建立一种用于动态预测脓毒症患者诱发PICS的机器学习方法

华东理工大学
白栋栋 成昊南 黄海骅 霍松泽

摘要

TODO:

目录

1 项目背景

TODO:

2 材料和方法

2.1 数据来源

TODO:

2.2 选择数据

TODO:

2.3 定义输出

TODO:

2.4 计算输出

TODO:

2.5 数据分析

TODO:

3 模型结果

3.1 基准特征

从eICU数据库中提取出100,308条数据，包含17,729名不同的脓毒症患者。其中，3,866 (3.85%)条数据为正例，96,442 (96.15%)条数据为反例。

经过比较，正例数据拥有更长的ICU入住天数（21.067 vs. 10.852, $p < 0.001$），更少的血浆蛋白（2.109 vs. 2.520, $p < 0.001$），更少的淋巴细胞数目（9.931 vs. 12.473, $p < 0.001$），更高的心率（93.337 vs. 88.458, $p < 0.001$），更高的呼吸频率（21.814 vs. 21.019, $p < 0.001$），更少的血清总蛋白（5.578 vs. 5.928, $p < 0.001$），更低的红细胞比容（27.808 vs. 29.888, $p < 0.001$），更少的肌酸酐（1.489 vs. 1.610, $p < 0.001$），更高的白细胞计数（13.218 vs. 12.189, $p < 0.001$），更多的血小板（260.259 vs. 226.342, $p < 0.001$），更低的平均动脉压（79.727 vs. 82.055, $p < 0.001$）。

3.2 模型比较

排名	模型名称	平均准确率	平均AUC[1]
1	CatBoost	0.996(±0.001)	0.996(±0.001)
2	Light Gradient Boosting	0.995(±0.001)	0.996(±0.001)
3	Extreme Gradient Boosting	0.995(±0.001)	0.994(±0.002)
4	Hist Gradient Boosting	0.994(±0.002)	0.996(±0.002)
5	Ada Boost	0.993(±0.002)	0.995(±0.002)
6	Decision Tree	0.989(±0.002)	0.949(±0.013)
7	Multi-Layer Perceptron	0.982(±0.004)	0.975(±0.008)
8	SVM (RBF Kernel)	0.973(±0.003)	0.957(±0.011)
9	Logistic	0.966(±0.007)	0.956(±0.012)
10	Extra Trees	0.961(±0.006)	0.977(±0.006)
11	Naive Bayes	0.961(±0.006)	0.689(±0.034)
12	Ridge	0.961(±0.007)	0.952(±0.013)
13	Linear Discriminant Analysis	0.961(±0.010)	0.952(±0.013)
14	K-Nearest Neighbours	0.951(±0.006)	0.544(±0.025)

[1] AUC：Area Under Curve，接受者操作特性曲线下与坐标轴围成的面积。

表 1: 14种模型的交叉验证结果比较（按平均准确率排序）

用提取出的数据训练预测模型，各种模型的交叉验证结果如表1所示。Logistic回归表现良好（平均准确率：0.966，平均AUC：0.956），而集成学习方法拥有更高的平均准确率和平均AUC。其中，CatBoost的预测结果最好（平均准确率：0.996，平均AUC：0.996），故选择CatBoost进入下一步。

3.3 完整模型与紧凑模型

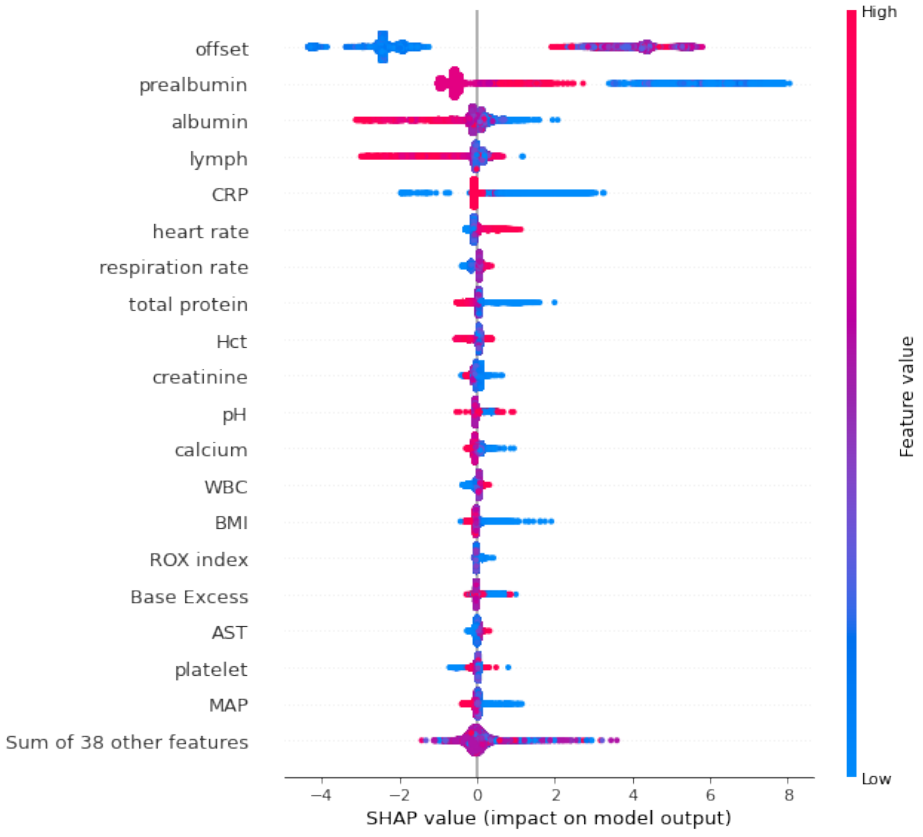

图 1: 完整模型中各变量的平均SHAP值比较

根据预测结果比较，选择含57个输入变量的CatBoost模型为完整模型。计算完整模型中各变量的平均SHAP值，结果如图1所示。此摘要图展示了

各个变量对预测结果的影响情况分布。例如，ICU入住天数（offset）对结果影响明显，且ICU入住天数越长，发生ICU综合症的概率越大。

根据变量的平均SHAP值大小和数据获取的难易程度，选择了15个变量作为输入，建立更加易于使用的紧凑模型。使用默认超参数的紧凑模型平均AUC为90.219%。通过贝叶斯优化超参数后，紧凑模型的平均AUC达到了90.682%，同时平均准确率为96.120%。虽然预测结果的得分略低于完整模型，但是紧凑模型明显在临床上更加可行、更加易用。

3.4 性能分析

TODO: sensitivity analysis

3.5 模型解释

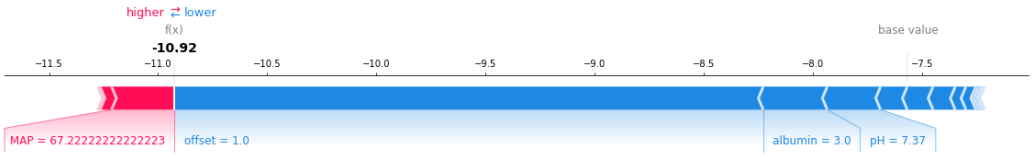

图 2: 个例(A)中主要变量的SHAP值

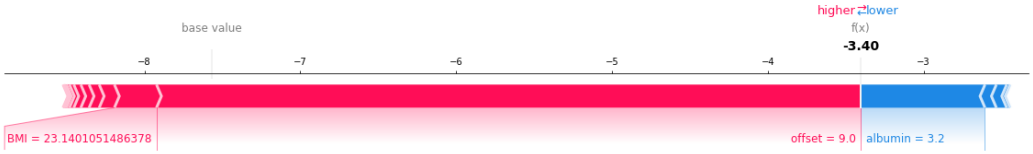

图 3: 个例(B)中主要变量的SHAP值

图1从整体上展示了各个变量对于预测结果的影响情况。而图2和图3展示了两个个例中主要变量的SHAP值。图中红色条和蓝色条分别表示危险因素和安全因素，它们共同作用决定了最终的结果。如图2，在个例(A)中，虽然患者的平均动脉压偏低，但是其ICU入住天数很短、血浆蛋白较多、

pH值也良好，所以模型准确预测了患者次日无ICU综合症风险。又如图3，在个例(B)中，虽然患者的血浆蛋白较多，但是其ICU入住天数较长、身体质量指数（BMI）偏低，所以模型准确预测了患者次日的ICU综合症。

3.6 H5预测工具

为了方便临床上对上述紧凑模型的测试，开发了一款预测脓毒症患者诱发ICU综合症的H5应用。只需在表单中输入指标数值，然后点击“提交”，就可以获得紧凑模型对患者次日发生ICU综合症概率的预测。目前部署应用在此网址上：`http://1.15.185.22/sepsis-pics-tool/`。

4 结论

TODO:

A 附录

TODO: