

Heart PCI Operation Effect Prediction

虞俊、万若斯

北京大学 数学科学学院

March 14, 2015

- 数据

- ▶ 待处理数据集一共有 2581 个观测，涉及43个入院即刻检查的变量和1 个因变量。
- ▶ 已知发生复流的案例有528个，对是否发生复流没有观测的有7个；
- ▶ 数据有大量缺失，完整的样本有1232个。

- 目标

- 我们主要要考虑入院即刻检查 43 个信息（即所有医院整合（术前）这张表中的信息）。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。

- 数据

- ▶ 待处理数据集一共有 2581 个观测，涉及43个入院即刻检查的变量和1 个因变量。
- ▶ 已知发生复流的案例有528个，对是否发生复流没有观测的有7个；
- ▶ 数据有大量缺失，完整的样本有1232个。

- 目标

- 我们主要要考虑入院即刻检查 43 个信息（即所有医院整合（术前）这张表中的信息）。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。

- 数据

- ▶ 待处理数据集一共有 2581 个观测，涉及43个入院即刻检查的变量和1 个因变量。
- ▶ 已知发生复流的案例有528个，对是否发生复流没有观测的有7个；
- ▶ 数据有大量缺失，完整的样本有1232个。

- 目标

- 我们主要要考虑入院即刻检查 43 个信息（即所有医院整合（术前）这张表中的信息）。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。

- 数据

- ▶ 待处理数据集一共有 2581 个观测，涉及43个入院即刻检查的变量和1 个因变量。
- ▶ 已知发生复流的案例有528个，对是否发生复流没有观测的有7个；
- ▶ 数据有大量缺失，完整的样本有1232个。

- 目标

- 我们主要要考虑入院即刻检查 43 个信息（即所有医院整合（术前）这张表中的信息）。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。

- 数据

- ▶ 待处理数据集一共有 2581 个观测，涉及43个入院即刻检查的变量和1 个因变量。
- ▶ 已知发生复流的案例有528个，对是否发生复流没有观测的有7个；
- ▶ 数据有大量缺失，完整的样本有1232个。

- 目标

- 我们主要要考虑入院即刻检查 43 个信息（即所有医院整合（术前）这张表中的信息）。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。

- 数据

- ▶ 待处理数据集一共有 2581 个观测，涉及43个入院即刻检查的变量和1 个因变量。
- ▶ 已知发生复流的案例有528个，对是否发生复流没有观测的有7个；
- ▶ 数据有大量缺失，完整的样本有1232个。

- 目标

- 我们主要要考虑入院即刻检查 43 个信息（即所有医院整合（术前）这张表中的信息）。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。

- ▶ 医院之间并无系统性差别 (no systematic difference)
- ▶ 所有数据都是随机缺失的 (missing at random)

- ▶ 医院之间并无系统性差别 (no systematic difference)
- ▶ 所有数据都是随机缺失的 (missing at random)

- step 1 预处理数据：删除缺失过多变量；
- step 2 初步选择变量：对数据完整的样本运用随机森林以挑选重要因变量；
- step 3 缺失数据填充：使用随机填补，均值填补，聚类填补，K近邻填补等方法填补初步选择选出的变量的缺失数据；
- step 4 模型评价：对填补好的数据集挑选变量并用随机森林进行最终预测，比对预测结果与真实结果，对模型做出评价

- step 1 预处理数据：删除缺失过多变量；
- step 2 初步选择变量：对数据完整的样本运用随机森林以挑选重要因变量；
- step 3 缺失数据填充：使用随机填补，均值填补，聚类填补，K近邻填补等方法填补初步选择选出的变量的缺失数据；
- step 4 模型评价：对填补好的数据集挑选变量并用随机森林进行最终预测，比对预测结果与真实结果，对模型做出评价

- step 1 预处理数据：删除缺失过多变量；
- step 2 初步选择变量：对数据完整的样本运用随机森林以挑选重要因变量；
- step 3 缺失数据填充：使用随机填补，均值填补，聚类填补，K近邻填补等方法填补初步选择选出的变量的缺失数据；
- step 4 模型评价：对填补好的数据集挑选变量并用随机森林进行最终预测，比对预测结果与真实结果，对模型做出评价

- step 1 预处理数据：删除缺失过多变量；
- step 2 初步选择变量：对数据完整的样本运用随机森林以挑选重要因变量；
- step 3 缺失数据填充：使用随机填补，均值填补，聚类填补，K近邻填补等方法填补初步选择选出的变量的缺失数据；
- step 4 模型评价：对填补好的数据集挑选变量并用随机森林进行最终预测，比对预测结果与真实结果，对模型做出评价

删除缺失过多变量

下表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: 数据缺失情况，黑体标识的变量缺失超过1000.

Genuer R (2010) 在他的Variable selection using random forests中提出了利用随机森林模型中的特征重要性 (VI) 来进行变量筛选的方法。其主要步骤为:

step 1 初步估计和排序

- a 对随机森林中的特征变量按照VI (Variable Importance) 降序排序。
- b 确定删除比例,从当前的特征变量中剔除相应比例不重要的指标,从而得到一个新的特征集。
- c 用新的特征集建立新的随机森林,并计算特征集中每个特征的VI,并排序。
- d 重复以上步骤,直到剩下m个特征。

step 2 根据1中得到的每个特征集和它们建立起来的随机森林,计算对应的袋外误差率(OOB err),将袋外误差率最低的特征集作为最后选定的特征集。

初步选择变量结果

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: 利用随机森林法挑选出的重要变量

随机填补

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。

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。

聚类填补

由于聚类算法CLARA(Clustering Large Applications)不考虑整个数据集, 而是选择数据的一小部分作为样本, 从而可以对不完备的数据进行聚类。因此考虑首先运用CLARA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。

CLARA 算法的步骤:

- (1) 对于 $i = 1$ 到 v (选样的次数), 重复执行下列步骤((2) \sim (3)) :
- (2) 随机地从整个数据库中抽取一个 N 个对象的样本, 调用PAM方法从样本中找出样本的 k 个最优的中心点。
- (3) 将这 k 个中心点应用到整个数据库上, 对于每一个非代表对象 O_j , 判断它与从样本中选出的哪个代表对象距离最近。
- (4) 计算上一步中得到的聚类的总代价。若该值小于当前的最小值, 用该值替换当前的最小值, 保留在这次选样中得到的 k 个代表对象作为到目前为止得到的最好的代表对象的集合。

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 $\tilde{X}_j, j \in \{1, 2, \dots, K\}$ （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 $\frac{\sum_{j=1}^K \tilde{X}_{ij}}{K}$ 。

最终模型建立与预测

数据集中一共有2581 个，因变量“无复流”有7 个缺失，删除相应的7 个观测。剩余2574 次观测。从中进行随机抽样，抽取2059 次观测作为训练集（其中包含1232个完整样本），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。

对训练集采取4折交叉验证的办法，利用随机森林对变量再次做出选择（基于填补好的数据集），并用随机森林对训练集数据建模。最后将所得模型用于测试集做预测。

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: 预测结果

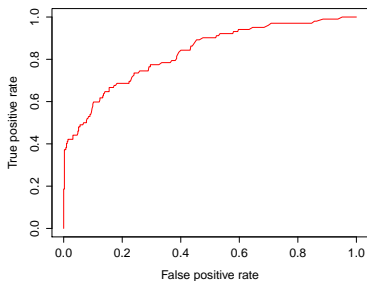


Figure: ROC曲线(随机)

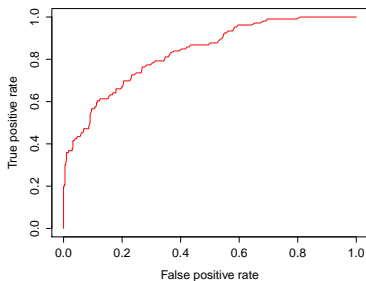


Figure: ROC曲线(均值)

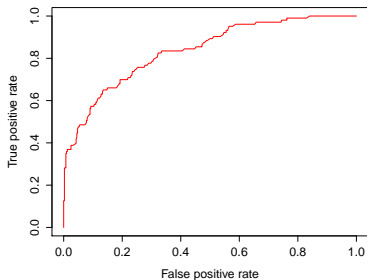


Figure: ROC曲线(聚类)

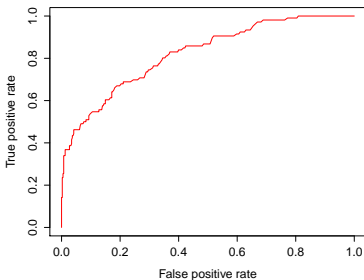










Figure: ROC曲线(K近邻)

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。

-  Genuer R, Poggi J M, Tuleau-Malot C. Variable selection using random forests[J]. Pattern Recognition Letters, 2010, 31(14): 2225-2236.
-  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.
-  Kaufman L R, Rousseeuw P. PJ (1990) Finding groups in data: An introduction to cluster analysis[J]. Hoboken NJ John Wiley & Sons Inc.
-  Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K. (2014). cluster: Cluster Analysis Basics and Extensions. R package version 1.15.2.
-  Robin Genuer, Jean-Michel Poggi and Christine Tuleau-Malot (2014). VSURF: Variable Selection Using Random Forests. R package version 0.8.2. <http://CRAN.R-project.org/package=VSURF>

-  Stef van Buuren, Karin Groothuis-Oudshoorn (2011). mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, 45(3), 1-67. <http://www.jstatsoft.org/v45/i03/>.
-  A. Liaw and M. Wiener (2002). Classification and Regression by randomForest. R News 2(3), 18–22.
-  Sing T, Sander O, Beerenwinkel N and Lengauer T (2005). "ROCR: visualizing classifier performance in R." Bioinformatics, 21 (20) , pp. 7881. <http://rocr.bioinf.mpi-sb.mpg.de>.