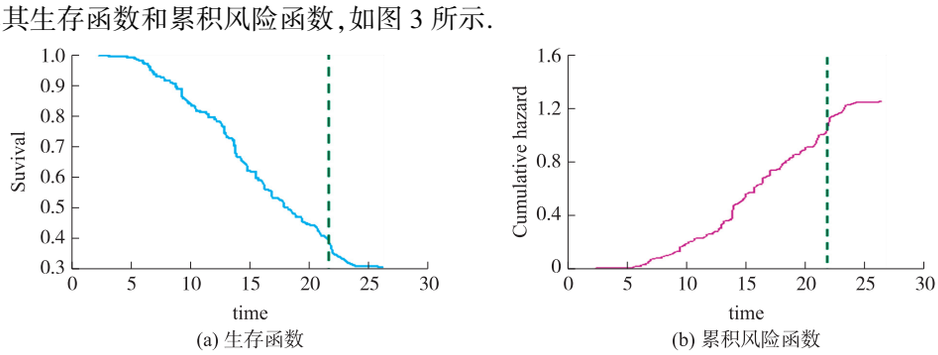

图 3 公司的生存函数(左)及累积风险函数(右)

Fig. 3 Survival function(left) and cumulative hazard function(right) of a company

图 3 绿色虚线横坐标为该公司被 ST 的时间. 从图 3 可以看出该公司在上市 20 年后生存概率已不到 40%,更重要的是在上市 21 年左右该公司的累积风险函数陡升,也就是说这一时间段风险函数很大即被 ST 的概率极大,这一结论和该公司在上市 21.7 年后被 ST 的事实相吻合. 所以基于随机生存森林,根据公司的财务指标可以计算其生存函数和累积风险函数,从而判断该公司被 ST 的风险.

3　结论

本文以沪深两市 A 股制造业上市公司为研究样本,从公司的偿债能力等方面初选 50 个财务指标,将随机生存森林模型引入企业财务危机研究中. 通过计算变量最小深度和 VIMP 得到两种度量下重要性排名前 10 的变量,发现前 4 位完全相同,分别是营业收入增长率、息税前利润、应收账款周转率和每股未分配利润,其中营业收入增长率和息税前利润对财务危机的影响最大,后 6 位变量基本相同只是顺序不同,所以公司若想对财务危机做出预警,需要重点关注这些财务指标. 本文将随机生存森林与 Cox 模型、后向逐步 Cox 模型和 Lasso-Cox 模型进行对比. 通过计算 C-index、Brier score 和 IBS 3 种模型预测评价指标,发现随机生存森林要优于 3 种 Cox 模型.

此外在随机生存森林模型下,根据公司的财务指标计算其生存函数和累积风险函数,从而得到财务危机的动态时点预测. 证券交易所可以据此随时评估公司的财务危机,必要时给公司预警,敦促其采取相应措施改善经营管理,故而将随机生存森林应用到财务危机预警中是可行的,且有很强的现实意义.

[参考文献]

[1] 李扬,李竟翔,马双鸽. 不平衡数据的企业财务预警模型研究[J]. 数理统计与管理,2016,35(5):893-906.

[2] 马超群,何文. 基于 Cox 的财务困境时点预测模型研究[J]. 统计与决策,2010(21):38-42.

[3] 王小燕,袁欣. 基于惩罚组变量选择的 Cox 财务危机预警模型[J]. 系统工程,2018,36(3):113-121.

[4] KALBFLEISCH J D,PRENTICE R L. The statistical analysis of failure time data[M]. 2nd ed. New Jersey:John Wiley & Sons,Inc,2002.

[5] SUN J. The statistical analysis of interval-censored failure time data[M]. New York:Springer,2006.

[6] 高珍,柯阿香,余荣杰,等. 基于随机生存森林的交通事件持续时间预测[J]. 同济大学学报(自然科学版),2017,45(9):1304-1310.

[7] ISHWARAN H,KOGALUR U B,BLACKSTONE E H,et al. Random survival forests[J]. The annals of applied statistics,2008,2(3):841-860.

[8] 王呈斌,方匡南,郑陈璐. 基于随机生存森林的房屋贷款逾期研究[J]. 上海金融,2020(2):59-63.

[9] MOGENSEN U B,ISHWARAN H,GERDS T A. Evaluating random forests for survival analysis using prediction error curves[J]. Journal of statistical software,2012,50(11):1-23.

[10] KIM Y,PARK S,LEE J. Integrated survival model for predicting patent litigation hazard[J]. Sustainability,2021,13(4):1763.

[11] ISHWARAN H,KOGALUR U B,GORODESKI E Z,et al. High-dimensional variable selection for survival data[J]. Journal of the American statistical association,2010,105(489):205-217.

[12] 鲍新中,陶秋燕,傅宏宇. 基于变量聚类和 Cox 比例风险模型的企业财务预警研究[J]. 系统管理学报,2015,24(4):517-523,529.

[责任编辑:陆炳新]