Final Project: Heart PCI Operation Effect Prediction

虞俊
万若斯

2015 年 1 月 20 日

摘要

本文研究了自变量有大量缺失时的分类问题和变量选择问题。本文采用两个阶段来求解该问题: 在第一个阶段中整理数据，首先删除缺失过多的变量，然后利用随机森林算法删除不显著变量，最后对剩下的变量的缺失数据进行填充; 第二个阶段为数据分析阶段，对已填补的数据集再利用随机森林算法进行变量选择，得到预测模型。

Keyword: 缺失数据, 随机森林

目录

1 分析目标

本文处理的数据是来自两个医院的 2581 个病人在接受心脏支架手术（PCI）前、手术中一共 75 项生理指标与环境因素和一项手术效果评价（是否处于无复流状态？无复流状态表示手术失败）。我们相信手术成功率与病人自身身体素质、手术前和手术中途各项生理指标以及外在环境都有很大关系。因此，我们期望能够通过统计手段，找到与手术效果（是否复流）高度相关的因素，并发展一种方法，通过检测病人的这些因素，评估其接受心脏支架手术的效果。通过对历史数据的统计，可以帮助医院在缺少相关严格的医学理论指导下的情况下，寻找未知的与 PCI 手术结果高度相关的因素，改进手术方法。并且可以识别出潜在的危险病人，并给予他们特别关注与看护，以降低手术风险。因此这是一项十分有意义的工作。

2 数据描述

待处理数据集一共有一共有 2581 个观测，涉及 73 个自变量，1 个因变量。其中自变量类型根据采集时间分为：入院即刻、PCI 手术前、PCI 手术中及其它（所在医院等）；根据数据类型分为定性变量（0、1 取值，例如男女）和定量变量（例如年龄）。自变量是定性变量，取值为 0、1，表示复流情况。我们主要要考虑入院即刻检查 43 个信息。注意到数据里有大量的缺失，因此在进行统计前，我们需要根据原数据的特征对数据集经行填补。

3 数据处理与数值结果展示

3.1 删除缺失过多变量

数据集中一共有 2581 个病人，涉及 43 个自变量，1 个因变量。因变量“无复流”有 7 个缺失，需要删除相应的 7 个观测。剩余 2574 次观测。从中进行简单随机抽样，抽取 2059 次观测作为训练集，剩余的 515 次观测作为最终的测试集用于计算模型的判错率。这里需要说明的是：每一次建模分析过程中都要重新随机抽取训练集和测试集，计算一次判错率。对于每一种方法，计算 100 次判错率，以平均判错率作为评价模型或方法的最终标准。

表 1 中为所有变量缺失情况的初步统计。其中 11 个自变量缺失的数据量超过 1000 , 由于这些变量的缺失数据接近一半，填补缺失十分困难，并且填补的数据对后期的建模分析也会有较大的干扰，因此，在这里我们人为的将缺失量超过 1000 的变量删除，剩余 32 个自变量。

变量名	缺失量	变量名	缺失量	变量名	缺失量	变量名	缺失量
性别	20	年龄	4	**身高**	2422	**体重**	2421
吸烟史	236	糖尿病史	123	高血压史	108	PCI 史	121
脑梗塞史	175	既往调脂药	127	既往阿司匹林	126	既往 ADP 拮抗剂	127
既往 ACEI	127	既往利尿剂	555	既往 β 受体阻滞剂	127	既往 CA 拮抗剂	127
梗死前心绞痛	106	收缩压	121	舒张压	117	心率	176
入院诊断	17	killip 分级	158	梗死部位	49	中性粒细胞	395
血红蛋白	1668	**白蛋白**	1667	肌酐	741	总胆固醇	447
甘油三酯	447	LDLC	457	HDLC	455	随机血糖	7
apoa1	2491	**apob**	2491	**LPa**	1677	高敏 C 反应蛋白	971
BNP	947	**TNI**	1667	PCI 前 CK	630	**PCI 前 CKMB**	1667
内皮素	2116	症状到 PCI 时间	28	无复流	7		

Table 1: 数据缺失情况，黑体标识的变量缺失超过 1000.

3.2 初步选择变量

为保证准确性，我们首先只对数据完整的样本运用随机森林[1] 以挑选重要因变量，再由这些重要因变量的数值分布填补缺失数据。数据完整的样本有 1232 个，我们直接使用 R 语言包“*VSURF*”中的 *VSURF.thres* 函数对预处理过的数据集做随机森林。表 2是一次数值计算的结果，经过初步筛选，剩余 17 个的变量。

变量名	随机血糖	甘油三酯	既往利尿剂	BNP	高敏 C 反应蛋白
因子重要性方差	0.0245	0.0064	0.0051	0.0047	0.0038
变量名	中性粒细胞	年龄	总胆固醇	LDLC	症状到 PCI 时间
因子重要性方差	0.0035	0.0032	0.0028	0.0026	0.0026
变量名	killip 分级	PCI 前 CK	糖尿病史	梗死部位	HDLC
因子重要性方差	0.0024	0.0021	0.0012	0.0011	0.0010

Table 2: 利用随机森林法挑选出的重要变量

3.3 缺失数据填充

数据的填补方法我们测试了多种方法，并与最简单的随机填补方法进行对比。

3.3.1 随机填补

随机填补方法十分简单：如果某一自变量的第 j 个变量缺失，那么从该变量已有的变量中任取一个填入。

3.3.2 均值填补

将完整样本集的每一个自变量已知样本的求平均值 μ，然后该自变量缺失的样本添上 $\mu+\epsilon$，其中 ϵ 为一个服从正态分布的随机扰动。

3.3.3 聚类填补

基于谱方法聚类的 PAM 算法[2] 可以对不完备的数据进行聚类。因此考虑首先运用 PAM 将初次挑选的重要因变量的全部样本分类，再在同一类中进行均值填补。聚类数目需要预先设定，经过实验，发现目标聚类数取为 6 时比较合适。

3.3.4 K 近邻填补

对一个记录有缺失的观测 X_i，找出与它最相近的 K 个完整观测 $\tilde{X}_j, j \in \{1,2,\ldots,K\}$（本文取 $K=10$）, 距离函数定为非缺失因变量维度的欧式距离。即

$$d(X_i,\tilde{X}_j)=\sqrt{\sum_{k\in\Lambda}(X_{ik}-\tilde{X}_{jk})^2}$$

其中 Λ 为 X_i 非缺失因变量的指标集。则 X_i 的缺失变量 X_{ik} 补全为 $\frac{\sum_{j=1}^{K}\tilde{X}_{ij}}{K}$。

3.4 再次挑选变量与预测

得到补全的数据集，再次运用随机森林挑选其中的重要因变量挑选变量见表格，并根据挑选的因变量，我们做回归预测因变量。因变量（不复流?）是一个逻辑值，因此，当回归结果大于 0.5 时，因变量为 1, 当不大于 0.5 时，因变量为 0。我们用测试集测试模型的预测效果。以下是四种分类方式的测试结果：

方法	0/0	1/0	0/1	1/1	correct rate
随机填充	398	10	67	39	0.8502
均值填充	397	11	67	39	0.8425
聚类填充	396	4	65	38	0.88568
K 近邻填充	397	11	66	40	0.8502

Table 3: 预测结果

其中,a/b 表示真实复流状态为 b 的样本预测复流状态为 a，correct rate 表示预测正确的样本数占总测试集的比例。

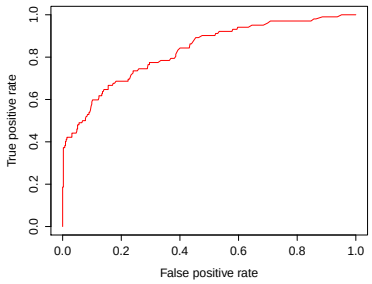

Figure 1: ROC 曲线 (随机)

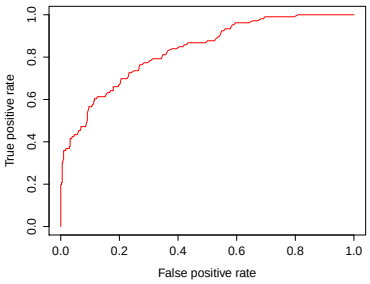

Figure 2: ROC 曲线 (均值)

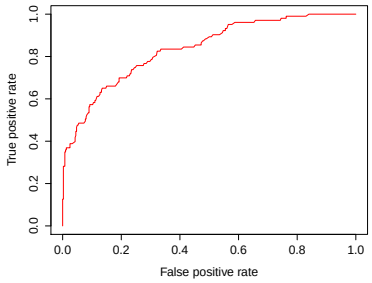

Figure 3: ROC 曲线 (聚类）

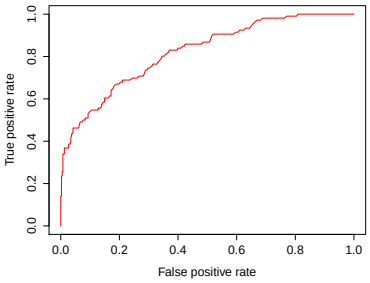

Figure 4: ROC 曲线 (K 近邻)

4 方法评估与讨论

从以上分析可知，这四种填补数据的方法效果十分接近，预测成功率都在 85% 左右，AUC 的取值也十分接近，但相对来说聚类方法填补数据预测成功率最高。另一方面，利用随机森林对数据集两重挑选两次重要变量再用回归预测的方式成功率很高。但是，通过表3发现，虽然对随机森林对预测不复流状态的病人成功率很高，但对复流状态的病人成功率很低不足 50%。这原因有两方面，一方面数据是与不复流的样本复流样本本身就少很多，因此相对而言不复流病人的预测会更不准确；另一方面我们仅分析了手术前的各项数据，而没有考虑手术中采集的数据，也没有考虑医院的因素，而这些是应该是反应手术成功与否的重要指标。

参考文献

[1] Genuer R, Poggi J M, Tuleau-Malot C. Variable selection using random forests[J]. Pattern Recognition Letters, 2010, 31(14): 2225-2236.

[2] Reynolds A P, Richards G, de la Iglesia B, et al. Clustering rules: a comparison of partitioning and hierarchical clustering algorithms[J]. Journal of Mathematical Modelling and Algorithms, 2006, 5(4): 475-504.