Statistics and Application 统计学与应用, 2018, 7(2), 99-110
Published Online April 2018 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2018.72013

Hans汉斯

Prognosis of Breast Cancer Based on Cox Proportional Hazards Regression Model, LASSO and Survival Tree

Li Wang, Juan Zhang

North China Electric Power University, Beijing
Email: wangli2_3_6@163.com

Received: Mar. 13th, 2018; accepted: Apr. 1st, 2018; published: Apr. 8th, 2018

Abstract

Traditional pathological examination methods are not enough to predict the treatment outcome of breast cancer. Therefore, it is of great significance to study the pathogenesis of breast cancer by molecular biology. By predicting the risk of recurrence in patients with breast cancer, high-risk cancer patients can benefit from adjuvant therapy, while low-risk cancer patients can be protected from unnecessary treatment. The microarray data of ER+ breast cancer and ER− breast cancer were analyzed in this paper. Univariate Cox proportional hazards regression mode was used to preliminary screening the genes, then the LASSO was further used to screen the genes and applied the genes to the survival tree for prediction and classification, Kaplan-meier curve and log-rank test were used to prove the validity of the result. The model in this paper has a good prediction effect in the classification of breast cancer patients. Some of the genes we screened have been reported in the relevant literature, indicating that it is closely related to the occurrence and development of breast cancer. Other genes need further experiments to verify the role they play in breast cancer.

Keywords

Cox Proportional Hazards Regression Model, Kaplan-Meier Curve, Log-Rank Test, LASSO, Survival Tree

基于Cox比例风险回归模型、LASSO与生存树的乳腺癌预后

王　莉，张　娟

华北电力大学，北京
Email: wangli2_3_6@163.com

文章引用: 王莉, 张娟. 基于 Cox 比例风险回归模型、LASSO 与生存树的乳腺癌预后[J]. 统计学与应用, 2018, 7(2): 99-110. DOI: 10.12677/sa.2018.72013

收稿日期：2018年3月13日；录用日期：2018年4月1日；发布日期：2018年4月8日

摘　要

传统的病理检查方法不足以预测乳腺癌的治疗结果，因此从分子生物学上研究其发病机制具有重要意义。通过对乳腺癌患者复发风险的预测，高风险标记的肿瘤患者可以从辅助治疗中获益，而低风险标记的患者可免遭不必要的治疗。本文分别对ER+乳腺癌和ER−乳腺癌的基因芯片数据进行分析，采用单因素Cox比例风险回归模型初步筛选基因，然后进一步使用LASSO方法对基因进行筛选，再利用这些基因通过生存树方法对患者进行预测和分类。本文使用Kaplan-meier曲线及对数秩检验对结果进行验证。本文的模型对乳腺癌复发风险具有良好的预测效果，所筛选出的基因部分已被相关文献报道其确实与乳腺癌的发生和发展密切相关，其它基因尚需进一步实验来验证其在乳腺癌中发挥的作用。

关键词

Cox比例风险回归模型，Kaplan-Meier曲线，对数秩检验，LASSO，生存树

1. 引言

近年来，全球乳腺癌发病率已居女性恶性肿瘤之首[1]。乳腺癌是性激素受体依赖的肿瘤之一，雌激素的存在可促进此类癌细胞的生长和增值[2] [3]。目前的方法尚不足以预测乳腺癌的治疗结果，即使是具有相同病历类型的患者，在经过手术与放射治疗后，预后也可能存在很大差别，这是因为乳腺癌具有异质性[4]，因此当前关于乳腺癌的分类标准还有待提高。而基因表达谱对肿瘤患者的分类和预后是有效的，在其海量数据中，使用数理统计模型有效地挖掘信息也受到业界越来越多的关注。对于分类问题，支持向量机、贝叶斯网络、人工神经网络、决策树等机器学习方法在各领域中都得到广泛的应用，并取得了良好的预测效果。但对于肿瘤患者来说，这些方法考虑了事件的结果而未充分应用出现这一结果所经历的时间，因此本文在此选择了对肿瘤患者更优的预后分组方法——Cox 比例风险回归模型、LASSO 以及生存树。本文利用乳腺癌基因芯片数据筛查影响乳腺癌患者预后的基因，通过筛选基因对患者进行分类可以发现不同类别患者的基因特征。在未来的乳腺癌治疗中，以期可对患者选择具有针对性的基因治疗方案，从而提高乳腺癌患者的生命质量。

2. 研究对象和方法

2.1. 研究对象

本文下载 GEO 数据库(https://www.ncbi.nlm.nih.gov/geo)乳腺癌基因芯片数据(GSE2034)，该数据集共含有 286 个样本。其中 209 例为雌激素受体阳性(ER+)患者，该组患者随访时间为 2 个月至 171 个月，中位随访时间为 86 个月，本组 80 例患者出现复发；77 例为雌激素受体阴性(ER−)患者，该组患者随访时间为 6 个月至 161 个月，中位随访时间为 84 个月，本组 27 例患者出现复发。上述患者均为淋巴结阴性，每个样本均含有 22,283 个探针。

2.2. 研究方法

本文分别对 ER+乳腺癌样本和 ER−乳腺癌样本进行研究。对于 ER+乳腺癌患者，将 209 个样本随机分为训练集和测试集，其中训练集含 90 个样本，测试集含 109 个样本；对于 ER−乳腺癌患者，将 77 个样本同样进行随机分组，其中训练集含 42 个样本，测试集含 35 个样本。训练集用于模型的建立，测试集用于检验训练好的模型的分辨能力。本文对 ER+与 ER−乳腺癌基因芯片数据的分析采用 R 语言编程来实现。基因初步筛选使用单因素 Cox 比例风险回归模型，随后使用 LASSO 方法进一步筛选基因并建立生存树，使用 Kaplan-meier 曲线和对数秩检验对分类结果进行验证，以 $P < 0.05$ 为差异具有统计学意义。

3. 模型建立

3.1. 数据预处理

本文下载的乳腺癌基因表达矩阵如表 1 所示，行名表示探针，列名为每一例患者的编号；表 2 为患者基本信息。将两个表格通过患者编号进行合并，并通过“Status”将患者分为 ER+乳腺癌组与 ER−乳腺癌组分别进行研究。

Table 1. The microarray data of breast cancer
表 1. 乳腺癌基因芯片数据

ID_REF	GSM36777	GSM36778	GSM36779	GSM36780	GSM36781	...
1007_s_at	3848.1	6520.9	5285.7	4043.7	4263.6	...
1053_at	228.9	112.5	178.4	398.7	417.7	...
117_at	213.1	189.8	269.7	312.4	327.1	...
121_at	1009.4	2083.3	1203.4	1104.4	1043.3	...
1255_g_at	31.8	145.8	42.5	108.2	69.2	...
1294_at	551.5	802.8	557.5	568.5	653.2	...
1316_at	176.7	278.4	183.3	187.7	185.8	...
1320_at	11.9	28.3	56.4	42.1	21.8	...
1405_i_at	309.3	449	101.9	899.1	3629.3	...
1431_at	49.9	122.9	85.9	90.7	96	...
...	...	...	...	...	...	...

Table 2. The basic information of breast cancer
表 2. 乳腺癌患者基本信息

ID	Time	Relapse	Status
GSM36777	79	0	ER+
GSM36778	50	1	ER+
GSM36779	132	0	ER+
GSM36780	84	0	ER-
GSM36781	147	0	ER+
GSM36782	66	0	ER+
GSM36783	52	0	ER+
GSM36784	57	1	ER+
GSM36785	57	0	ER+
GSM36786	66	0	ER+
...	...	...	...

将上述分组后的表格分别与表 3 所示的探针与基因匹配表结合进行整理，转换为每一个基因的表达值进行研究。若探针与基因为“一对一”的关系(即一个探针对应一个基因)，则将相应表达值作为基因的表达值；若探针与基因为“一对空”或“一对多”，此时由于不能确定探针对应的是哪个基因的表达值，因此将其删除；若探针与基因为“多对一”，则取表达量较高的值作为此基因的表达值。经处理之后，ER+乳腺癌组与 ER−乳腺癌组分别得到 12,548 和 11,923 个基因的表达值。

初始数据一般都具有冗余性、不完整性和不规范性，无法直接进行数据分析。一些无意义的数据的存在会严重影响算法的执行，若存在噪音干扰，还会造成结果的偏差。因此，对不理想的原始数据预处理是进行数据分析的首要步骤。为了去除芯片间的系统误差，本文对数据进行了分位数标准化[5]；同时为了减少背景噪音，将小于 50 的基因表达值赋值为 50；接着对数据进行以 2 为底的对数化变换；再将变异系数小于 3%的基因剔除，此时 ER+乳腺癌组与 ER−乳腺癌组分别剩余 11,960 和 11,846 个基因的表达值。最后对每一个基因表达值进行编码，计算每组全部基因表达值的 25%、50%、75%分位数，小于等于 25%分位数的编码为 1，大于 25%分位数且小于等于 50%分位数的编码为 2，大于 50%分位数且小于等于 75%分位数的编码为 3，大于 75%分位数的编码为 4 [6]。经过预处理的基因编码矩阵如表 4 与表 5 所示。

Table 3. The matching table of probes and genes
表 3. 探针与基因匹配表

ID_REF	Gene
1053_at	RFC2
117_at	HSPA6
121_at	PAX8
1255_g_at	GUCA1A
1316_at	THRA
1320_at	PTPN21
1405_i_at	CCL5
1431_at	CYP2E1
1438_at	EPHB3
1487_at	ESRRA
...	...

Table 4. The encoding matrix of ER+ breast cancer
表 4. ER+乳腺癌基因编码值矩阵

ID	A1CF	A2M	A4GALT	A4GNT	AAAS	...
GSM36777	2	4	2	2	2	...
GSM36778	3	4	1	2	1	...
GSM36779	2	4	1	2	2	...
GSM36781	3	4	1	2	1	...
GSM36782	2	4	2	2	2	...
GSM36783	2	4	1	2	2	...
GSM36784	3	4	2	2	1	...
GSM36785	2	4	1	2	2	...
GSM36786	2	4	1	2	1	...
GSM36787	3	4	1	2	1	...
...	...	...	...	...	...	...

Table 5. The encoding matrix of ER− breast cancer
表 5. ER−乳腺癌基因编码值矩阵

ID	A1CF	A2M	A4GALT	A4GNT	AAAS	...
GSM36780	2	4	1	2	1	...
GSM36788	3	4	1	2	2	...
GSM36791	3	4	1	2	1	...
GSM36793	3	4	1	2	2	...
GSM36795	3	4	2	2	2	...
GSM36797	3	4	1	1	1	...
GSM36798	2	4	1	2	1	...
GSM36800	2	4	2	2	3	...
GSM36808	3	4	2	2	1	...
GSM36809	2	4	2	1	1	...
...	...	...	...	...	...	...

3.2. 单因素 Cox 比例风险回归模型初步筛选基因

Cox 比例风险回归模型[7]是由英国的生物统计学家 Cox D R 提出的比例风险模型。风险函数(hazard function)是描述生存时间分布的一个重要函数。如终点事件为死亡(复发)，风险函数表示 t 时刻仍存活的病人在 t 时刻的瞬间死亡(复发)率：

$$h(t)=\lim_{\Delta t\to 0}\frac{\text{在时间}t\text{生存的病人死(复发)于区间}(t,t+\Delta t)\text{的概率}}{\Delta t}$$

称为瞬时死亡(复发)率或条件死亡(复发)速率。

Cox 提出的比例风险模型是：病人具有 $X_{i1},X_{i2},\cdots,X_{ip}$ 的伴随变量值，则第 i 名病人生存(复发)到时间 t 的风险函数是基础风险函数与自变量的函数的乘积：

$$h_i(t)=h_0(t)\times f\left(\beta_1X_{i1}+\cdots+\beta_pX_{ip}\right)$$

$h_0(t)$ 即当所有的伴随变量都为 0 时的风险函数。其中定义伴随变量的函数 $f\left(x_{ij}\beta\right)$ 为指数形式，因此

$$h_i(t)=h_0(t)\times\exp\left(\beta_1X_{i1}+\cdots+\beta_pX_{ip}\right)$$

$$\ln\frac{h_i(t)}{h_0(t)}=\beta_1X_{i1}+\cdots+\beta_pX_{ip}$$

Cox 模型是一个风险比对数的线性模型，β_j 实际意义为当伴随变量 X_j 每改变一个单位时所引起的相对风险度的自然对数的改变量。β_j 不仅反映出协变量的作用强度，而且反映它的作用方向[8]。

我们对训练集数据使用单因素的 Cox 比例风险回归模型进行基因初步筛选，对于 ER+乳腺癌组，以 $P<0.05$ 作为入选标准。经过筛查，有 999 个基因对 ER+乳腺癌的复发具有显著性影响，结果见表 6 (结果保留两位小数)。如基因 ABCB8，其β值为 0.96，e^{β}值为 2.61，表明每当 ABCB8 的编码值增加一个单位，其复发风险比扩大到原来的 2.61 倍，说明对于 ER+乳腺癌患者来说，ABCB8 是其复发的风险基因；对于基因 ABLIM1，其β值为−0.97，e^{β}值为 0.38，表明每当 ABCB8 的编码值增加一个单位，其复发的风险比缩小到原来的 0.38，说明对于 ER+乳腺癌患者来说，ABLIM1 是其复发的保护基因。

对于 ER−乳腺癌组，以 $P<0.001$ 作为入选标准(在此分别使用 $P<0.05$ 、 $P<0.01$ 、 $P<0.001$ 为限定条件对基因进行筛选，通过对最终结果进行比较，此步将 P 值严格限定为 0.001)。经过筛查，有 13 个基因对 ER−乳腺癌的复发具有显著性影响，结果见表 7。

Table 6. Primary screening genes by univariate Cox proportional hazards regression model (ER+)
表 6. 单因素 Cox 比例风险模型初步筛选基因(ER+)

Gene	coef	exp(coef)	z	Pr(>|z|)	lower.95	upper.95
ABCB8	0.96	2.61	2.14	0.03	1.08	6.28
ABCC1	0.83	2.28	2.14	0.03	1.07	4.86
ABCC5	1.10	3.00	2.53	0.01	1.28	7.04
ABCE1	0.77	2.16	2.26	0.02	1.11	4.20
ABCF2	1.09	2.98	2.57	0.01	1.29	6.86
ABI3BP	3.37	29.16	2.92	0.00	3.03	280.37
ABLIM1	−0.97	0.38	−2.58	0.01	0.18	0.79
ABLIM3	−0.50	0.60	−2.31	0.02	0.39	0.93
BAG5	−1.17	0.31	−2.19	0.03	0.11	0.88
BARD1	0.72	2.05	2.74	0.01	1.23	3.42
...	...	...	...	...	...	...

Table 7. Primary screening genes by univariate Cox proportional hazards regression model (ER−)
表 7. 单因素 Cox 比例风险模型初步筛选基因(ER−)

Gene	coef	exp(coef)	z	Pr(>|z|)	lower.95	upper.95
ABCC2	1.40	4.08	3.66	0.000252594	1.92	8.65
AKT1	−1.72	0.18	−3.57	0.000362096	0.07	0.46
ATP4B	1.72	5.56	3.43	0.000608925	2.08	14.82
BTN3A2	−1.48	0.23	−3.37	0.000753891	0.10	0.54
CD200	1.60	4.95	3.40	0.000671633	1.97	12.44
FICD	−1.91	0.15	−3.57	0.000356132	0.05	0.42
MAPKAP1	−2.18	0.11	−3.29	0.000992481	0.03	0.41
PARP4	−1.97	0.14	−3.64	0.000276686	0.05	0.40
POLDIP2	−2.44	0.09	−3.79	0.000150915	0.02	0.31
RECQL5	−2.45	0.09	−3.36	0.000778005	0.02	0.36
...	...	...	...	...	...	...

3.3. 计算风险分数，对患者进行分类

通过上一步的筛查，有 999 个基因与 ER+乳腺癌患者复发显著相关，13 个基因与 ER−乳腺癌患者复发显著相关。本文使用所筛选的基因计算患者的风险分数，每例患者的风险分数为以相应 Cox 回归系数为权重的基因编码值的线性组合[9] [10]。接着使用风险分数对每一例患者进行高风险和低风险标记的划分，由于两组数据复发人数分别占每组人数的 38.28%和 35.06%，因此在此都选择风险分数的 60%分位数作为对患者进行分类的阈值(≤60%分位数为低风险标记；>60%分位数为高风险标记)，如表 8 与表 9 所示，其中 0 代表低风险标记，1 代表高风险标记。

本文将训练集建模过程中得到的 999 个基因和 13 个基因的 Cox 回归系数和分类阈值直接应用于每组的测试集，计算测试集中每例患者的得分并进行高风险标记和低风险标记的划分。

最后，对分类结果进行检验，如图 1 与图 2 Kaplan-meier 曲线显示，两组数据的训练集和测试集中，低风险标记的患者相较于高风险标记的患者都有着较高的中位未复发时间。对数秩检验(log-rank test)显示，低风险标记与高风险标记的 ER+乳腺癌患者的未复发率差异显著(训练集：$\chi^2 = 76.4$，$P < 0.001$；测试集：$\chi^2 = 6.6$，$P = 0.01$)，低风险标记与高风险标记的 ER−乳腺癌患者的未复发率差异同样也显著(训

Table 8. Calculate the risk scores and classify the ER+ breast cancer of training set
表 8. ER+乳腺癌训练集患者风险分数及分类情况

ID	Time	Relapse	Score	Group
GSM36945	92	0	−412.13	0
GSM37036	15	1	−197.87	1
GSM36850	121	0	−403.01	0
GSM36882	99	0	−404.50	0
GSM36998	5	1	−305.18	0
GSM37041	9	1	−202.26	1
GSM37035	14	1	−14.10	1
GSM36986	77	1	−276.82	1
GSM36913	109	0	−32.40	0
GSM36792	71	1	10.41	1
...	...	...	...	...

Table 9. Calculate the risk scores and classify the ER− breast cancer of training set
表 9. ER−乳腺癌训练集患者风险分数及分类情况

ID	Time	Relapse	Score	Group
GSM36886	94	0	−25.73	0
GSM36808	58	0	−25.73	0
GSM36926	24	1	−23.81	0
GSM37040	6	1	−11.71	1
GSM37042	14	1	−14.59	1
GSM36978	148	0	−25.73	0
GSM36891	87	0	−25.73	0
GSM37045	123	0	−27.33	0
GSM36875	7	1	−6.97	1
GSM37056	82	0	−25.61	0
...	...	...	...	...

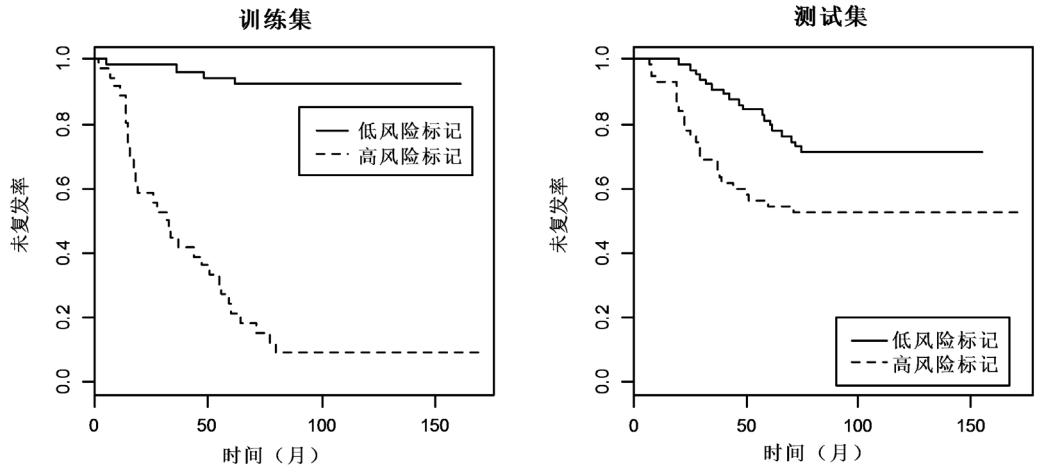

Figure 1. Kaplan-Meier estimates of survival of ER+ breast cancer according to the 999-gene signatures
图 1. ER+乳腺癌 999-基因标记 Kaplan-Meier 曲线

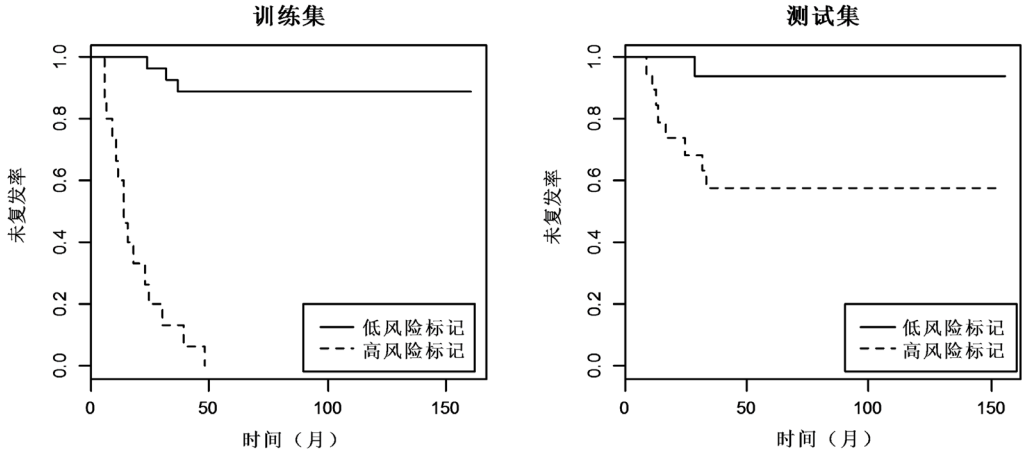

Figure 2. Kaplan-Meier estimates of survival of ER− breast cancer according to the 13-gene signatures
图 2. ER−乳腺癌 13-基因标记 Kaplan-Meier 曲线

练集：$\chi^2 = 49.1$，$P < 0.001$；测试集：$\chi^2 = 5.8$，$P = 0.02$)。

3.4. LASSO 方法筛选基因

由于数据中基因数量多而样本量较少，并且各基因间可能存在交互作用，因此选用 LASSO 方法对基因进行进一步筛选。LASSO 对于高维度、强相关、小样本的生存资料数据较为适用。LASSO 的基本思想是在回归系数的绝对值之和小于一个常数的约束条件下，使残差平方和最小化，从而使某些回归系数严格等于 0，来得到可以解释的模型[11] [12]。

$$\min_{\beta}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2, \quad \text{subject to } \sum_{j=1}^{p}\left|\beta_j\right| \le \lambda$$

该方法的估计参数

$$\hat{\beta}(\text{Lasso}) = \arg\min\left\{\sum_{i=1}^{n}\left\|y_i - \sum_{j=1}^{p}x_{ij}\beta_j\right\|^2 + \lambda\sum_{j=1}^{p}\left|\beta_j\right|\right\},\ \lambda\text{为调整参数}$$

随着λ的增加，$\sum_{j=1}^{p}\left|\beta_i\right|$项就会减小，这时候一些自变量的系数就逐渐被压缩为 0，以此达到对高维资料进行降维的目的。LASSO 方法的降维是通过惩罚回归系数的数量来实现的。

交叉验证法[13]是比较常用的推测估计调整参数λ的方法，如 k 折交叉验证。在 k 折交叉验证中，所有的数据观测值被大致分为 k 等份，然后轮流以其中所有可能的 k-1 份为训练集对数据进行拟合，剩下 1 份作为测试集，即测试集与训练集的观测值数目之比约为 1:(k-1)，一共计算 k 次，得到拟合测试集时的误差率(或其他指标)那样的 k 个指标，再做平均。在对每个模型都做一遍之后，最后选择误差率最小的模型。

我们选择 10 折交叉验证的方法来估计调整参数λ。对于 ER+乳腺癌组，当 $\ln\lambda = -1.50, \lambda = 0.22$ 时，误差率取得最小值，此时进一步筛选出了 5 个基因：CCDC69、KIF18A、KIF23、PLA2G15、RAI2；对于 ER−乳腺癌组，当 $\ln\lambda = -2.53, \lambda = 0.08$ 时，误差率取得最小值，此时进一步筛选出了 10 个基因：ABCC2、AKT1、ATP4B、CD200、FICD、PARP4、POLDIP2、RECQL5、THPO、XRCC1。

3.5. 生存树方法进行预测

生存树方法是由 Gordon 和 Olshen [14]在分类与回归树的基础上改进而成的，不同于普通的分类树，

其对截尾生存资料同样适用。通过递归分割计算，树逐渐进行生长。即选择某一截断点将根节点分为两部分，在分开的两个子样本中，其生存分布差异达到最大，即两组人群的预后相差达到最大。最优划分 s^* 满足

$$G\left(s^*,h\right)=\max_{s\in S_h} G\left(s,h\right)$$

这里，$G\left(s,h\right)$ 表示两样本 log-rank 检验统计量，其中 s 表示节点 h 内所有可能的截断方式，重复应用这个规则进行划分。

树的停止规则为：1) 结点内样本例数太小；2) 划分函数的测度不够充分，即划分得到的两个子结点的生存分布无差别。这时就得到了一棵初始树(initialtree) [15]。一般情况下初始树过大，会产生过度拟合的现象，在对未来样本进行预测时较为不准确，而且不容易解释，因此，需要通过剪枝过程对初始树的节点进行删减，控制树的复杂度[16]。

对于 ER+乳腺癌，本文使用所筛选的 5 个基因对训练集数据建立生存树模型，并将结果应用于测试集，得到了如图 3 所示的一棵生存树。将第 1、2、3、4、5、7 个叶节点作为低风险标记，第 6、8、9 个叶节点作为高风险标记。

对于 ER−乳腺癌，本文使用所筛选的 10 个基因对训练集数据建立生存树模型，进一步筛选出了 5 个基因：AKT1、CD200、FICD、THPO、XRCC1。将此结果应用于测试集，得到了如图 4 所示的一棵生存树。然后将第 1、2、3、5、6 个叶节点作为低风险标记，第 4、7、8 个叶节点作为高风险标记。

同样，对分类结果进行检验，如图 5 与图 6 Kaplan-meier 曲线显示，两组数据的训练集和测试集中，低风险标记的患者相较于高风险标记的患者都有着较高的中位未复发时间。对数秩检验显示，低风险标记与高风险标记的 ER+乳腺癌患者的未复发率差异显著(训练集：$\chi^2=60.2$，$P<0.001$；测试集：$\chi^2=9$，$P=0.003$)，低风险标记与高风险标记的 ER−乳腺癌患者的未复发率差异同样也显著(训练集：$\chi^2=52.3$，$P<0.001$；测试集：$\chi^2=4.3$，$P=0.04$)。

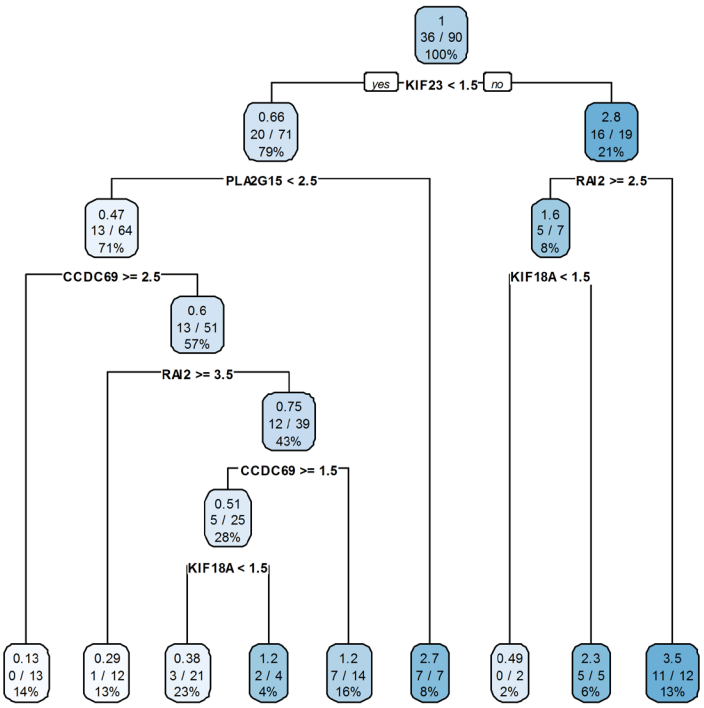

Figure 3. Survival tree of ER+ breast cancer
图 3. ER+乳腺癌生存树

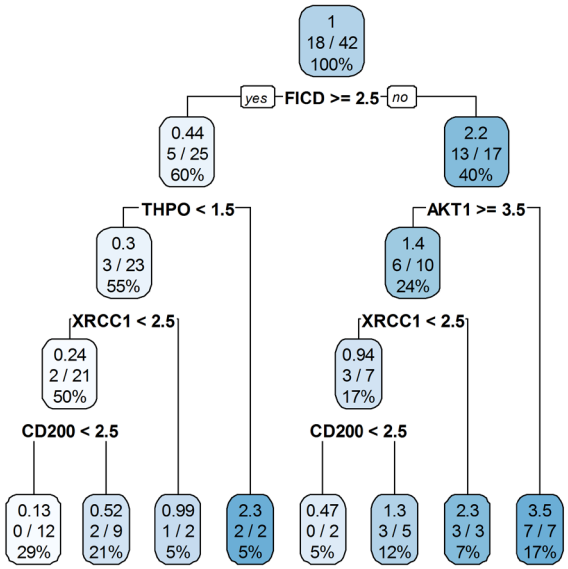

Figure 4. Survival tree of ER− breast cancer
图 4. ER−乳腺癌生存树

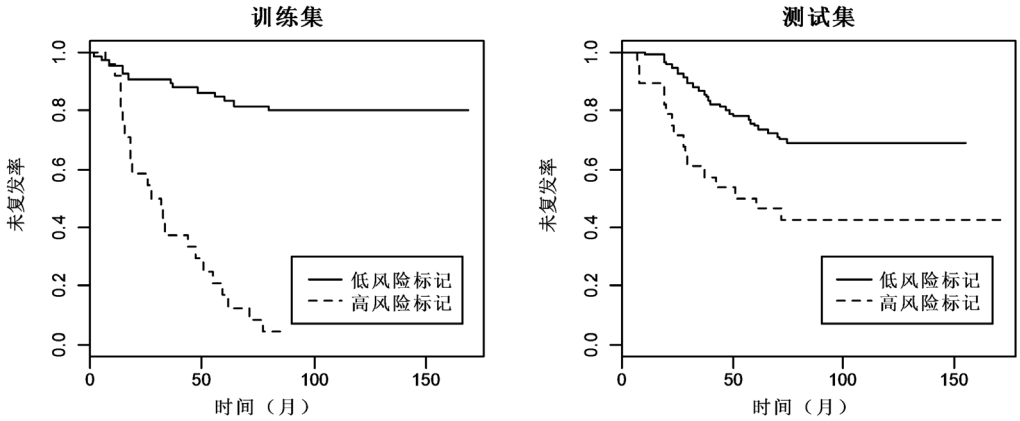

Figure 5. Kaplan-Meier estimates of survival of ER+ breast cancer according to the 5-gene signatures
图 5. ER+乳腺癌 5-基因标记 Kaplan-Meier 曲线

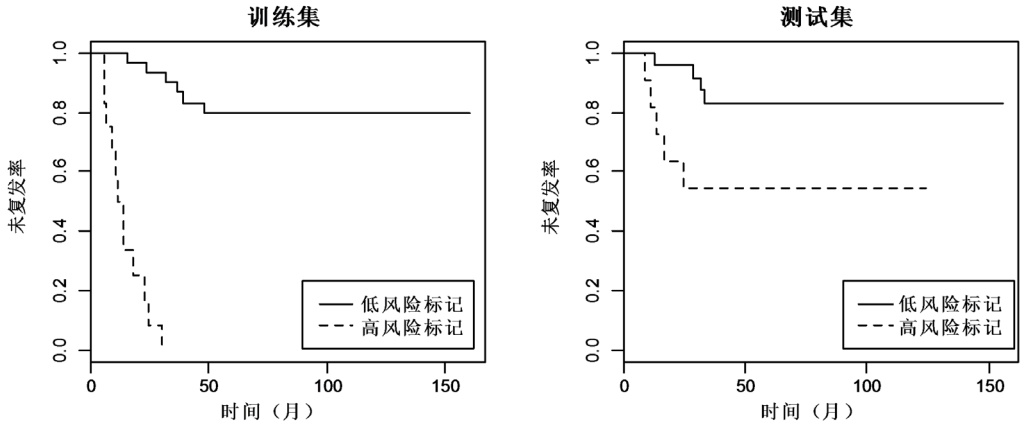

Figure 6. Kaplan-Meier estimates of survival of ER− breast cancer according to the 5-gene signatures
图 6. ER−乳腺癌 5-基因标记 Kaplan-Meier 曲线

4. 讨论

本研究对互联网上下载得到的乳腺癌基因芯片实验数据建模分析，分别对 ER+乳腺癌和 ER−乳腺癌患者的复发进行了研究，筛选对其复发具有影响的基因。对于高维生存数据，采用 Cox 比例风险回归模型进行了初步筛选，经检验，由此步所筛选基因编码值的线性组合所进行的分类效果差异显著。由于我们希望可用较少的基因达到较优的预测效果，因此继续进行下一步筛选。LASSO 方法在处理高维数据的生存分析方面效果显著，因此将由 Cox 比例风险回归模型筛选出的基因应用到 LASSO 方法分析，得到了对 ER+乳腺癌和 ER−乳腺癌复发有显著性影响的基因，且各基因间可能存在交互作用。我们将其应用于生存树对患者进行高风险标记与低风险标记的分类。为了验证所建立模型的有效性，本文使用测试集数据进行了检验，最终在对 ER+乳腺癌与 ER−乳腺癌两组数据复发风险的预测中，结果均具有统计学意义，且得到如下结论：

1. CCDC69、KIF18A、KIF23、PLA2G15、RAI2 对 ER+乳腺癌的复发具有显著性影响；

2. AKT1、CD200、FICD、THPO、XRCC1 对 ER−乳腺癌的复发具有显著性影响。

在 ER+乳腺癌最后筛选的 5 个基因中，有 2 个基因在相关文献中已被证明在乳腺癌中发挥着重要作用。Zhang C，Zhu C 等[17]的结果表明 Kif18a 参与乳腺癌中且可作为乳腺癌的潜在治疗靶点；Werner S，Borgmann K 等[18]发现 RAI2 功能的丧失与有丝分裂保真度降低有关，除了维持激素依赖性乳腺肿瘤的分化外，RAI2 还作为维持基因组完整性的一般肿瘤抑制因子。在 ER−乳腺癌最后筛选的 5 个基因中，有 3 个基因在相关文献中已被证明在乳腺癌中发挥着重要作用。Kabraji S, Solé X 等[19]表明 AKT1 低沉降癌细胞可能成为三重阴性乳腺癌患者新辅助化疗后持续存在的非遗传细胞状态，并值得进一步研究。Spears M，Cunningham C A [20]的研究也显示，AKT1 激活与早期乳腺癌的不良预后有关。Erin N，Podnos A 等[21]的研究结果支持 CD200 表达可以改变免疫反应的假说，并且可以抑制诱导全身和局部炎症反应的肿瘤细胞的转移生长。增加 CD200 活性/信号可能是治疗侵袭性乳腺癌的重要治疗策略。Moullan N，Cox D G 等[22]的研究显示 XRCC1 基因多态性的不同的组合似乎与乳腺癌风险增加有关，或与某些乳腺癌患者出现不良放疗反应的可能性相关。

CCDC69、KIF23、PLA2G15 这 3 个基因在 ER+乳腺癌中发挥的作用及 FICD 和 THPO 在 ER−乳腺癌中发挥的作用仍需通过进一步生物实验验证，其可能为 ER+乳腺癌和 ER−乳腺癌复发的潜在影响因子。总之，由我们筛选的基因对 ER+与 ER−乳腺癌患者风险高低的分类与患者临床结果密切相关。希望这些基因能为进一步实验提供理论依据，以研究这些基因在乳腺癌的预后中发挥的潜在作用。

基因项目

国家自然科学基金(11271125)。

参考文献

[1] 蒋定锋, 高峻, 赵耐青. 乳腺癌基因芯片数据分析[J]. 复旦学报(医学版), 2005, 32(2): 169-172.

[2] Higa, G.M. and Fell, R.G. (2013) Sex Hormone Receptor Repertoire in Breast Cancer. *International Journal of Breast Cancer*, **2013**, 284036.

[3] 陈慧, 莫淋, 徐晓帆, 等. 雌激素受体阳性乳腺癌预后的相关因素分析[J]. 临床肿瘤学杂志, 2015, 20(4): 333-337.

[4] 刘宁. 乳腺癌基因分型的研究进展[J]. 中国普通外科杂志, 2010, 19(5): 556-559.

[5] Bolstadm B,M,, Irizarrym R,A,, Astrandm M,, *et al.* (2003) A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Variance and Bias. *Bioinformatics*, **19**, 185-193. https://doi.org/10.1093/bioinformatics/19.2.185

[6] Hsiao, L.L. (2007) A Five-Gene Signature and Clinical Outcome in Non-Small-Cell Lung Cancer. *New England Jour-*

nal of Medicine, **356**, 11. https://doi.org/10.1056/NEJMoa060096

[7] Cox, D.R. (1992) Regression Models and Life-Tables. Breakthroughs in Statistics. Springer, New York, 187-220. https://doi.org/10.1007/978-1-4612-4380-9_37

[8] 李元章，何春雄. 实用生存模型：不完全数据分析[M]. 广州：华南理工大学出版社, 2015: 95-103.

[9] Warnke, R. (2004) Prediction of Survival in Diffuse Large-B-Cell Lymphoma Based on the Expression of Six Genes. *The New England Journal of Medicine*, **350**, 1828-1837. https://doi.org/10.1056/NEJMoa032520

[10] Beer, D.G., Kardia, S.L., Huang, C.C., *et al*. (2006) Gene-Expression Profiles Predict Survival of Patients with Lung Adenocarcinoma. *The Journal of Evidence-Based Medicine*, **8**, 816-824.

[11] Tibshirani, R. (1997) The Lasso Method for Variable Selection in the Cox Model. *Statistics in Medicine*, **16**, 385-395. https://doi.org/10.1002/(SICI)1097-0258(19970228)16:4<385::AID-SIM380>3.0.CO;2-3

[12] 闫丽娜，覃婷，王彤. LASSO 方法在 Cox 回归模型中的应用[J]. 中国卫生统计, 2012, 29(1): 58-60, 64.

[13] Simon, N., Friedman, J., Hastie, T., *et al*. (2011) Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. *Journal of Statistical Software*, **39**, 1. https://doi.org/10.18637/jss.v039.i05

[14] Gordon, L. and Olshen, R.A. (1985) Tree-Structured Survival Analysis. *Cancer Treatment Reports*, **69**, 1065-1069.

[15] 郎素平，余红梅，王彤，等. 生存树方法及其在预后分析中的应用[J]. 中国卫生统计, 2006, 23(1): 13-15.

[16] Atkinson, E.J. and Therneau, T.M. (2000) An Introduction to Recursive Partitioning Using the RPART Routines. Rochester Mayo Foundation.

[17] Zhang, C., Zhu, C., Chen, H., *et al*. (2010) Kif18A Is Involved in Human Breast Carcinogenesis. *Carcinogenesis*, **31**, 1676-1684. https://doi.org/10.1093/carcin/bgq134

[18] Werner, S., Borgmann, K., Pantel, K., *et al*. (2016) Abstract 2733: Novel Function of the RAI2 Protein in Genomic Integrity of Breast Cancer Cells. *Cancer Research*, **76**, 2733-2733.

[19] Kabraji, S., Sole, X., Ying, H., *et al*. (2017) AKT1low Quiescent Cancer Cells Persist after Neoadjuvant Chemotherapy in Triple Negative Breast Cancer. *Breast Cancer Research*, **19**, 88.

[20] Spears, M., Cunningham, C.A., Taylor, K.J., *et al*. (2012) Proximity Ligation Assays for Isoform-Specific Akt Activation in Breast Cancer Identify Activated Akt1 as a Driver of Progression. *Journal of Pathology*, **227**, 481-489. https://doi.org/10.1002/path.4022

[21] Erin, N., Podnos, A., Tanriover, G., *et al*. (2015) Bidirectional Effect of CD200 on Breast Cancer Development and Metastasis, with Ultimate Outcome Determined by Tumor Aggressiveness and a Cancer-Induced Inflammatory Response. *Oncogene*, **34**, 3860-3870. https://doi.org/10.1038/onc.2014.317

[22] Moullan, N., Cox, D.G., Angele, S., *et al*. (2003) Polymorphisms in the DNA Repair Gene XRCC1, Breast Cancer Risk, and Response to Radiotherapy. *Cancer Epidemiology, Biomarkers & Prevention*, **12**, 1168-1174.