基于患者检查数据的出院状态预测模型

233345 姚雪健，233352郑晨曦，233464金智豪

东南大学 交通学院，江苏 南京 210096

摘要：

1. 研究简介
2. 数据统计性分析

## 2.1 基线数据统计性分析

选取性别(gender)、年龄(admission\_age)、体重(weight\_admit)、身高(height)和种族(ethnicity)这五个基线数据分析患者出院状态(hospital\_expire\_flag)的分布。

首先对不同基线数据分布进行数据的分组处理。考虑到性别、种族为离散变量，而年龄、体重和身高为连续变量，同时医疗数据可能存在记录错误等意外情况（如数据中出现了80岁体重1kg的明显错误），在分组前先通过3sigma法则筛去异常值，后对于初步清洗的数据，若为离散变量则选取类别为分类标准，若为连续变量则选取合适的区间，将其转变为离散数据，最后按照分组计算死亡比例，不同基线数据下患者出院状态分布结果如图1所示。

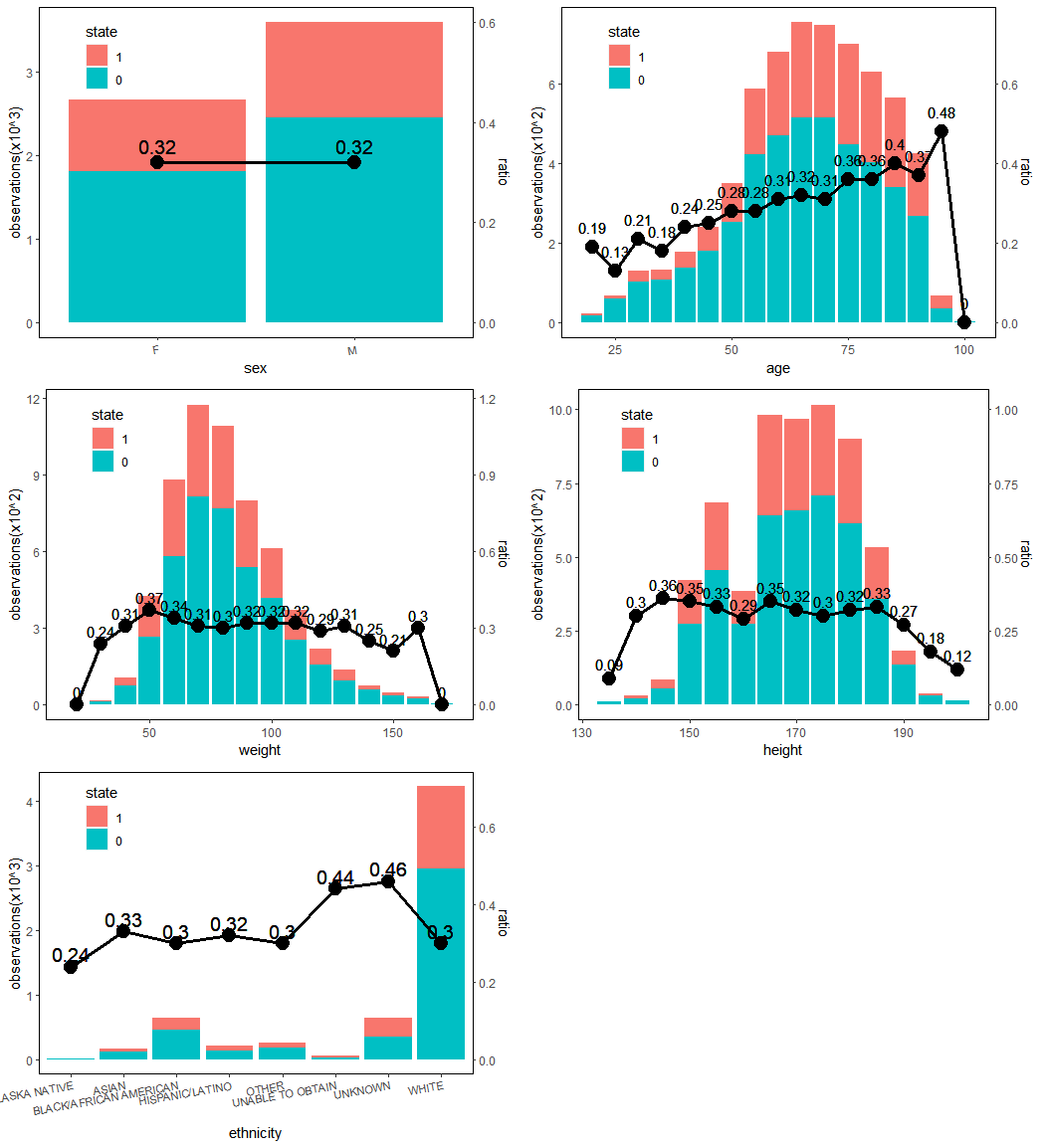
对于性别，患者中男性数量偏多，但是否死亡与性别关系较小，男女死亡比例均为0.32。

对于年龄，患者年龄从20岁覆盖到100岁，近似呈现出正态分布状，其中50-90岁人数较多，年龄均值为66.36岁。死亡率随年龄增长不断上升。

对于体重，患者体重从30kg覆盖到160kg，近似呈现出正态分布状，其中50-100kg人数较多，体重均值为83.02kg。死亡率在体重区间的分布较为均匀，基本维持在0.3左右。

对于身高患者身高从135cm覆盖到200cm，近似呈现出正态分布状，其中160-180kg人数较多，身高均值为168.8cm。死亡率在身高区间的分布较为均匀，基本维持在0.3左右。

对于种族，患者中白人数量最多，但是否死亡与种族关系较小，不同种族除未知种族外，死亡比例均维持在0.3左右。



**图1** 不同基线数据下患者出院状态分布图

## 2.2 不同基线数据下出院状态的差异性分析

为了更深入了解患者的基线数据与出院状态之间的关系，本文对性别、年龄、身高、体重和种族这些基线数据进行了差异性分析。

对于连续的基线数据，如身高、体重和年龄，选用ANOVA（方差分析）作为分析方法。ANOVA允许比较独立样本均值，从而帮助确定出院状态是否对这些连续变量均值产生影响。

而对于分类的基线数据，如性别和种族，则使用卡方检验。这是因为卡方检验适用于评估两个分类变量之间的关联性或独立性，允许检查这些分类变量与出院状态之间是否存在关系。分析结果如表1所示。

由表1可以得到以下结论：

1. 年龄、身高、体重和种族与患者的出院状态存在显著的关系。
2. 性别与患者的出院状态之间没有显著的关系。

**表1** 差异性分析结果

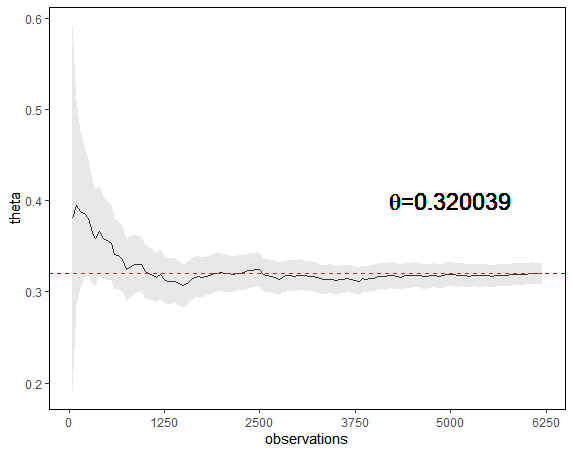
|  |  |  |  |
| --- | --- | --- | --- |
| 基线数据 | 分析方法 | F 值/χ² 值 | p值 |
| 性别(gender) | Chi-squared | 0.015798 | 0.9 |
| 年龄(admission\_age) | ANOVA | 85.34 | <2e-16 |
| 身高(height) | ANOVA | 4.495 | 0.034 |
| 体重(weight\_admit) | ANOVA | 6.16 | 0.0131 |
| 种族(ethnicity) | Chi-squared | 74.164 | 2.12e-13 |

## 2.3 患者出院状态死亡概率贝叶斯估计

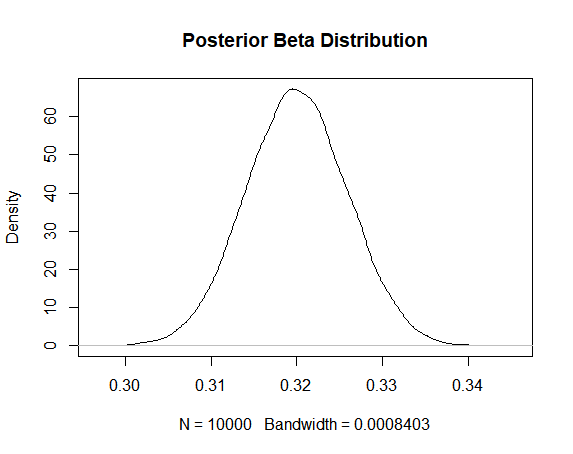
每个患者出院状态只有死亡和存活两种情况，且死亡概率，可视为伯努利分布；数据集中所有患者的出院状态则可视为二项分布。考虑到beta分布为伯努利分布和二项分布的共轭先验分布，因此选择Beta分布作为死亡概率的先验分布

具体算法流程如伪代码1所示，首先选择前100个样本的出院状态，选择合适的先验Beta分布的初始值，后不断加入后续样本序列中的出院状态作为新的观测，并更新后验Beta分布的参数值，同时基于当前的后验Beta分布计算随机变量在每次更新后的可信区间上下界。绘制得到的死亡状态的点估计曲线和95%可信区间如图2所示，最后一次更新得到的后验Beta(1982,4211)分布如图3所示，死亡概率的后验点估计值为。

|  |
| --- |
| 算法 1 贝叶斯估计 |
| 输入：患者出院状态数据框  输出：患者出院死亡概率的贝叶斯估计曲线   1. 计算出院状态数据框前个数据的均值和标准差 2. # 参数和满足 3. 选取合适的的先验分布的参数和 4. # 后验均值向量，贝叶斯可信区间上界向量，贝叶斯可信区间下界向量 5. # flag只为0或1,其中1表示死亡，0表示存活 6. # 后验分布迭代,为可信性水平 7. **for** flag **in** hospital\_expire\_flag[:] **do**      14. **end for** 15. # 序列采样，采样间隔为step 16. picplot # 绘制点估计曲线及可信区间 |
| 1. **return** pic |



**图2** 患者死亡概率的点估计曲线与可信区间



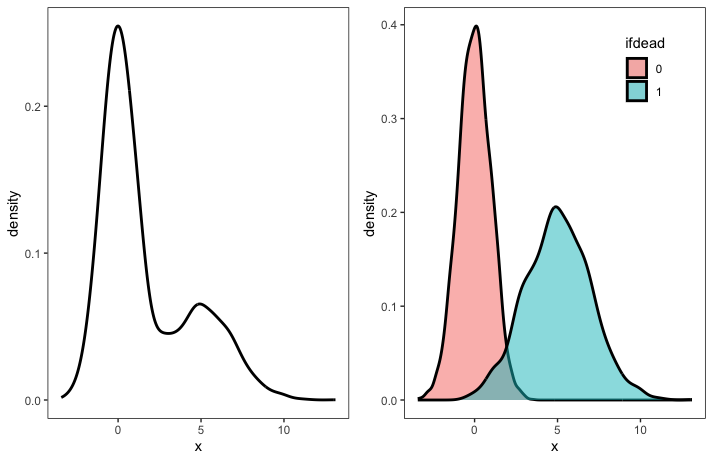
**图3** 后验Beta(1982,4211)分布

## 2.4 患者出院状态的软分类分析

患者出院状态只存在死亡和存活两种状态，可视为一个二分类问题。本节使用概率模型中的高斯混合模型对其进行软分类分析。

2.4.1 高斯混合分布数据生成

对于死亡(hospital\_expire\_flag=1)和存活(hospital\_expire\_flag=0)患者，分别生成和的高斯分布数据，后将两者混合得到高斯混合分布数据，与其对应的出院状态集合则为。完整数据概率密度函数图见图4左，给定出院状态标签下数据概率密度函数图见图4右。



**图4** 生成的高斯混合分布数据概率密度

（左）完整生成数据 （右）不同出院状态的生成数据

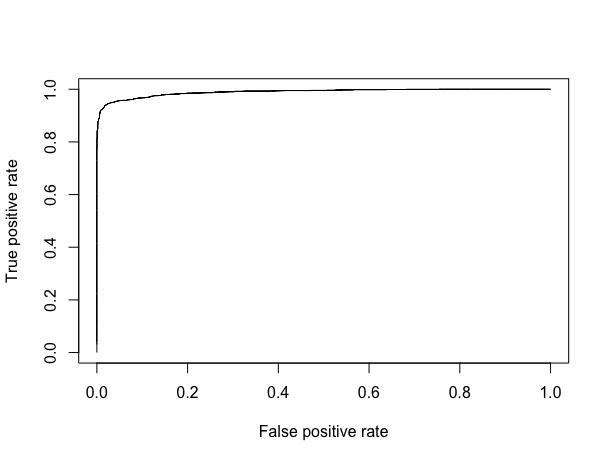
2.4.2数据软分类

软分类使用的是概率模型，输出不同类对应的概率，最后的分类结果取概率最大的类。使用高斯混合分布模型的分类就是一种典型的软分类。首先计算每个观测点数据分别属于存活和死亡两个标签对应高斯分布的概率和，后计算其属于死亡标签的高斯分布的相对概率。若该对应的出院状态，且其，则认为软分类结果正确，否则错误。分类结果如表2所示，P表示预测死亡，N表示预测存活。

**表2** 高斯混合模型软分类结果混淆矩阵

|  |  |  |
| --- | --- | --- |
| 预测  出院状态 | N | P |
| 0 | 4166 | 97 |
| 1 | 112 | 1898 |

从表2中可知，，第一类错误为真实死亡却被预测为存活的概率，即；第二类错误为真实存活却被预测为死亡的概率，即；准确率，召回率，F1分数。绘制ROC曲线如图5所示，ROC曲线下面积。



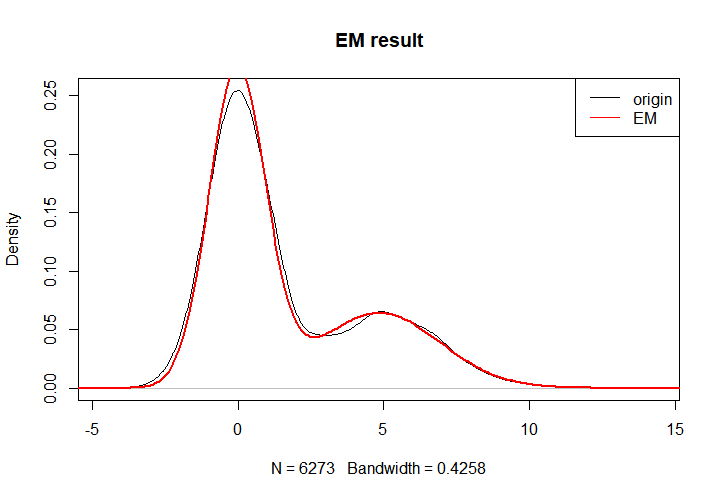
**图5** 高斯混合分布模型二分类ROC曲线

2.4.3 EM算法

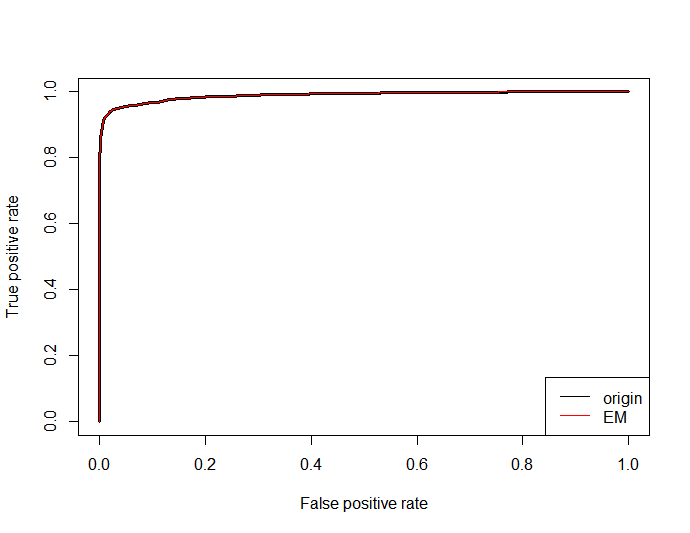
在实际情况中，往往只知道观测数据可能服从高斯混合分布，但不知道其对应标签。这种情况下为了进行软分类，可先赋予高斯混合模型参数值以初始值，采用如EM算法等的迭代算法，借助似然估计进行迭代逼近参数真值，从而在一定误差允许范围内建立估计高斯混合模型。下述伪代码为EM算法迭代求解高斯混合模型参数流程。

|  |
| --- |
| 算法 2 EM算法 |
| 输入：可以使用高斯混合分布拟合的数据  输出：估计的高斯混合分布参数   1. 初始化高斯混合分布参数：第个高斯模型均值，标准差，占整个GMM模型权重 2. 选取合适的迭代停止阈值error\_stop 3. **while** error > error\_stop **do** 4. # E步，估计隐变量分布 5. **for** **in** **do** 6. # 求第i个数据点属于第k个高斯分布的概率 7. **end for** 8. **#** M步，更新参数估计值 9. # 计算第k个高斯分布的数据点个数 10. # 更新第k个高斯分布占整个GMM模型权重 11. # 更新第k个高斯分布均值 12. # 更新第k个高斯分布方差 13. 计算更新前后参数值间误差error 14. **end while** |
| 1. **return** |

对前述生成的服从高斯混合分布的数据，模拟实际情况认为其标签未知，借助EM算法求解得到的高斯混合分布模型为，与真实值十分接近，如图6所示。估计模型与真实模型的AUC曲线对比图如图7所示。



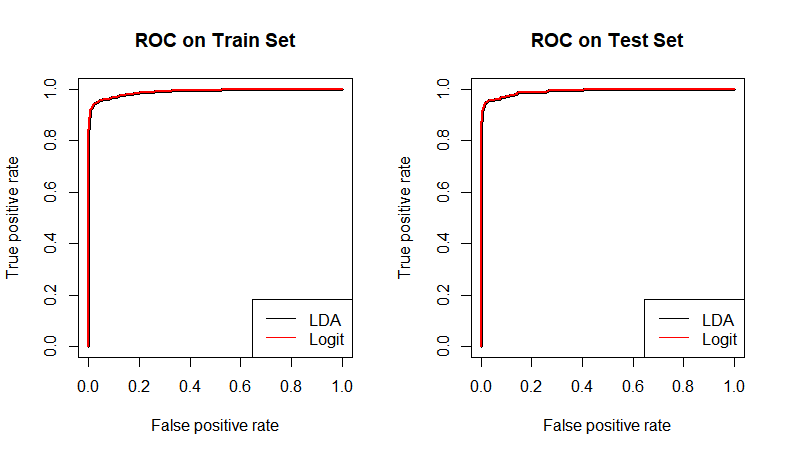
**图6** EM算法迭代结果

****

**图7** EM算法估计模型与原始模型二分类ROC曲线比较

2.4.4 模型比较

将生成的高斯混合分布数据按照8:2比例分成训练集和测试集，在训练集上拟合线性判别（Fisher判别）和Logistic模型，并在测试集上预测，最后分别绘制出两模型在训练集和测试集上的ROC曲线，如图7所示。比较结果发现，三种模型效果类似。



**图7** LDA和Logit模型在训练集和测试集上的ROC曲线

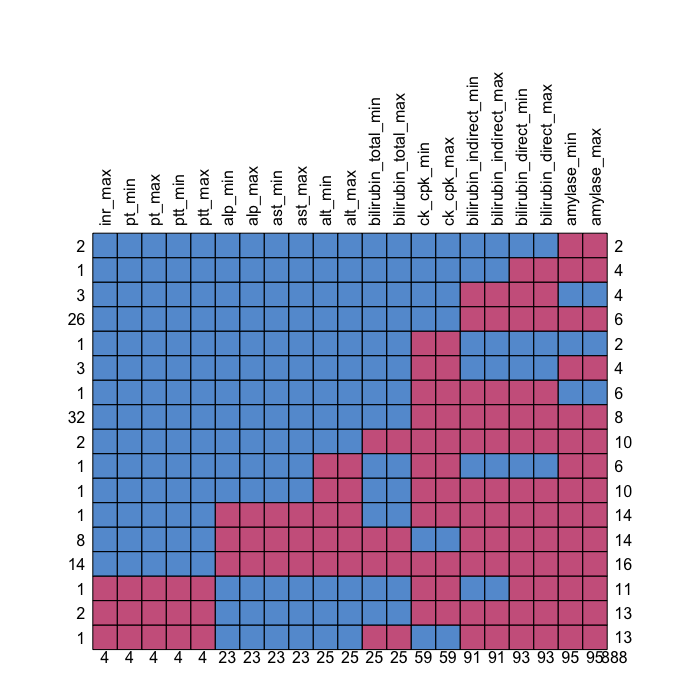
（左）模型在训练集上ROC曲线 （右）模型在测试集上ROC曲线

1. Logit分类模型

## 3.1 数据相关性分析

第二章中利用人工生成的随机数据进行高斯混合分布模型的软分类，该数据不存在出现错误值或缺失值等情况。但在实际医疗数据中，由于患者未做部分检查或隐私需要，存在部分医疗数据缺失的现象。因此，在基于实际医疗数据建立分类模型之前，首先需要对其进行数据缺失情况的分析。

缺失数据存在完全随机缺失(MCAR)、随机缺失(MAR)和非随机缺失(NMAR)三种情况。利用mice包对原始医疗数据进行缺失模式分析后发现，几乎所有数据都存在缺失，大多数缺失数据都与其他变量相关（可能都为同一张报告单数据），根据缺失数据的分类，可认为属于随机缺失，图8为部分原始医疗数据缺失模式。因此针对缺失超过40%的变量，认为其是无效特征并将其删除，最后108个首日检验数据对应变量还剩74个有效变量。



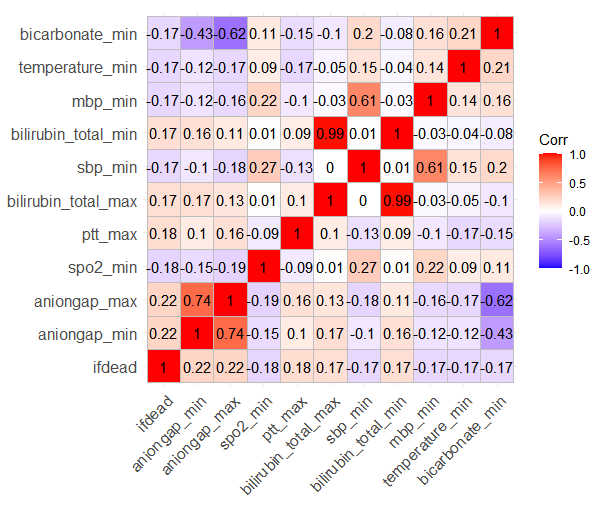
**图8** 部分数据缺失模式

考虑到原始数据存在缺失值，为保证相关性分析的准确性，选择在不进行数据填补的情况下，将单一变量去除NA值后分别与患者出院状态进行相关性分析。由于患者出院为0-1的二分类变量，其余74个有效的待分析变量为连续变量，因此选择Point-biserial相关系数进行相关系数的计算。将计算完的相关系数按绝对值排序后选出相关性最高的前10个变量，结果如表3所示。

**表3** 与出院状态相关性最高的前10个变量

|  |  |  |  |
| --- | --- | --- | --- |
| 变量名 | Point-biserial  相关系数 | 变量名 | Point-biserial  相关系数 |
| aniongap\_min | 0.225 | aniongap\_max | 0.220 |
| spo2\_min | -0.180 | ptt\_max | 0.177 |
| bilirubin\_total\_max | 0.175 | bilirubin\_total\_min | 0.173 |
| sbp\_min | -0.174 | mbp\_min | -0.170 |
| temperature\_min | -0.167 | bicarbonate\_min | -0.165 |

表3中与出院状态相关性最高的前10个变量中，有部分变量分别出现在相同的报告单上，通常一张报告单上的数据都有一定的相关性，更何况有几组变量两两均为同一指标的最高和最低值，因此需继续分析这10个变量间的相关性。相关性热力图如图9所示，可以发现部分变量间呈现出很高的相关性（如aniongap\_min和aniongap\_max，bilirubin\_total\_max和bilirubin\_total\_min），如果直接使用这10个变量进行后续建模，将会出现多重共线性的干扰导致所建模型不可信。



**图9** 相关性热力图

## 3.2 数据共线性影响分析

借助上一节筛选出的与出院状态相关性最高的10个变量，本节研究共线性对模型拟合的影响及其解决方案。首先选用基础的Logit模型，用筛选出的10个变量对出院状态拟合，然后再分别用10个变量中的每一个与出院状态拟合，比较各变量系数和显著性的区别，具体数据如表4所示。

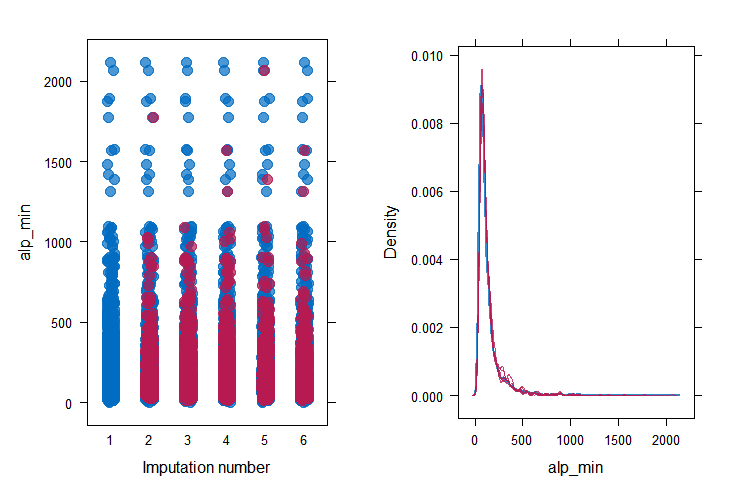
**表4** 单变量与共线性多变量的Logit模型拟合结果

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 变量名 | 多变量拟合时的系数 | 多变量拟合时的显著性Pr(>|z|) | 单变量拟合时的系数 | 单变量拟合时的显著性Pr(>|z|) |
| aniongap\_min | 0.0746122 | 2.37e-09 | -0.03975 | <2e-16 |
| aniongap\_max | 0.0208252 | 0.040084 | -0.029840 | <2e-16 |
| spo2\_min | -0.0169055 | 3.82e-05 | -0.0088083 | <2e-16 |
| ptt\_max | 0.0095495 | < 2e-16 | -0.0071545 | <2e-16 |
| bilirubin\_total\_max | 0.0105592 | 0.819630 | -0.0008436 | 0.88 |
| bilirubin\_total\_min | 0.0552260 | 0.284063 | 0.001311 | 0.835 |
| sbp\_min | -0.0117432 | 0.000145 | -0.0097550 | <2e-16 |
| mbp\_min | -0.0109857 | 0.000624 | -0.0152929 | <2e-16 |
| temperature\_min | 0.0054415 | 0.679928 | -0.0215956 | <2e-16 |
| bicarbonate\_min | 0.0005479 | 0.950951 | -0.041468 | <2e-16 |

## 3.3 数据缺失值处理

与出院状态相关性最高的10个变量之前存在相关性，可以认为完整的检查报告数据中也存在普遍的多重共线性情况。在进行多重共线性的处理之前，首先需要对缺失的医疗数据进行填补，以保证不同变量具有相同的样本长度，便于后续的降维处理。

使用mice包中的预测均值匹配（Predictive Mean Matching, PMM）对筛选出的74个有效变量中的缺失值进行插值。alp\_min变量反映出骨骼系统、肝胆系统出现一系列异常，是一个较为重要的检查指标，同时较为均匀得缺少20%的观测值，是较为有代表性的医疗指标，图10为 alp\_min变量填补结果。图中蓝色代表观测样本，红色代表填补值，可以发现PMM填补结果较好地遵从了观测值的分布，能够在一定程度上满足后续建模的需求。



**图10** alp\_min变量填补结果（左）和概率密度曲线（右）

（蓝色为观测值，红色为填补值）

## 3.4 数据降维

填补完缺失值后，考虑到检查数据变量较多，放入同一个模型中拟合会导致模型复杂度过大且准确性不一定高的问题，需对填补完的数据进行降维操作。常见的数据降维方法有Lasso回归、Ridge回归和主成分分析(PCA)。

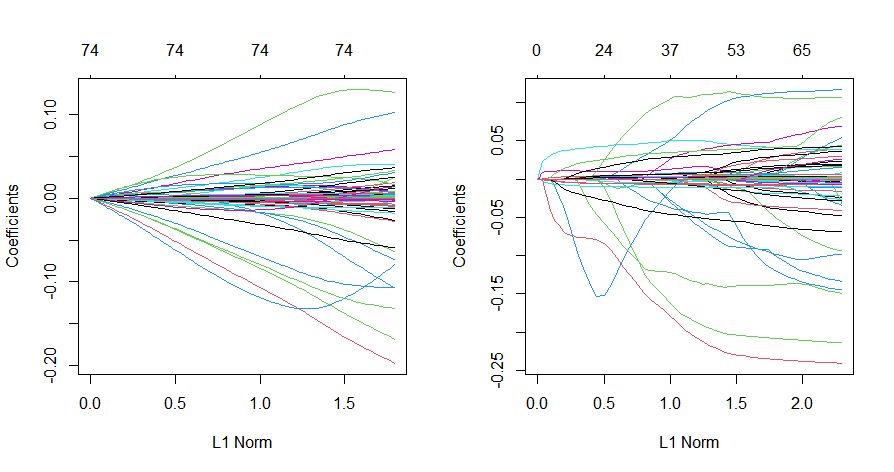
Lasso回归和Ridge回归均使用范数作为正则化项，并具有正则化强度参数。但Lasso回归通过L1范数惩罚模型中的绝对值，对与模型无关的特征系数施加较大的惩罚，从而使此类特征变量系数在求解过程中趋向0，识别和选择对目标变量有最大影响的特征，即实现特征选择；Ridge回归使用L2范数来惩罚模型中系数的平方值，L2正则化不会使系数趋向于0，而是将系数的大小限制在一个范围内，这有助于减小模型中系数的方差，提高模型的鲁棒性，同时保留了所有特征。

主成分分析是一种用于数据降维和特征提取的多变量统计方法,其基本原理是通过线性变换找到一组新的变量，即主成分，将高维数据转换为低维数据，同时最大程度地保留数据中的方差信息。对于主成分分析，首先标准化所有变量之后计算出各主成分，然后根据方差贡献率选取贡献达到80%的主成分作为新的拟合变量。

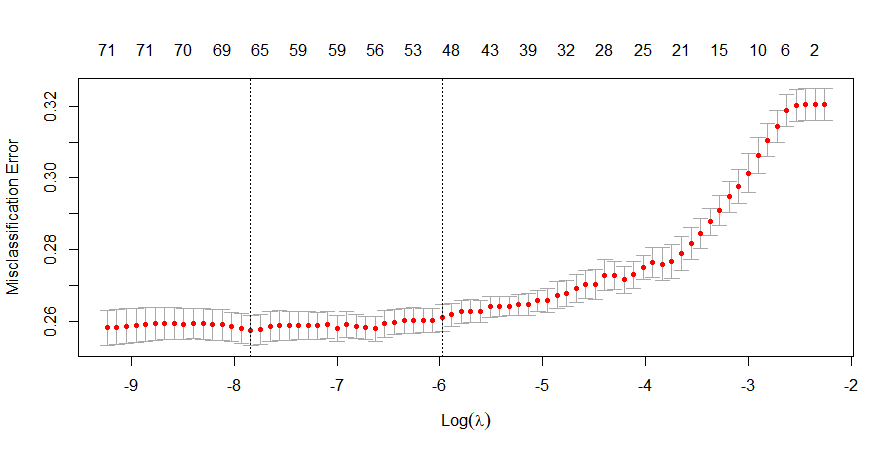
## 3.5 Logistic模型

降维后使用logistic模型进行出院状态的预测。Lasso-logistic回归结果和Ridge-logistic回归结果如图11所示，图中结果包括多个不同超参数取值下的模型。两种回归方法均使用了超参数正则化强度参数，因此通过10折交叉验证选择合适的，据此即可选择最优的预测模型，交叉验证曲线如图12所示。

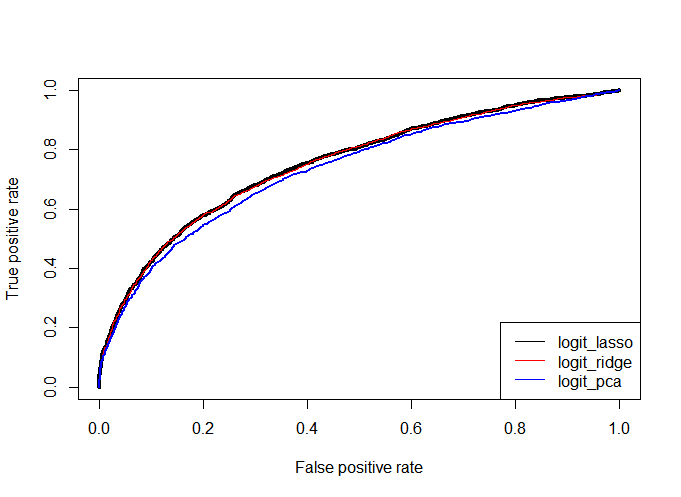
基于三种降维方法对应的的Logistic模型的ROC曲线如图13所示，基于三种降维方法的Logistic模型的ROC曲线十分接近，可认为表现相近。



**图11** Lasso-logistic回归结果（左）和Ridge-logistic回归结果（右）



**图12** 正则化强度参数交叉验证曲线



**图13** Lasso, Ridge, PCA方法下Logistic模型ROC曲线

## 3.6 模型比较

比较模型选取支持向量机(SVM)，随机森林(Random Forest)，AdaBoost和XgBoost四种机器学习方法。考虑到上一问使用了所有有效的74个变量进行降维建模操作，这些数据量对上述模型可能复杂度过大，因此选择与出院状态相关性最高的前10个变量进行降维，后按7:3划分训练集和验证集，并使用默认参数和k-folds交叉验证分别进行建模。

考虑到本次预测模型的背景，能够成功预测更多的出出院死亡病例，并采取相应医疗措施更加重要，因此模型的召回率(Recall)有着更加重要的参考价值，因此在这些模型中，综合参考其预测准确率(Accuracy)和召回率(Recall)。具体结果如表5所示：

**表5** 不同比较模型评价指标

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Evaluation  Index | SVM | | RF | | AdaBoost | | XgBoost | |
| default | best | default | best | default | best | default | best |
| Accuracy | 0.715 | 0.715 | 0.711 | 0.714 | 0.721 | 0.712 | 0.687 | 0.714 |
| Recall | 0.212 | 0.212 | 0.340 | 0.340 | 0.319 | 0.312 | 0.319 | 0.322 |

从上述的各项模型指标中可以发现，各个模型的预测准确率均达到了0.65以上，能够较好地根据已有特征变量来预测患者的出院状态。在这些模型中准确率最高的是AdaBoost模型，其次是SVM模型和随机森林模型。而在召回率评价指标方面，最优为Random Forest模型，其次是XGBoost模型。

对于使用交叉验证的