Heart PCI Operation Effect Prediction

统计学习 final project

陈翰轩 (1401110040), 邓思圆 (1401110045), 陆道旭 (1401210055), 徐来 (1401210065)

摘要

通过对在安贞医院进行心脏手术的患者数据进行分析, 我们期望根据患者的情况预测手术后无复流现象发生的概率. 数据集由 2581 个患者样本构成, 共有 73 个特征变量, 以及一个响应变量用以指示手术后是否有无复流现象, 从而这是一个分类问题. 具体的, 特征变量按数据获取时间分为三类: 入院时, 入院后至手术前和手术中. 我们的目标是通过前两类甚至仅第一类数据来进行预测. 数据集指标很杂且有很多缺失值, 因此数据清洗和缺失值填补是一道重要工序.

对预处理及填补后的数据, 我们首先通过 LASSO 等方法进行了特征选择, 剔除了一部分无显著影响的特征变量. 之后我们使用逻辑斯蒂回归 (LR), 支持向量机 (SVM), 神经网络 (ANN) 和随机森林 (RF) 等方法进行分类预测, 比较其结果, 择优选取了随机森林方法, 其入院即刻数据的分类准确率达 86%, 并进一步细致分析, 结合实际问题对模型进行解释.

更进一步, 无复流现象为 PCI 导致的并发症, 我们期望对无复流现象发生的预测尽可能准确 (例如若预测为会发生无复流现象, 则不进行手术), 甚至能全部准确预测, 并考虑在此基础上, 对有复流情形的预测准确率能达到多少 (或者反向考虑).

数据预处理与清洗

数据集中变量较杂, 其类型包括类别变量、连续变量和整数型变量. 此外, 数据中每项指标均有缺失, 缺失比例从 1% 到 90% 不等. 首先我们将对数据中的一些明显错误进行处理并对某些指标进行变换, 主要操作如下：

- 由于二值变量无复流是我们的预测目标, 故首先剔除变量无复流的无观测值的记录, 共 7 个.
- 由于有的指标缺失比例太高, 我们认为强行填补会对模型的预测效果造成影响, 故去掉缺失比例在 80% 以上的指标, 共有 6 个.
- 数据中存在一些变量的取值与说明不符, 比如 0-1 取值的变量取值为 2, 我们将这些样本剔除.
- 数据中存在一些记录中收缩压小于舒张压的情况, 这与常识不符, 我们将这些记录剔除, 共 4 个.
- 将一些明显的离群点 (坏点) 设为缺失值.
- 将一些取值范围比较大的指标通过 log(x+0.1) 进行变换.

经过上述清洗之后, 得到的数据集中包括 2507 条记录和 68 个指标（包含响应变量）.

缺失值填补

设数据集 D 由观测集 D_{obs} 和缺失值 D_{mis} 组成. 设缺失矩阵为 M, 其中 $m_{ij} = 1_{\{d_{ij} \in D_{mis}\}}$. 我们假设缺失数据满足 MAR(Missing at random) 条件:

$$P(M|D) = P(M|D_{obs})$$

在此基础上, 我们使用 4 种缺失值填补方法.

1. 完全数据分析 *(Complete Analysis)*
   仅保留所有特征均无缺失值的样本, 术前 + 术后的数据中完整的观测共 404 条 · 术前数据中完整观测共 630 条。
2. 完全随机填补 *(MCAR Imputation)*
   对于所有的缺失变量 · 从其观测到的个体中等概率抽取 · 填补缺失值。
3. *Multiple Imputation*[2]
   设 $D \sim N_k(\mu, \Sigma)$, 记 $\theta = (\mu, \Sigma)$. 由 MAR 假设, $P(M|D) = P(M|D_{obs})$, 且似然函数 $P(D_{obs}, M|\theta) = P(M|D_{obs})P(D_{obs}|\theta)$. 由贝叶斯原理,

$$P(\theta|D_{obs}) \propto P(D_{obs}|\theta) = \int P(D|\theta)dD_{mis}.$$

   该后验概率可以由 EM 算法求得 · 得到 μ 和 Σ 的估计后依下式抽样,

$$\tilde{D}_{mis} \sim P(D_{mis}|D_{obs}, \tilde{\mu}, \tilde{\Sigma}).$$

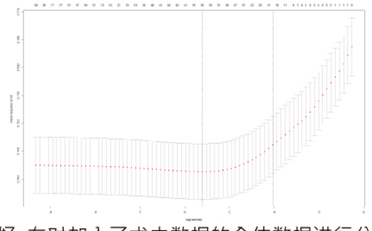
   从而完成数填补. 特别的, 这里对于类别变量进行特殊处理.
4. *Multiple Hot-deck*[3]
   Hot-deck 基础上的改进方法. 对缺失值, 从与当前样本相似度最高的样本中随机选值来填补. 将上述过程进行 m 次, 可得到 m 个完全数据集.

分类预测——结合特征选择

对于每种填补方法得到的数据集, 在使用 LASSO 进行特征选择后, 分别使用逻辑斯蒂回归 (LR), 支持向量机 (SVM), 神经网络 (ANN) 和随机森林 (RF) 进行分类. 具体的, 我们从样本中随机抽取 80% 作为训练集, 剩下的作为测试集, 对每种分类方法, 重复上述过程 100 次, 取预测正确率的平均值. 我们先对住院即刻与手术前的数据进行分析. 预测分类错误率结果汇总如下表所示, ⋆ 表示使用 LASSO 进行特征选择后再分类.

	MCAR Imp	Multiple hot deck	Multiple Imp	Complete Ana
ANN	0.193625	0.194422	0.19626	0.160494
ANN⋆	0.182072	0.18259	0.185179	0.159259
SVM	0.190637	0.19008	0.186255	0.158025
SVM⋆	0.181275	0.18498	0.185656	0.157901
LR	0.196175	0.196092	0.190159	0.18716
LR⋆	0.192052	0.195199	0.191673	0.159877
RF	0.151593625	0.15123506	0.151075697	0.161728395
RF⋆	0.147609562	0.151354582	0.154342629	0.162962963

使用 LASSO 进行特征选择对结果有所提升. 左起第一条竖线表示最优特征个数, 第二条竖线表示解释一半方差所需的特征个数.



在多种分类方法中, 随机森林表现较好, 在对加入了术中数据的全体数据进行分类预测时, 结果有显著提升. 加入术中数据的显著特征有血小板膜糖蛋白和钙拮抗剂等等.

	MCAR Imp	Multiple hot deck	Multiple Imp	Complete Ana
RF	0.124701195	0.124023904	0.120996016	0.096296296
RF⋆	0.119322709	0.119760956	0.124422311	0.096296296

进一步我们将随机森林方法应用住院即刻的数据上, 结果如下

	MCAR Imp	Multiple hot deck	Multiple Imp	Complete Ana
RF	0.151167315	0.147587549	0.154396887	0.208730159
RF⋆	0.143190661	0.150505837	0.147120623	0.201587302

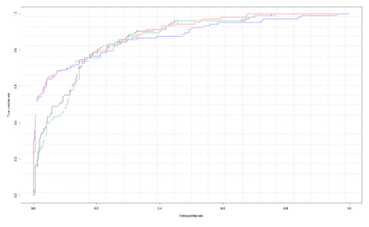
有趣的是, 对比住院即刻的数据与手术前的所有数据, 经过特征选择后, 后者并没有提供任何更多的显著特征, 所以二者结果相近. 最显著的 4 个特征为年龄, 中性粒细胞, 甘油三酯, 随机血糖. 其中年龄与无复流呈负相关, 说明 PCI 手术对高龄患者效果优于低龄患者, 这与临床经验相符合 [4].

侧重某一类预测的正确率

无复流现象为 PCI 手术导致的并发症, 从实际角度考虑, 相对于没有无复流现象, 对有无复流现象发生的病患我们希望能尽可能准确的预测 (例如若预测为会发生无复流现象, 则不进行手术), 甚至能全部准确预测, 并考虑在此基础上, 对有复流情形的预测准确率能达到多少 (或者反向考虑).

这里我们用受试者工作特征曲线 (ROC,receiver operating characteristic curve) 来刻画, 曲线上各点反映着相同的感受性, 是对同一信号刺激的反应. ROC 曲线越凸越近左上角表明其诊断价值越大, 曲线下面积可评价诊断准确性.

右图中横轴为发生无复流情形错判率, 纵轴为不发生情形的正确率, 较上方的两条曲线来自 RF 和 RF⋆, 下方的两条来自 LR 和 LR⋆. 可以发现,RF(随机森林) 的效果较优. 具体的, 当我们无复流情形错判率控制在 0.05 以下时, 不发生无复流情形的正确率能达到 60% 以上.



团队分工

陈翰轩: 特征选择, 分类预测
邓思圆: 预处理, 数据清洗与填补
陆道旭: 结果分析, 报告编辑制作
徐来: 特征选择, 分类预测

References

[1] Friedman, T Hastie, R Tibshirani. Elements of Statistical Learning. 2001.

[2] Dempster, Arthur P., N.M. Laird and D.B. Rubin. 1977. "Maximum likelihood estimation from incomplete data via the em algorithm." Journal of the Royal Statistical Society B 39:1–38

[3] Cranmer, S.J. and Gill, J.M.. (2013) "We Have to Be Discrete About This: A Non-Parametric Imputation Technique for Missing Categorical Data." British Journal of Political Science 43:2 (425-449).

[4] 谭保平. 年龄对冠心病患者 PCI 术后长期结果的影响 [J]. 中国医药导刊, 2011, 13(8):1304-1305.