从临床数据中早期预测脓毒症

**学号： SY2124108 姓名 孙浩哲**

（北京航空航天大学 中法工程师学院 北京 100191）

**摘 要** 脓毒症，又名败血症，是一种严重的血液类疾病，指的是致病菌侵入血液循环，并在血中生长繁殖，产生毒素而引发的急性全身性感染。由于脓毒症的严重性和致死率会随时间上升，脓毒症的早期发现至关重要。本文使用机器学习方法早期预测脓毒症，对输入数据进行建模，训练模型的分类能力，同时验证模型的可解释性。

**关键词** 脓毒症，疾病预测，机器学习

**Early Prediction of Sepsis from Clinical Data**

Haozhe Sun

(Ecole Centrale de Pékin, Beihang University, Beijing, 100191, China.)

**Abstract**: Sepsis is a serious blood-based disease that refers to an acute systemic infection caused by pathogenic bacteria that invade the bloodstream and grow in the blood, producing toxins. Since the severity and lethality of sepsis increase over time, early detection of sepsis is crucial. In this paper, we use a machine learning approach for early sepsis prediction, modeling the input data and training the model to detect sepsis, as well as verifying the interpretability of the model.

**Keywords:** Sepsis, Disease Prediction, Machine Learning

引言

脓毒症是一种会威胁生命的疾病，往往在身体对感染产生反应时发生，会导致组织损伤、器官衰竭甚至死亡。在美国，每年有近 170 万人患脓毒症，27 万人死于脓毒症；在美国医院死亡的人中，超过三分之一患有脓毒症。在国际上，每年约有 3000 万人患脓毒症，600 万人死于脓毒症；约有 420 万新生儿和儿童也会受到影响。脓毒症每年给美国医院造成的损失高达 240 亿美元（占美国医疗保健费用的 13%），其中大部分费用用于入院时未确诊的脓毒症患者。在全球范围内，脓毒症的成本甚至更高，而发展中国家的风险最大。总之，脓毒症凭借其显著的发病率、死亡率和医疗费用成为世界主要的公共卫生问题之一。

脓毒症的早期发现和抗生素治疗对于治疗脓毒症至关重要，每延迟治疗一小时，死亡率就会增加大约 4-8%。为了解决这个问题，临床医生提出了脓毒症的新的诊断方法Sepsis-3 指南，但早期发现和治疗脓毒症的基本需求仍然存在，关于早期检测限制的基本问题仍未得到解答。

本文旨在使用统计学习方法解决早期发现脓毒症问题。我们使用的数据集来自两个独立医院系统的 ICU 患者，每一个样本代表患者一小时的数据。我们将两个医院系统的数据分别设为训练集和测试集，分别含有790215和761995条数据。可用的患者数据变量包括人口统计学数据、生命体征数据和实验室测量数据，详细定义及单位见下表。

|  |  |
| --- | --- |
| **生命体征（第 1-8 栏）** | |
| HR | 心率（每分钟心跳数） |
| O2Sat | 脉搏血氧仪 （％） |
| Temp | 温度（摄氏度） |
| SBP | 收缩压 (mm Hg) |
| MAP | 平均动脉压 (mm Hg) |
| DBP | 舒张压 (mm Hg) |
| Resp | 呼吸频率（每分钟呼吸次数） |
| EtCO2 | 呼吸末二氧化碳 (mm Hg) |
| **实验室值（第 9-34 栏）** | |
| BaseExcess | 过量碳酸氢盐(mmol/L) |
| HCO3 | 碳酸氢盐 (mmol/L) |
| FiO2 | 吸入氧比例 (%) |
| pH | pH值 |
| PaCO2 | 动脉血二氧化碳分压 (mm Hg) |
| SaO2 | 动脉血氧饱和度 (%) |
| AST | 天冬氨酸转氨酶 (IU/L) |
| BUN | 血尿素氮 (mg/dL) |
| Alkalinephos | 碱性磷酸酶（IU/L） |
| Calcium | (毫克/分升) |
| Chloride | (毫摩尔/升) |
| Creatinine | (毫克/分升) |
| Bilirubin\_direct | 直接胆红素 (mg/dL) |
| Glucose | 血清葡萄糖 (mg/dL) |
| Lactate | 乳酸 (mg/dL) |
| Magnesium | (毫摩尔/分升) |
| Phosphate | (毫克/分升) |
| Potassium | (毫摩尔/升) |
| Bilirubin\_total | 总胆红素 (mg/dL) |
| TroponinI | 肌钙蛋白 I (ng/mL) |
| Hct | 分血器 （％） |
| Hgb | 血红蛋白 (g/dL) |
| PTT | 部分凝血活酶时间（秒） |
| WBC | 白细胞计数（计数\*10^3/µL） |
| Fibrinogen | (毫克/分升) |
| Platelets | (计数\*10^3/µL) |
| **人口统计（第 35-40 栏）** | |
| Age | 年（90 岁或以上患者为 100 年） |
| Gender | 女性 (0) 或男性 (1) |
| Unit1 | ICU 单元 (MICU) 的管理标识符 |
| Unit2 | ICU 单元 (SICU) 的管理标识符 |
| HospAdmTime | 入院到 ICU 入院之间的时间 |
| ICULOS | ICU 住院时间（ICU 入院后的小时数） |
| **结果（第41栏）** | |
| SepsisLabel | 对于脓毒症患者，SepsisLabel为1如果时间在脓毒症时间6小时以内，为0如果时间在脓毒症发病时间6小时以外；对于非脓毒症患者，SepsisLabel为 0。 |

使用的机器学习模型

* 1. 对数几率回归（Logistic Regression）

对数几率回归是最基本的机器学习模型之一，广泛应用于二分类任务。对数几率回归模型用逻辑斯谛函数（logistic function或sigmoid function）来建模二分类中正类的概率，这样做等价于对几率（odds）的对数进行线性回归，因此称为对数几率回归模型。损失函数一般使用负的对数似然函数（Negative Log Likelihood，NLL），这等价于用极大似然估计法（Maximum Likelihood Estimation，MLE）进行参数估计。

我们使用内存有限BFGS（limited-memory BFGS，L-BFGS）法进行模型的优化。

* 1. 朴素贝叶斯法（Naïve Bayes）

朴素贝叶斯法是一种基于贝叶斯定理的机器学习模型。我们将监督学习视为后验概率建模问题，由于一般机器学习问题中输入模式很复杂，所以后验概率的直接建模是不可行的，需要改变思路。根据贝叶斯定理，后验概率正比于先验概率和条件概率（有的地方会将条件概率称作似然，但是我们一般还是认为似然函数是对于模型参数来讲的）。朴素贝叶斯采取条件独立假设，将复杂的条件概率拆解成对每一个特征独立的条件概率，从而大大降低了模型的复杂度。之后我们只需对先验概率和分解之后的条件概率进行建模。对于条件概率，我们继续简化模型，认为对每一个特征来说，条件概率都服从正态分布（这称为Gaussian Naïve Bayes）。

由于对先验和条件概率进行建模可以得到联合分布，而从联合分布中可以对数据进行采样，因此朴素贝叶斯属于生成模型。但是注意，虽然朴素贝叶斯模型的分类性能尚可，其对概率分布的建模能力是很差的，原因便是条件独立假设过强。朴素贝叶斯模型也可以视为是一种特殊的贝叶斯网络。

* 1. 支持向量机（Support Vector Machine）

支持向量机是一种使用超平面来进行分割输入空间，从而进行二分类（支持向量机也可以进行多分类，这里不进行阐述）的机器学习模型。在数据集线性可分时，我们希望选择的超平面能够将所有数据正确分类，并且使得两类数据中，与超平面最近的样本（称为支持向量）与超平面之间的距离的和尽可能大（称为硬间隔（hard margin）最大化）。在数据集线性不可分时，我们引入铰链损失函数（hinge loss），模型的优化目标为铰链函数最小化（一般也会引入正则化项，从数学上，正则化项体现了距离尽可能大的理念，称为软间隔（soft margin）最大化）。

一般来说，支持向量机是一种线性模型，但是我们可以使用核技巧将支持向量机扩展为非线性模型（能拟合的非线性函数种类取决于核函数）。采用核技巧后的非线性支持向量机一般使用序列最小化算法（Sequential Minimal Optimization，SMO）进行最优化（工具包如LIBSVM [6]），但其时间复杂度较高，因此不适用于较大的数据集。在数据集较大时，一般仍然使用线性支持向量机，和坐标下降法（Coordinate Descent）进行优化（工具包如LIBLINEAR [7]）。

本文使用线性支持向量机进行分类。同时，由于样本数远大于特征数，我们直接在原始问题上进行优化。

* 1. 随机森林（Random Forest）

随机森林是一种引导聚集算法（Bootstrap Aggregating，简称Bagging），是集成学习（Ensemble Learning）的一种。除了将Bagging的技术应用在决策树（Decision Tree）上之外，随机森林在每次抽出的子集上训练决策树时，都只使用一部分特征，以降低决策树之间的关联性。

* 1. 梯度提升法（Gradient Boosting）

梯度提升法是一类提升（Boosting）方法，是集成学习的另一种主要形式（集成学习最主要的两个分支便是Boosting和Bagging）。提升算法迭代使用弱分类器组成，并将其结果加权求和，得到一个最终的强分类器。加和弱分类器之后，数据通常会被重新加权，来强化对之前分类错误数据点的分类。

比起第一种被发明的提升方法Adaboost，梯度提升法更加灵活，因为Adaboost使用的损失函数是指数损失函数，而梯度提升法则可以使用任何可微的函数作为损失函数。当使用的弱学习器是决策树时，梯度提升法又叫梯度提升树（Gradient Boosting Tree，GBT或Gradient Boosting Decision Tree，GBDT）。

* 1. 多数基线（Majority Baseline）

多数基线并不属于我们一般认为的“机器学习模型”范畴，但却是很多分类任务中常见的基线选择。多数基线模型在训练集中搜寻样本数最多的类别，并在测试时将所有的样本都标为那个类别。很明显，多数基线并没有任何分类能力。因此，如果我们的机器学习模型性能不明显高于多数基线，或者多数基线在我们给定的评价标准下得分异常的高，那么就应当思考是哪里出了问题。常见的问题是模型内部有设计问题从而没有发挥出应有的性能，或者数据集本身类别非常不平衡。

实验

数据预处理

本文中使用的数据集是现实医疗应用中的数据集，其中含有很多噪声和无效信息，因此首先需要对数据集进行预处理。我们使用sklearn包进行缺省值处理和标准化输入变量，用imbalanced-learn（基于sklearn开发的适用于不平衡数据集的机器学习工具包）来处理类别不平衡的问题。

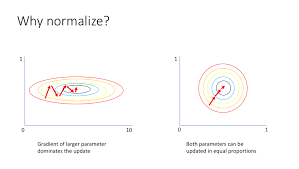
3.1.1 缺省值处理

我们观察到数据集中含有大量的NaN（Not a Number，非数，又叫缺省值）项目，这些无用数据大量占据数据集的空间，给机器学习模型的学习带来了不便。一般常见的做法包括依据同样本的其它变量或者其它样本的同变量进行填充，或者直接丢弃该变量或样本。

我们统计不同输入变量所含NaN的比例，发现8-34列含有的NaN比例超过85%。这无疑给填充法带来了极大的困难，因为缺失样本过多。于是我们选择直接丢弃8-34列的特征，只保留1-7列，35-40列的输入变量和41列的输出变量。

去除这些输入特征后，剩下的特征仍然含有少部分缺省值。我们采用平均值填充的方法，即用各特征的平均值来代替NaN。

3.1.2 标准化输入变量

输入变量的尺度不同是机器学习问题中常见的问题：虽然不会改变最终的最优解，但是尺度不同会给最优化算法带来学习上的困难，见下图[4]：

因此我们统一将各个输入特征标准化到平均值为0，方差为1的区间中，以方便模型进行学习。

3.1.3 类别不平衡问题

随后我们观察到，由于脓毒症的发病率较低，数据集中正类与负类的比例非常不均匀：以测试集为例，测试集中的761995个样本中只有10780个样本是正类，即该ICU患者患有脓毒症。这不仅给模型训练带来了障碍（模型很难从少数类的错误分类中受到足够的惩罚，从而少数类的准确率会较低），也让准确率不能作为模型的合理评价标准，因为准确度容易夸大模型的判别能力（例如在我们的数据集上，多数基线也可以获得约98%的准确率）。为此，我们使用两种另外的评价标准来衡量模型的好坏：

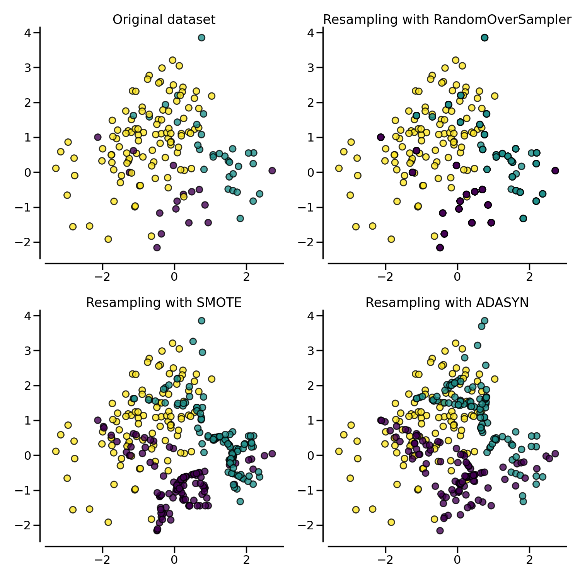
平衡准确率[3]（balanced accuracy）：定义为在不同的类上取得的召回率（recall）的平均值（例如在二分类中，多数基线的平衡准确度为0.5）；

经验风险（empirical risk）：定义一个损失函数，对于假正类取值1，对于假负类取值1000（这样选取的直观理解是对于严重疾病，漏诊比误诊严重的多），对于测试集所有数据求平均值。

而对于模型训练，我们则尝试以下方法对数据集进行处理：

对于负类进行随机的欠采样（random undersampling），采用不放回抽样的方法从负类数据中抽取数量等同于正类的样本；

对于正类进行随机的过采样（random oversampling），使得正类数据和负类数据数量相同；

SMOTE [1]（Synthetic Minority Oversampling Technique）和ADASYN [2]（Adaptive Synthetic sampling）：这两种方法属于过采样的范畴，但比起普通的采样，SMOTE和ADASYN使用插值法来生成新的正类数据。SMOTE和ADASYN过采样的结果可参考下图[5]：

实验结果

我们使用sklearn中实现的模型进行实验。在第一部分实验中，我们首先检验在没有使用欠采样或过采样时各模型的性能（取三位有效数字）：

|  |  |  |  |
| --- | --- | --- | --- |
|  | 准确率 | 平衡准确率 | 经验风险 |
| 多数基线 | **0.986** | 0.500 | 14.1 |
| 对数几率回归 | 0.985 | 0.503 | 14.1 |
| 朴素贝叶斯 | 0.958 | **0.576** | **11.6** |
| 线性支持向量机 | **0.986** | 0.500 | 14.1 |
| 随机森林 | 0.984 | 0.510 | 13.8 |
| 梯度提升 | 0.972 | 0.509 | 13.7 |

如果观察准确率，我们会错误地做出结论：多数基线与线性支持向量机取得了最好的分类效果。然而，正如我们在3.1.3中讨论的，准确率在类别不平衡的情况下并不能作为合理的评价标准。从平衡准确率和经验风险的数据可以看出：

在类别不平衡时表现最好的模型是朴素贝叶斯模型；

准确率夸大了模型的分类能力；

学习后的模型分类能力很差，因为我们观察到除了朴素贝叶斯模型之外的模型性能都只等同或略好于多数基线。

我们给出一个对1)的解释：朴素贝叶斯模型是生成模型，通过对类别的先验建模，模型可以减轻类别不平衡带来的参数估计的困难。

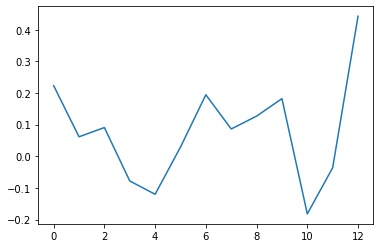
接下来我们考察采用过采样或欠采样（以下可统称重采样，resampling）后的模型的性能：

平衡准确率：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 随机过采样 | 随机欠采样 | SMOTE | ADASYN |
| 对数几率回归 | 0.643 | 0.643 | 0.642 | 0.634 |
| 朴素贝叶斯 | 0.657 | 0.656 | 0.656 | 0.659 |
| 线性支持向量机 | 0.641 | 0.641 | 0.640 | 0.631 |
| 随机森林 | 0.501 | 0.663 | 0.504 | 0.503 |
| 梯度提升 | 0.647 | 0.667 | 0.535 | 0.529 |

经验风险：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 随机过采样 | 随机欠采样 | SMOTE | ADASYN |
| 对数几率回归 | 6.67 | 6.67 | 6.65 | 6.35 |
| 朴素贝叶斯 | 7.97 | 7.96 | 8.17 | 7.73 |
| 线性支持向量机 | 6.84 | 6.83 | 6.82 | 6.48 |
| 随机森林 | 14.1 | 7.60 | 9.84 | 9.39 |
| 梯度提升 | 8.20 | 7.40 | 2.59 | 2.29 |



我们观察发现：

总体来看，大多数模型在使用重采样的方法后性能都得到了明显的提升；

并没有哪一种重采样方法在任何模型上都有优势，也没有一种模型在任何重采样方法上都胜过其它模型。同样，比较两种评价方式，会发现以不同的评价方式为基准，做出的判断也不同；

总的来说，非集成方法对于重采样方法具有健壮性，即使用不同的方法对最终分类能力的影响不大（但使用比不使用要好得多）。而集成方法则对重采样方法很敏感，在陌生的不平衡数据集上应用集成方法应当谨慎；

过采样对随机森林的性能几乎没有影响，而使用欠采样则很有提升；SMOTE和ADASYN作用于梯度提升法明显减少了假负类，但是以假正类的增加为代价的；朴素贝叶斯比起对数几率回归和支持向量机来说假负类要略多一些，这与没有进行重采样时的情况相反；猜测有可能是因为没有进行先验的重新调整，导致训练集与测试集先验不一致所致。

1. 可解释性

可解释性是机器学习的一种非量化的性能评价标准，近年来愈发受研究人员和工业界重视。可解释性可以定义为模型对于输出可以给出令人信服的解释，最经典的方法便是分析输入特征与输出之间的关联度，即找出对做出判断最有影响力的输入变量。在这里我们通过分析对数几率回归模型的参数向量，来对其输出进行解释。由于对数几率回归模型是线性模型，其参数能够很直观地反映输入特征的重要性。

从图表来看，对判断对象是否患有脓毒症来说，最重要的特征是ICU住院时间，其次则是心率和呼吸频率，三者均是越高越有可能是正例（即罹患脓毒症）。对于ICU住院时间，我们认为，人体在感染的条件下暴露的越久，就越容易引发脓毒症。而对于心率和呼吸频率，查阅资料可得脓毒症常见的症状包括心跳过速和呼吸急促，因此我们的模型给出的解释与现实基本是一致的。

**结束语**

脓毒症作为一种威胁生命的严重疾病，我们希望使用机器学习模型来早期预测其发病状态。本文简要介绍了使用的模型与学习算法，然后对于遇到的类别不平衡问题，提出了新的评价标准并重采样数据集，提升了模型的性能。本文还验证了模型的可解释性，证明了机器学习模型在预测脓毒症问题上给出的解释与现实基本一致。本文存在的局限性有：使用现有的机器学习模型，基本没有调整超参数（包括朴素贝叶斯模型中的先验）；没有专门对类别数据（如性别）进行处理而是将其视为连续数据，等等。

**参考文献**

[1] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research. 2002 Jun 1;16:321-57.

[2] He H, Bai Y, Garcia EA, Li S. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence) 2008 Jun 1 (pp. 1322-1328). IEEE.

[3] Brodersen KH, Ong CS, Stephan KE, Buhmann JM. The balanced accuracy and its posterior distribution. In2010 20th international conference on pattern recognition 2010 Aug 23 (pp. 3121-3124). IEEE.

[4] https://ai-pool.com/a/s/normalization-in-deep-learning

[5] <https://imbalanced-learn.org/stable/auto_examples/over-sampling/plot_comparison_over_sampling.html>

[6] Chang CC, Lin CJ. LIBSVM: a library for support vector machines. ACM transactions on intelligent systems and technology (TIST). 2011 May 6;2(3):1-27.

[7] Fan RE, Chang KW, Hsieh CJ, Wang XR, Lin CJ. LIBLINEAR: A library for large linear classification. the Journal of machine Learning research. 2008 Jun 1;9:1871-4.