心脏手术预测的相关问题

吴昊 王计玥 王添 陈建宏
上海交通大学
2013/7/28

摘要

本文主要讨论了如何用有缺失的术前术中指标来估计手术后是否有回流。最终的目的是挑选出尽量少的术前指标来对手术结果有一个精确的预测。鉴于对于完整的数据，我们有 Logistic 回归模型，Lasso 方法，SVM（支持向量学习机）等多种方法来建立预测模型，故这个问题的主要难点在于缺失数据。另一方面，对于数据缺失，传统的 KNN（k-nearest neighbor）的方法对于高维向量将不适用，故我们有如下想法:

1. 算出所有术前指标和是否回流指标的 precision matrix(精确度矩阵)，此时就可以根据术前指标与是否回流指标的关联强弱来选择变量。
2. Precision matrix(精确度矩阵) 同时也揭示了术前指标之间的相关性，故可以在对某个指标进行数据补全的时候，根据与其相关性强的指标进行 KNN。

数据预处理

1. 数据介绍

共有样本数 n=1214 术前指标变量共有 p=41，样本矩阵为 $X=(x_{ij})_{n\times p}$，反应变量 $Y=(y_i)_{n\times 1}$代表手术后是否回流。

2. 明显有错误的数据

此次总数据量达十万级别，其中难免有因为输入错误而产生的异常数据。

我们采取的方法是将每个指标的数据画出散点图，找出其中远超出正常范围的特殊点。

比如 249 号病人的舒张压只有 5。再看收缩压，252 号病人 15，394 号病人 11，634 号病人 10，甚至都低于各自的舒张压。而血红蛋白指标中有 11 个病人为 0（具体为 71,231,475,581,684,744,809,942,945,1082,1164）。

另外，在“高敏 C 反应蛋白”指标中，安贞医院的数据在 0~45 直接近似呈指数分布，有五分之一的数据落在 15~45 之间。但是另外两家医院的病人在此指标上均在 0~15 之

间。这个现象不知是什么原因造成，尚待探究。

3. 对于属性变量，由于其指标数据之间并不存在强度关系，故采用哑指标的方法。对于非属性变量。考虑到每个变量的单位，变化范围均不同，为了合理的表现出其与是否回流指标的关联强弱，一个合理的方法为做 z-normalization:

$$\tilde{x}_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}$$

其中 μ_j 为术前指标 j 的平均值，σ_j 为其方差。

对于有数据缺失的情况，及有观测到的值为样本即可。

理论

1. 精确度矩阵（precision matrix）

精确度矩阵即为协方差矩阵的逆。从数学上来说，给定一些随机变量 $(X_1, \cdots X_d)$，若他们的联合分布是联合正态分布的话，则有：

$$P(X_1 = x_1, \cdots X_d = x_d) = \frac{1}{(2\pi)^{d/2} \left|\det(\Sigma)\right|^{1/2}} \exp(-\frac{1}{2}(\vec{x} - \vec{\mu})^T \Sigma^{-1}(\vec{x} - \vec{\mu}))$$

其中，$\vec{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix}$，$\vec{\mu}$ 为 $\vec{x}$ 的平均值，Σ 为协方差矩阵，$P = \Sigma^{-1}$ 为精确度矩阵。更为重要的是，精确度矩阵的每个元素都有重要意义即：

$$P(i, j) = Cov(X_i, X_j \mid \cdots X_{i-1}, X_{i+1}, \cdots X_{j-1}, X_{j+1}, \cdots)$$

即精确度的第 (i, j) 分量代表了 X_i, X_j 两个随机变量的条件协方差

从这个角度来说，精确度矩阵可以表现出任意两个随机变量之间的直接联系，而过滤了因为由第三个随机变量作为桥梁的联系。这样，若我们可以做出术前术中指标和最后是否回流指标的精确度矩阵的话，那么我就可以直观的看出那些指标与术后回流指标有直接的强关联关系。这些信息对我们选取变量有指导作用，我们可以去精确度矩阵回流指标对应列中选择绝对值高变量再进行回归。这样可以排除很多噪音。

选择用精确度矩阵的另外一个重要原因为在有数据缺失的情况下，不能直接使用各种回归方法。而用 KNN 方法，会代入误差。而只选择各种指标的全的样本进行回归，

会大大减少数据的利用率。而，精确度矩阵的方法，可以在尽量增加数据利用率的情况下，不引入额外的误差。在有数据缺失的情况下，precision matrix 的估计理论可以参考[1]

定义矩阵 $K \in R^{n\times p}$，其中 $K(i,j)=1$ 代表第 i 个样本的第 j 个指标观测值未缺失。相反的，$K(i,j)=0$ 代表第 i 个样本的第 j 个指标观测值未缺失。

则，定义经验协方差矩阵 $\hat{\Sigma}$：

$$\hat{\Sigma}(i,j)=\frac{\sum_{t=1}^{n} k_{ti}k_{tj}(x_{ti}-\hat{\mu}_i)(x_{tj}-\hat{\mu}_j)}{\hat{\sigma}_i\hat{\sigma}_j\sum k_{ti}k_{tj}}$$

其中 $\hat{\mu}_i$ 是第 i 个指标的平均值，定义为

$$\hat{\mu}_i=\frac{\sum_{t=1}^{n} k_{ti}x_{ti}}{\sum k_{ti}}$$

其中 $\hat{\sigma}_i$ 为第 i 个指标的标准差，定义为

$$\hat{\sigma}_i^{\,2}=\frac{\sum_{t=1}^{n} k_{ti}(x_{ti}-\hat{\mu}_i)^2}{\sum k_{ti}}$$

这样我们就可以得到经验协方差矩阵。取逆，则可以得到估计的协方差逆矩阵。当然，若经验协方差矩阵不可逆，或者结果不好，这里也可以考虑稀疏性质。

2. KNN 方法

对于高维数据，传统的 KNN 算法失效，故我们希望可以通过 precision matrix，揭示的指标之间的关系，将高维 KNN 问题，转化成低维的填补问题。即根据指标之间的关系强弱，来定义距离中的权重。

原来的 kNN 算法中，譬如我们要填补第 i 位病人的第 j 个指标，那么原先的 kNN 算法是按照欧氏距离选出 k 个邻居；尽管我们必须对原先的数据矩阵归一化，但是在计算欧氏距离时我们还是认为每种指标对于两个病人之间的距离是一样的，但是我

们可以改进距离的算法：通过 Precision matrix，我们可以得到其他指标和指标 j 的关系，Precision matrix 归一化后得到 Partial correlation matrix P，可以近似地认为 P(m,j)数值代表了指标 m 和指标 j 的 correlation；故在填补第 i 位病人的第 j 个指标时，我们采用下式计算第 k 个病人和第 i 个病人的距离：

$$\rho_j(\mathrm{k,i}) = \mathrm{sqrt}(c_j \sum_{m\neq j} \frac{d_m(k,i)}{abs(P(m,j))})$$

这里 m 的范围是除了 j 以外的指标，c_j 是归一化系数，其中 $d_m(k,i)$ 取法如下:

如果第 k 个病人和第 i 个病人的 m 指标都缺失，则[3]：

$$d_m(k,i) = E\big(X(k,m) - X(i,m)\big)^2$$

如果第 k 个病人和第 i 个病人的 m 指标都不缺失，则：

$$d_m(k,i) = (X(k,m) - X(i,m))^2$$

如果第 k 个病人和第 i 个病人的 m 指标有一个缺失，譬如 data(i,m)=NA，则：

$$d_m(k,i) = E(X(k,m) - X(i,m) \mid X(k,m))^2$$

3. Logistic 回归

结果

1. 变量选择

在 precision matrix 中选出最后一列绝对值最大的二十个指标及数据如下。

PCI 术中钙拮抗剂	0.874055
PCI 术中 2b3a	0.475942
罪犯血管血栓数量	0.344792
IABP	0.233557
PCI 前 CKMB	0.219559
随机血糖	0.185473
PCI 前 CK	0.166979
侧枝循环分级	0.164264
PCI 术中血栓抽吸	0.154491
killip 分级	0.128908
LDLC	0.119273
舒张压	0.107099
TNI	0.105653
脑梗塞史	0.102515
PCI 前硝酸酯	0.097739
PCI 前溶栓	0.083994
高血压史	0.083887
症状到 PCI 时间	0.082322
球囊扩张次数	0.072674

预扩张	0.067945

我们选取前十名，作为我们 logistic 回归的变量。

与赵靖康、苗旺[2]的报告对比：

其关联性最强的四个指标是“PCI 术中钙拮抗剂”、“PCI 术中 2b3a”、“罪犯血管血栓数量”、“IABP”与我们相同，除了“罪犯血管血栓数量”和“IABP”两指标顺序交换。但其第五个指标“侧枝循环分级”在我们的结果中排名第八。

与陆宇的报告对比：

其关联性最强的两个指标“PCI 前 ARB”、“HDLC”在我们的结果中均未出现，但之后的五个指标均在我们结果的前六名中出现。

以上比较可以认为结果基本可靠。

2. Logistic 回归和预测

变量选择之后，取样本总量的 4/5 进行 logistic 回归，系数如下，下列数据是由 lasso 系数 λ=0.01时得到：

指标名称	回归系数
(Intercept)	-2.17707472
PCI 术中钙拮抗剂	1.06008520
PCI 术中 2b3a	0.70146928
罪犯血管血栓数量	0.72079666
IABP	0.27612353
PCI 前 CKMB	0.08447686
随机血糖	0.18422740
PCI 前 CK	0
侧枝循环分级	-0.27928805
PCI 术中血栓抽吸	0.36116250
killip 分级	0.15534445

剩下的 1/5 的样本进行检验，得到不同 threshold 下的误判率（每个数据均是 100 次实验的统计结果）。

threshold	mean	Standard deviation
0.31	0.091535	0.017059
0.32	0.094803	0.016099
0.33	0.093976	0.01603
0.34	0.092913	0.016703
0.35	0.092795	0.016536
0.36	0.092992	0.016255
0.37	0.090709	0.016815
0.38	0.091693	0.014401
0.39	0.09122	0.01789
0.4	0.09122	0.015057

0.41	0.089724	0.016939
0.42	0.093071	0.017478
0.43	0.093189	0.017239
0.44	0.092598	0.017369
0.45	0.091654	0.016334
0.46	0.091339	0.016286
0.47	0.092087	0.016227
0.48	0.089764	0.016843
0.49	0.093504	0.016902
0.5	0.09315	0.018821

看到，方差在一个合理的范围内。

结果分析

可以看到，我们的结果与已知最好的结果相差不远，而且注意到，我们只采用了线性模型，而在以前的工作中，离散的强度量成为哑指标以后，会出现回归系数并不同号的结果,即哑指标的引入相当于引入了回流指标和术前术中指标的非线性关系。这方面我们的模型还可以改进。

未来方向

我们的方法在术前预测中的表现并不好，误判率大约 20%，我们现在的想法是通过术前指标来预测重要的术中指标。总而言之，还有很多等着我们去做。

引用

[1] *Estimating Sparse Precision Matrices from Data with Missing Values* **Mladen Kolar Eric P. Xing**

[2]《自变量有缺失的分类问题》 **李艳芳 苗旺**

[3] Final Project Report **Yu Lu, Zhongyu Wang, Yuting Wei**