心脏衰竭致死相关因素的研究

2021 年 8 月

邱文杰　　陈清旺　　陈万驰
191870145　191870014　181840033

摘要：心脏衰竭对于人类健康构成了重大威胁，研究心脏衰竭致死的相关因素对于疾病的治疗和预防具有重要意义。基于原始数据集，本文关注了 12 个相关因素，从三个角度递进式的进行分析。首先，对数据进行可视化处理，直观的展现相关关系；其次，利用统计学方法分析了各因素与致死之间的相关性关系，并利用 Lasso 方法得到了更为重要的因素。最后，分别使用机器学习中的逻辑回归方法, 支持向量机（SVM）和随机森林三种模型构建分类器，训练得到用于预测的模型。

关键词: Lasso 方法，逻辑回归，支持向量机

1　基本介绍

心脑血管疾病是人类健康的第一大威胁，每年约有 1790 万人死于该病，占据了全球总死亡人数的 31%。心血管疾病导致易导致心脏衰竭，对人体健康构成直接威胁，分析心脏衰竭致死相关因素，预测心脏衰竭的致死率具有重要价值。相关因素包括年龄，患病时间，各种生理化学指标等等。

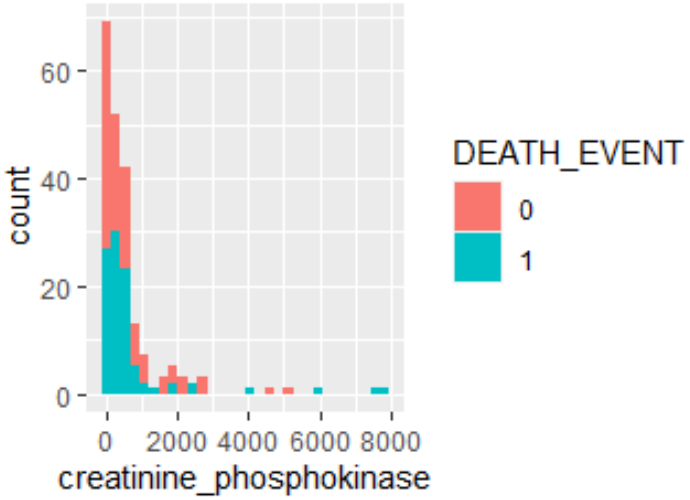
本文基于 Davide Chicco, Giuseppe Jurman 的研究调查的数据集，包括了约 300 个病人的信息，对于每个病人，关注了　12　个不同的因素，最后还包括了是否死亡这一重要信息。在下表中我们对 12 个特征进行了汇总解释。

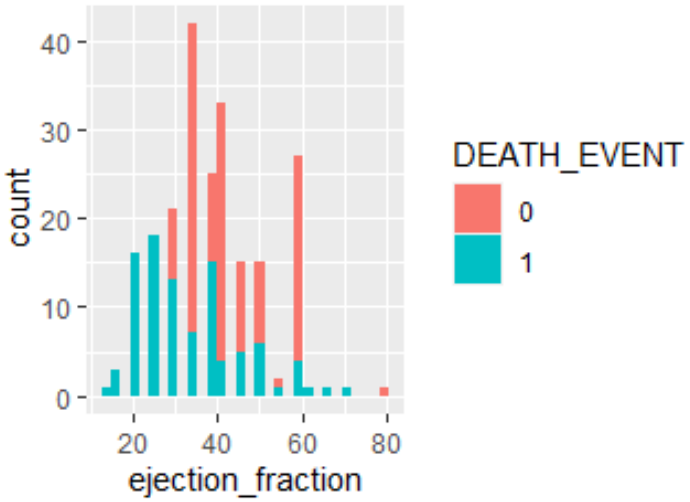
数据标签	解释	变量类型
age	年龄	数值
anaemia	是否患贫血症（0 表示不患，1 表示患病）	类别
creatinine_phosphokinase	肌酐 _ 磷酸激酶（化学指标）	数值
diabetes	是否患糖尿病（0 表示不患，1 表示患病）	类别
ejection_fraction	射出分率（化学指标）	数值
high_blood_pressure	是否高血压（0 表示否，1 表示是）	类别
platelets	血小板水平（化学指标）	数值
serum_creatinine	血清肌酸酐（化学指标）	数值
serum_sodium	血清钠（化学指标）	数值
sex	性别（0 表示女，1 表示男）	类别
smoking	是否吸烟（0 表示不吸，1 表示吸）	类别
time	患病时间（单位：年）	数值
DEATH_EVENT	是否死亡（0 表示存活，1 表示死亡）	类别

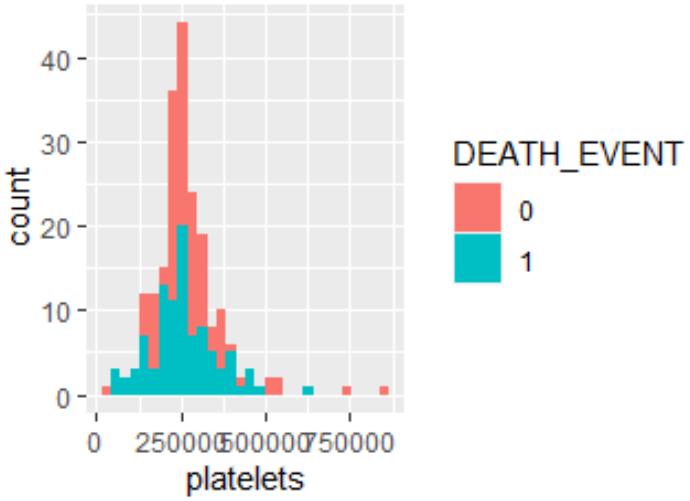
我们从三个角度递进式的展开研究，首先进行了数据的可视化处理，直观的展现了各因素与死亡之间的相关关系，也展示了各因素人数的分布关系。其次，利用统计方法分析了各因素与死亡之间的相关性关系，使用 Lasso 方法得到具有重要价值的因素。最后，筛除不重要因素，利用机器学习中的逻辑回归和 SVM 方法构建两种分类器，分别进行训练，得到可用于预测的模型。

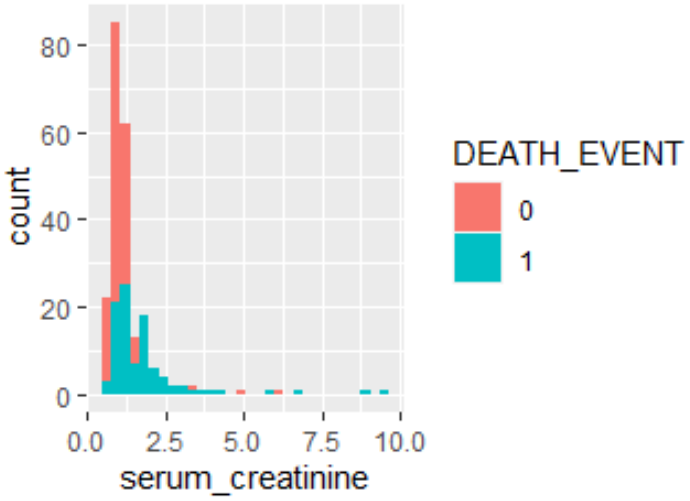
2 数据可视化

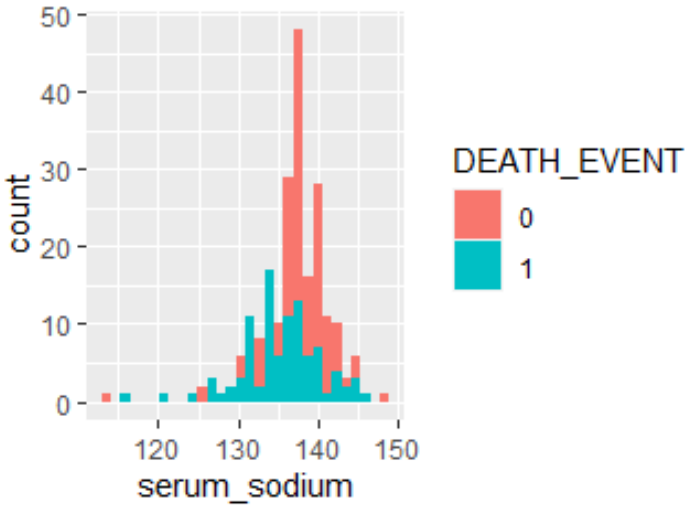
对几种数值型生化指标变量绘制频率分布直方图




count
40
30
20
10
0
20
40
60
80
ejection_fraction
DEATH_EVENT
0
1


count
40
30
20
10
0
0
250000
500000
750000
platelets
DEATH_EVENT
0
1


count
80
60
40
20
0
0.0
2.5
5.0
7.5
10.0
serum_creatinine
DEATH_EVENT
0
1



3 数据预处理

3.1 变量相关性分析

首先对单个自变量与因变量（死亡与否）进行相关性分析，由于因变量属于类别变量，自变量中既有数值变量也有类别变量。对于类别自变量，我们采用卡方检验的方法分析其与因变量的相关性；对于数值变量，我们采用 Spearman 等级相关分析方法。

经过两种计算方法，对于原假设（两个变量独立），得到的 p 值如表所示：

自变量	age	anaemia	creatinine_phosphokinase	diabetes
p 值	0.00014	0.30732	0.68423	0.92672
自变量	ejection_fraction	high_blood_pressure	platelets	serum_creatinine
p 值	4.50928E-7	0.21410	0.42606	3.61215E-11
自变量	serum_sodium	sex	smoking	time
p 值	0.00026	0.95605	0.93176	2.39326E-24

以 0.05 作为 p 值最高界限，可得 age,ejection_fraction,serum_creatinine,serum_sodium,time 这五个因素与死亡率相关性较强。

3.2 LASSO 回归

我们使用 LASSO 方法的目的是找出影响死亡率的最重要的一些因素，达到降维的效果，从而为进一步的预测工作节约成本。类似于一般的回归分析，LASSO 回归的目标也是最小化残差平方和，只是比一般的回归多了一个关于系数向量的一范数限制条件，LASSO 可以用下面的数学描述加以表示：

$$\min_{\beta_0,\beta}\{\sum_{i=1}^{N}(y_i-\beta_0-x_i^T\beta)^2\} \quad \text{s.t.}\sum_{j=1}^{p}|\beta_j|\leqslant t$$

其中共有 N 个样本，$x_i := (x_1, x_2, \ldots, x_p)_i^T$ 是第 i 个样本，y_i 为其对应的输出结果，每个样本是 p 维的。β_0 是常量系数，$\beta := (\beta_1, \beta_2, \ldots, \beta_p)$ 是系数向量，t 是预先设定的自有参数，决定正则化程度。

借助 python 的 sklearn.linear_model 中 LassoCV 和 Lasso 函数，我们计算出最合适的正则化参数以及回归系数，结果如下图:

```
利用Lasso交叉检验计算得出的最优alpha：2.2401020380085237
[ 0.00000000e+00  0.00000000e+00  2.62937954e-05  0.00000000e+00
 -0.00000000e+00 -0.00000000e+00 -2.17783734e-07  0.00000000e+00
 -0.00000000e+00 -0.00000000e+00 -0.00000000e+00 -2.79619080e-03]
Lasso回归后系数不为0的个数：3
Y = 0.0 * X0 + 0.0 * X1 + 0.0 * X2 + 0.0 * X3 + -0.0 * X4 + -0.0 * X5 + -0.0
* X6 + 0.0 * X7 + -0.0 * X8 + -0.0 * X9 + -0.0 * X10 + -0.003 * X11
```

对比前面十二个变量，可见 creatinine_phosphokinase,platelets 和 time 这三个因素都在 LASSO 中被保留了下来。

4 预测

4.1 逻辑回归

逻辑回归是一种广义的线性回归，常用来解决分类问题。设 X 是数据矩阵，逻辑回归的一般形式可表述为：

$$P = L(wX^T + b)$$

其中 w, b 为待确定的参数，L 为逻辑函数。机器学习中通常使用的逻辑函数为 sigmoid 函数:

$$L(z) = \frac{1}{1 + e^{-z}}$$

则可得到 $P(X) = L(wX^T + b) = \frac{1}{1 + e^{-(wX^T + b)}}$。

sigmoid 函数输出的 P 位于区间 $(0, 1)$，表示了在参数 w, b 条件下 P 取值为 1 的概率，对于二分类任务，通常在 $P \geqslant 0.5$ 时预测为类别 1，在 $P < 0.5$ 时预测为类别 0；为得到 w, b 的值，采用交叉熵函数作为损失函数，记为 $J(w, b)$，形式如下：

$$J(w, b) = -\frac{1}{n} \sum_{i=1}^{n} y^{(i)} log P(x^{(i)}) + (1 - y^{(i)}) log(1 - P(x^{(i)}))$$

逻辑回归的目标为：

$$\min_{w,b} J(w, b)$$

$x^{(i)}$ 表示第 i 组训练样本的取值，$y^{(i)}$ 表示第 i 组训练样本的真实类别，n 为训练样本总数。对于参数 w, b 在机器学习中有多种方法学习得到，常采用的一种方式是梯度下降法，具体表述如下:

$$w = w - \alpha \frac{\partial}{\partial w} J(w, b)$$

$$b = b - \alpha \frac{\partial}{\partial b} J(w, b)$$

其中 α 为学习率，是该学习任务中的超参数。

4.2 支持向量机 (SVM)

支持向量机是机器学习中常见的一种分类器，可用来完成二分类或多分类任务。支持向量机是一种最大间隔分类器，其目标是找到一个划分平面，使得样本点到划分平面的距离最大。转化为优化问题如下：

$$\min_{w,b} \frac{1}{2}\|w\|^2$$

$$s.t. y^{(i)}(wx^{(i)} + b) - 1 \geqslant 0, \quad i = 1, 2, \ldots, n$$

上述模型适用于线性可分的情况，对于线性不可分的情形，可通过核函数方法处理。在线性可分的上述模型中，求解 w, b 有许多成熟的算法包可供调用, 如 SMO(Sequential Minimal Optimization)。

4.3 随机森林

随机森林（Random Forests）是对样本进行训练和预测的分类器模型，它是一个包含多个决策树的分类器。

随机森林模型是 Bagging 方法的改进，Bagging 方法的基本流程为，给定一个包含 m 个样本的数据集后，通过 bootstrap 方法可以得到含 m 个样本的一个采样集，重复 bootstrap 方法 n 次，可以得到 n 个含 m 个样本的采样集，在每一个采样集上都可以训练得到一个基学习器，执行具体任务时，将基学习器的结果结合起来，对于分类则采用“简单投票法”，回归问题则采用“简单平均法”。随机森林在 Bagging 方法中训练决策树时引入了随机属性，不同于传统决策树划分属性时选择当前节点属性集合的最优属性，随机森林模型在每个决策树节点属性集合中选择一个子集，再从子集中进行进一步划分。

随机森林具有速度快，准确度高的优点，得到了广泛的应用。在 python 的 sklearn 包中可直接调用相关模型。

4.4 结果

数据集中的 DEATH_EVENT 是一个二值变量，表示了在病人是否死亡，依此作为目标变量，在上面的分析中，我们通过 LASSO 方法找到了对于 DEATH_EVENT 影响更为重要的特征。以下我们将通过逻辑回归、支持向量机，随机森林构建分类器，并在效率和准确率上进行比较。首先进行数据集的分割，按 7：3 的比例划分训练集与测试集。在训练集进行训练，测试集上进行模型比较评估。由于特诊变量的类型同时包括了整数、实型小数和二值变量，为方便模型的训练，对训练集的数据进行了归一化处理，归一化方式采用了均值归一化。比较上采用了准确率 (Accuracy), 查全率 (Recall), 查准率 (Precision) 三个指标。同时为了评估 LASSO 方法的性能，我们也针对数据集中的包括的全部特征进行了训练，得到了更为复杂的模型。将原始特征、相关性分析得出的五个特征、LASSO 得出的三个特征这三种数据集分别用上面三种分类器进行学习，最终得出如下的表格。

在上述三个模型的训练过程中，逻辑回归未进行正则化，支持向量机使用了核函数方法，随机森林中使用了 1000 棵子树，未设置高度限制。

使用三种不同模型进行预测，在上述三个指标评估下，具有一定的差别。不难发现，逻辑回归方法的准确率和查准率是最低的，这表示模型在预测的精度上表现不佳，但具有最高的查全率，在全部的死亡病例中，相对而言预测出了更多的“死亡”结果。支持向量机在查准率和准确率的表现都不如随机森林。

Logistic Regression	Accuracy	Precision	Recall
原始数据	0.83	0.67	0.75
相关性分析	0.86	0.69	0.83
LASSO	0.82	0.68	0.62

SVM	Accuracy	Precision	Recall
原始数据	0.81	0.73	0.46
相关性分析	0.87	0.71	0.83
LASSO	0.80	0.62	0.67

Random Forest	Accuracy	Precision	Recall
原始数据	0.87	0.77	0.71
相关性分析	0.86	0.74	0.71
LASSO	0.84	0.69	0.75

在实际应用中，我们往往希望能够对于一个病人进行更为准确的预测，而非关注在一个整体中能够更多精确预测“死亡”，因而在上述偏向性下，选择随机森林模型来进行分类和预测能有更好的效果。同时，虽然用相关性分析和 LASSO 方法选出部分特征构建的分类器准确率并不如原始数据高，但是考虑到这两种方法降低数据维度效果明显，能够有效节省成本，同时准确率差别很小，因此更应该选择这两种方法选取出的特征。

参考文献

[1] 周志华. 机器学习 [M]. Qing hua da xue chu ban she, 2016.

数据来源及代码

数据集链接：https:// www.kaggle.com/andrewmvd/heart-failure-clinical-data
代码链接：https://github.com/wingrise/Statistics-Homework