机器学习方法在早产和低出生体重儿预测中的应用*

蒋雯音

（宁波卫生职业技术学院 宁波 315100）

〔摘要〕 应用机器学习方法构建早产儿和低出生体重儿的预测模型，包括逻辑回归、支持向量机和随机森林算法，运用交叉验证法得到不同算法的最优模型，综合准确率、F1 值和 AUC 值评估 3 种模型的预测性能，结果表明基于随机森林算法的模型预测效果最好。

〔关键词〕 机器学习；早产儿；低出生体重儿；逻辑回归；支持向量机；随机森林

〔中图分类号〕R－056 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2019.04.014

Application of Machine Learning Methods in Prediction of Preterm Birth and Low Birth Weight Infants *JIANG Wenyin, Ningbo College of Health Sciences, Ningbo 315100, China*

〔**Abstract**〕 The prediction model for Preterm Birth (PTB) and Low Birth Weight (LBW) infants has been built by adopting Machine Learning (ML) techniques including Logical Regression (LR), Support Vector Machine (SVM) and Random Forest (RF) algorithm. The paper figures out the optimal models of different algorithms through using cross validation and then evaluates the predication performance of 3 models according to accuracy rate, F1score and AUC value. The result shows that the model based on RF algorithm features the best prediction effect.

〔**Keywords**〕 Machine Learning (ML); Premature Brith (PTB); Low Birth Weight (LBW); Logical Regression (LR); Support Vector Machine (SVM); Random Forest (RF)

1 引言

随着全面“二孩”政策的放开，高龄和高危孕妇明显增多，这给妇幼保健管理和妇产科相关人员带来前所未有的挑战。高危妊娠中早产（PTB）和低出生体重儿（Low Birth Weight，LBW）成为重点关注的问题。世界卫生组织将早产儿定义为胎龄小于 37 周出生的新生儿，低出生体重儿是指出生体重在 2 500 克以下者。早产儿、低出生体重儿生命脆弱，抵抗力低下，容易患病导致死亡，是新生儿的脆弱人群，是围产儿死亡的重要原因之一。世界卫生组织呼吁社会更多地关注早产问题，加强相关研究，采取有效行动，从而增加预防、应对早产的办法。

随着存储技术的发展和医院信息化的深入，医院数据库积累大量的数据，传统的数据分析和

〔修回日期〕 2018－10－20

〔作者简介〕 蒋雯音，硕士，讲师，发表论文 10 篇，参编专著 1 部。

〔基金项目〕 浙江省教育厅一般科研项目“基于机器学习的高危妊娠中早产儿和低体重儿预测研究”（项目编号：Y201738666）。

处理方法无法获得数据之间的隐藏信息和内在关联，但通过数据分析挖掘潜在知识、信息和规律的需求广泛存在。机器学习（Machine Learning，ML）是利用已有数据训练出模型，然后使用模型预测的一种方法。充分利用积累的原始数据，应用机器学习方法建立早产儿、低出生体重儿的预测模型，以期对高危妊娠中早产和低出生体重儿预测和实施有效干预提供辅助决策，从而降低早产儿、低出生体重儿的出生率和病死率，提高出生人口质量。

2 相关技术与理论

2.1 机器学习方法

2.1.1 逻辑回归（Logical Regression，LR）

机器学习中的一种常用分类模型，主要用于二分类问题（即输出只有两种，分别代表两个类别），通常是利用已知的自变量来预测一个离散型因变量的值。逻辑回归基于 Sigmoid 函数，算法主要思想是：用极大似然估计法来求模型参数，通过建立似然函数 L（θ），对其进行简单变换后为损失函数 J（θ），再用优化方法（如梯度下降）迭代求解出最优（即最小化损失函数）的模型参数（θ）。由于算法具有计算量小、简单、高效等特点，实际应用非常广泛，主要应用于预测某些事件发生的概率。

2.1.2 支持向量机（Support Vector Machine，SVM）

在统计学习理论和结构风险最小原理基础上发展起来的一种机器学习方法，其机器学习策略是结构风险最小化原则。主要思想是寻找一个最优分割超平面，使得该平面两侧距超平面最近的两类样本之间的距离最大化，通过分割面将数据进行分类。具有全局最优、最大泛化能力、推广能力强等优点，在解决小样本、非线性及高维模式识别中表现出许多特有的优势，能够推广应用到函数拟合中，较好地解决许多实际预测问题[1]。

2.1.3 随机森林（Random Foresrt，RF）

一种以决策树为基本分类器，将多个决策树进行组合来提高预测精度的集成学习方法。随机森林的构建过程大致如下[2]：从原始训练集中随机有放回采样选出 m 个样本，共进行 n 次采样，生成 n 个训练集；分别训练生成 n 个决策树模型；对于单个决策树模型，假设训练样本特征的个数为 n，那么每次分裂时根据信息增益（或 GINI 指数）选择最好的特征进行分裂；每棵树一直分裂直到该节点的所有训练样例都属于同一类；将生成的多棵决策树组成随机森林。对多重共线性不敏感、准确率高，不容易过拟合，而且能处理维度很高的数据集，所以在众多算法中较为突出。

2.2 研究方法

本研究采用以上 3 种机器学习方法分别建立预测模型并进行比较，使用 PyCharm 作为集成开发环境，利用 Python 2.7 中 numpy、pandas、matplotlib 包进行数据处理和分析，应用机器学习包 scikit - learn（简称 sklearn）构建预测模型[3]，进行模型测试和评估。

3 构建预测模型

3.1 数据预处理

本研究数据来自宁波市某妇幼保健院的 725 例高危孕妇病例，根据《宁波市高危妊娠评分标准》并结合样本数据，初步选出高危妊娠因素共 26 个（其中每种危险因素分轻、中、重 3 个等级）。原始数据一般存在缺失值、冗余重复、格式不一致、维度高等问题，需要进行数据清洗和预处理工作。对于本研究数据集经过清洗异常样本、插补缺失值、数据变换等处理，使数据格式和质量达到建模要求，生成最终可用于研究使用的样本 705 条，数据集正负样本比例，见图 1。在现实机器学习任务中通常还要做进一步的特征选择[4]，即选取对问题研究有用的属性（相关特征），用于后续建立预测模型的输入变量。本研究通过过滤低方差特征并根据自变量与目标变量间的关联大小选择特征，最终选取出 14 个特征作为预测模型的输入变量，包括年龄、体重指数、遗传病史、流产、异常分娩史、异常妊娠史、消化系统（肝损 ALT、肝炎病毒）、内

分泌系统（糖尿病、甲亢、甲低）、肿瘤（子宫肌瘤、卵巢肿瘤）、胎位、胎儿、试管婴儿、不孕时间、宫高等常见高危因素。另外在数据集最后添加标签列，设置生产早产儿或低体重儿的样本标签为 1，正常婴儿体重的样本标签为 0。可见预测目标实际上就是一个二分类问题。

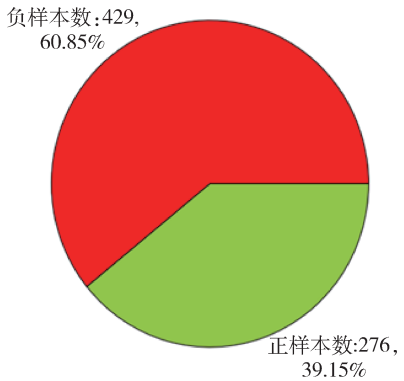

图 1 数据集正负样本比例

3.2 分割数据集

建立模型前用 sklearn. model_ selection 库中的 train_ test_ split 函数先将数据集按比例随机划分为训练集和测试集，语句如下：

```
X_ train, X_ test, y_ train, y_ test =
train_ test_ split (r_ data, labels, test_ size = 0. 3, random_ state = 20)
```

其中 r_ data 是数据样本的特征集，labels 是样本标签（取值为 1 或 0）；参数 test_ size = 0. 3 表示随机抽取样本的 30% 作为测试集用来测试模型性能，样本的 70% 作为训练集用于模型训练；设定 random_ state 参数，是该组随机数的编号，在需要重复试验时可以保证得到一组相同的随机数，即保证每次试验划分的训练集和测试集都是相同的。函数返回划分后的训练集和测试集以及各自对应的标签，其中 X_ train 和 y_ train 分别是训练数据的特征集及对应标签，X_ test 和 y_ test 分别是测试数据的特征集和标签。

3.3 建立和训练模型

数据准备好后分别调用 sklearn 中的 linear_ model. LogisticRegression、svm. svc 和 ensemble. RandomForestClassifier 类，即逻辑回归、支持向量机和随机森林 3 种分类模型分别对数据集进行训练和预测。通过以下方式建立模型：

```
from sklearn import linear_ model, svm, ensemble
lr_ model = linear_ model. LogisticRegression ()    #建立逻辑回归模型
svm_ model = svm. SVC ()    #建立支持向量机模型
rf_ model = ensemble. RandomForestClassifier ()    #建立随机森林模型
```

每种模型都有多个参数，当使用不同的参数配置时会产生预测性能不同的分类器，需要进行参数选择（即调参）。在实际训练中结果对于训练集的拟合程度通常较好，但是对于训练集之外的数据的拟合程度通常较差。本研究利用交叉验证方法[5]，基本思想是将训练数据集分为训练集和验证集，首先用训练集对分类器进行训练，再利用验证集来测试训练得到的模型性能。如 10 折交叉验证就是将数据集分成 10 份，轮流将其中 9 份做训练、1 份做验证，将 10 次结果的均值作为对算法精度的估计，由此可相对客观地判断这些参数对训练集之外的数据的拟合程度，从而得到最优参数和模型。利用 GridSearchCV 可系统地遍历多种参数组合，通过网格搜索确定最佳效果参数（即模型调参）[6-7]，形式如下：

```
from sklearn. grid_ search import GridSearchCV
#通过 GridSearchCV 来寻求最佳参数空间
gs_ model =
GridSearchCV (model_ name, param_ grid = params, cv = 6, scoring = 'roc_ auc')
```

其中 model_ name 是上述建立的 3 种模型名称（即 lr_ model、svm_ model 和 rf_ model），param_ grid 是需要优化的参数取值，cv = 6 表示采用 6 折交叉验证，scoring = 'roc_ auc' 表示以 auc 值作为评价标准，选取其值最高时的参数为模型最佳参数。各模型需要优化的参数及其取值范围名，见表 1。

表 1 各模型优化参数取值

模型名称（model_ name）	需优化参数（param_ grid）
逻辑回归（lr_ model）	params = {'C': [1e - 4, 1e - 3, 1e - 2, 0. 1, 1]}

续表 1

支持向量机（svm_ model）	params = {'C': [0.1, 1, 10, 100, 1000], 'gamma': [1e-5, 1e-4, 1e-3, 1e-2, 0.1]}
随机森林（rf_ model）	params = {'n_ estimators': [20, 40, 60, 80, 100]}

设置好模型和评价指标后用不同的参数通过 fit（）方法训练模型，交叉验证训练模型后利用 gs_ model. best_ params_ 属性得到最佳参数，运行结果如下。

```
逻辑回归模型：
交叉验证...
最优参数：{'C': 1}
支持向量机模型：
交叉验证...
最优参数：{'C': 10,'gamma': 0. 1}
随机森林模型：
交叉验证...
最优参数：{'n_ estima tors': 100}
```

用 gs_ model. best_ estimator_ 属性获得最优模型，然后使用最优模型的 predict（）方法进行预测，方法如下：

```
gs_ model. fit（X_ train, y_ train）    #训练模型
best_ model = gs_ model. best_ estimator_    #获得最优模型 best_ model
best_ model. predict（X_ test）    #利用最优模型进行预测
```

3.4 保存模型

训练好模型后使用 python 中的 pickle 模块来保存和读取模型，可利用此模型进行预测，实现语句如下：

```
import pickle
    pickle. dump（best_ model, f）    #保存模型
    model = pickle. load（f）    #加载模型
```

4 模型评估和分析

4.1 模型评估

模型训练后得到不同分类算法的最优模型，需比较不同算法模型的泛化性能，得到最适合本研究数据样本的预测模型。在训练模型前数据集的一部分作为测试集用来测试模型对新样本的预测性能，以评估模型在实际使用时的泛化能力。用模型预测结果和测试集真实值构建混淆矩阵，利用 sklearn. metrics 模块的 accuracy_ score、f1_ score、roc_ curve 函数分别计算出不同模型的准确率、F1 值、伪阳性率即错误命中率（False Positive Rate, FPR）和真阳性率即灵敏度（True Positive Rate, TPR），以 FPR 为横坐标、以 TPR 为纵坐标绘制 ROC 曲线，见图 2。求得 AUC 值，AUC 能够较好地评估模型的预测值与真实值之间的差异，3 种模型评估结果，见表 2。用 y_ test 表示测试集真实值，y_ train 表示测试集预测值，predict_ proba 表示测试集预测概率，具体实现如下：

```
from sklearn. metrics import precision_ score, f1_ score, roc_ curve, auc
    accuracy = accuracy_ score（y_ test, y_ train）    #准确率
    F1 = f1_ score（y_ test, y_ train）    #F1 值
    fpr, tpr, _ = roc_ curve（y_ test, predict_ proba）    #FPR 和 TPR
    auc（fpr, tpr）    #AUC 值
```

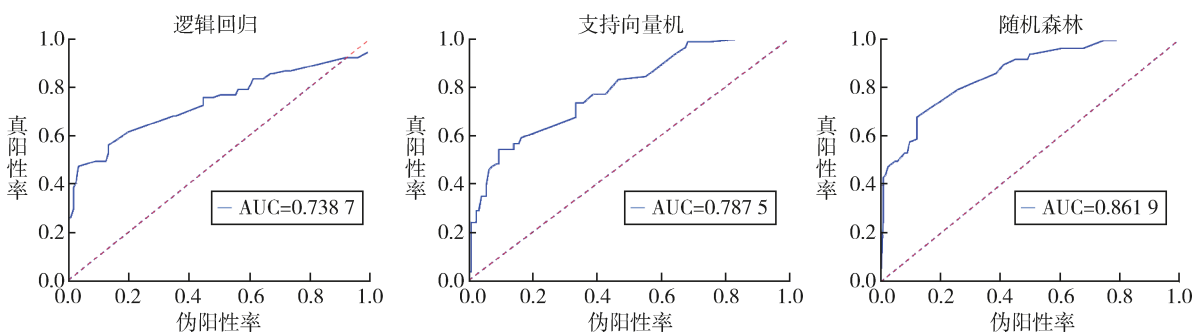

图 2 3 种模型的 POC 曲线比较

表 2 3 种预测模型评估结果

预测模型	准确率（%）	F1 值（%）	AUC 值
逻辑回归	81.50	53.11	0.738 7
支持向量机	90.56	73.43	0.787 5
随机森林	92.27	81.06	0.861 9

4.2 评估结果分析

根据各项评估指标对 3 类预测模型的性能比较后发现随机森林算法在测试结果中的各类指标上都高于其他两种算法，AUC 值为 0.861 9，若设定合适的阈值，模型具有一定的预测价值。因此用随机森林算法构建的模型更适合本研究数据集对于高危妊娠中早产和低体重儿的预测。最后利用基于随机森林算法的预测模型得到所有特征的权重值如下。

各 feature 的重要性：[0.115 777 33、0.023 856、0.078 223 5、0.115 411 45、0.057 786 33、0.125 978 77、0.078 623 07、0.035 271 67、0.036 850 01、0.015 649 09、0.186 131 75、0.030 612 83、0.016 419 54、0.083 399 06]

其大小反映与该特征对应的高危因素对早产及低出生体重儿影响程度的大小。结果显示在本数据集研究的各高危因素中导致早产和低出生体重儿的前几位高危因素分别是双胎及胎儿过大、疤痕子宫及附件手术史、年龄、流产（自然、人工）≥2 次及自然流产≥3 次。

5 结语

本研究应用机器学习方法探索效果最佳的高危妊娠中早产和低出生体重儿预测模型，以常用分类器评估指标来评估模型预测性能，确定基于随机森林算法的预测模型为适用于本研究的最优模型，利用模型分析得到导致早产、低出生体重儿的主要高危因素。实验结果一定程度上验证随机森林作为一种重要的基于 Bagging 的集成学习方法的优势，后期将会进一步研究和应用集成学习中的 Boosting 方法，继续探索预测性能更优的模型。此外还需要考虑与平台的整合以实现实时的数据分析和预测，辅助医生诊断，帮助医生及时关注高危人群并采取有效措施，为临床早期干预和妊娠期的关键监测提供参考，也希望为机器学习在医学辅助决策中的应用和发展做出一定的理论和实验贡献。

参考文献

1 吴鲲. 基于机器学习的学生成绩预警系统建模与研究［J］. 太原城市职业技术学院学报, 2016（12）：178－180.

2 CSDN 技术社区. 随机森林算法学习（RandomForest）［EB/OL］.［2017－10－21］. https://blog.csdn.net/qq547276542/article/details/78304454.

3 Fisman R, Iyengar S S, Kamenica E, et al. Gender Differences in Mate Selection: evidence from a speed dating experiment［J］. Quarterly Journal of Economics, 2006, 121（2）：673－697.

4 INRIA Feature Selection［EB/OL］.［2018－06－10］. http://scikit－learn.org/stable/modules/feature_selection.html.

5 Pang－NingTan, MichaelSteinbach, VipinKumar. 数据挖掘导论［M］. 北京：人民邮电出版社, 2011.

6 INRIA. Model Selection: Choosing estimators and their parameters［EB/OL］.［2018－06－10］. http://scikit－learn.org/stable/tutorial/statistical_inference/model_selection.html.

7 INRIA. Sklearn. Model_ Selection. GridSearchCV［EB/OL］.［2018－06－10］. http://scikit－learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html#sklearn.model_selection.GridSearchCV.