Statistics and Application 统计学与应用, 2020, 9(3), 345-356
Published Online June 2020 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2020.93037

Hans汉斯

Disease Risk Prediction Based on Stacking Integrated Learning Algorithm

—Using Data of Gestational Diabetes

Songqi Zhou*, Ying Bai

Mathematics and Statistics School of Northeastern University at Qinhuangdao, Qinhuangdao Hebei
Email: *328522803@qq.com

Received: May 11th, 2020; accepted: May 25th, 2020; published: Jun. 2nd, 2020

Abstract

In this paper, four missing value processing schemes are used for missing value processing, and the pros and cons of these four missing value processing schemes are compared and analyzed based on six machine learning algorithms. For each machine learning algorithm, this article gives the measures that should be taken to prevent the algorithm model from overfitting. By comparing the F1 values of the prediction results of each algorithm, the appropriate algorithm model is selected as the primary of the Stacking integrated learning algorithm. The learner then selects the logistic regression algorithm as the secondary learner of the ensemble learning algorithm. Finally, by adjusting the parameters of the logistic regression algorithm, a Stacking ensemble learning algorithm model based on the risk prediction problem of gestational diabetes is obtained with high accuracy and generalization ability.

Keywords

KNN, MLPC, GBDT, Random Forest, SVM, Naive Bayes, Stacking

基于Stacking集成学习算法的疾病风险预测

——以妊娠糖尿病为例

周颂奇*，白　颖

东北大学秦皇岛分校数学与统计学院，河北　秦皇岛
Email: *328522803@qq.com

收稿日期：2020年5月11日；录用日期：2020年5月25日；发布日期：2020年6月2日

*通讯作者。

文章引用: 周颂奇, 白颖. 基于 Stacking 集成学习算法的疾病风险预测[J]. 统计学与应用, 2020, 9(3): 345-356.
DOI: 10.12677/sa.2020.93037

摘　要

本文共采用了四种缺失值处理方案进行缺失值处理，并根据六种机器学习算法分析比较出了这四种缺失值处理方案的优劣程度。对于每一种机器学习算法，本文都给出了为防止算法模型过拟合所应采取的措施，并通过比较各算法预测结果的F1值，筛选出合适的算法模型作为Stacking集成学习算法的初级学习器，然后选取逻辑回归算法为该集成学习算法的次级学习器。最终，通过调节逻辑回归算法的参数得到精度高、泛化能力强的基于妊娠期糖尿病患病风险预测问题的Stacking集成学习算法模型。

关键词

KNN，MLPC，GBDT，随机森林，SVM，朴素贝叶斯，Stacking

1. 引言

2017 年国家二孩政策全面放开后，随着高龄、前次妊娠糖尿病史的妇女怀孕，妊娠糖尿病的发病率进一步增加，至此，妊娠糖尿病已经成为孕期最常见的并发症之一，形势极其严峻。所以，通过妊娠糖尿病的早筛查、早发现、早干预，减缓、阻止妊娠糖尿病的发生和发展具有重要的意义。本文利用天池精准医疗大赛复赛提供的数据集，对数据集中体检者的各项体检项目指标进行分析，并运用若干种机器学习算法进行结果的分析与比较，最终得到预测精准度颇高的 Stacking 集成学习算法模型。

2. 数据处理

2.1. 数据归一化

由于不同评价指标具有不同的量纲，这样会影响到数据分析的结果，为了消除指标之间量纲的影响，需要进行数据归一化处理，使各指标处于同一数量级，方便进行对比评价。本文采用离差标准化，即根据

$$x^* = \frac{x - \min}{\max - \min} \tag{1}$$

对原始数据进行归一化处理[1]。

2.2. 冗余数据清洗

去掉 id、孕前身高和体重特征。

2.3. 缺失值处理

由于样本数量较小，且缺失值占比较大。各特征中缺失值所占比例(见图 1)。若删除全部缺失值则会导致样本数量过小而无法进行数据分析与建模。故此次需要对缺失值进行填补。为比较不同处理方法对预测结果的影响，将分别采取以下方案进行预测精度的对比。

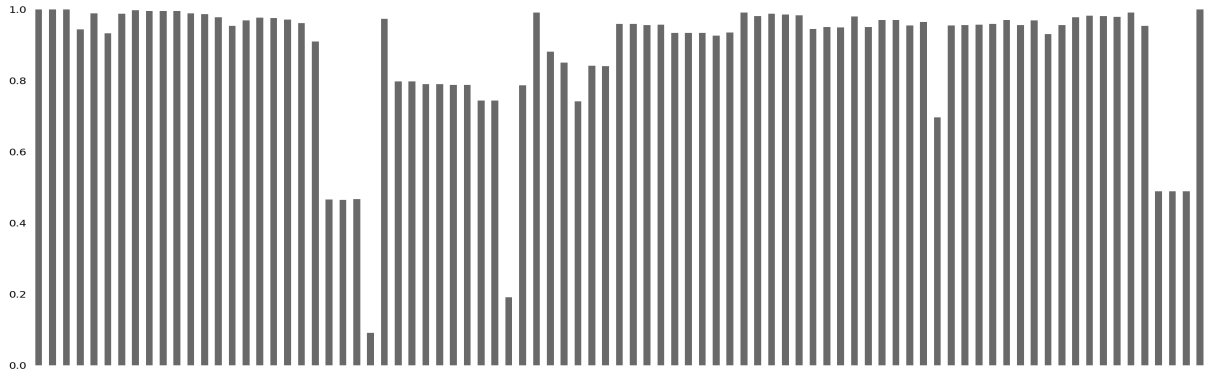

Figure 1. Curve: system result of standard experiment
图 1. 标准试验系统结果曲线

方案一：随机森林填补缺失值[2]

由于样本缺失较多，完整特征比较少，少数完整特征难以估计整体数据分布。因此，采用遍历所有特征的方式，从缺失值最少的特征开始填补(需填补缺失值较少的特征所需要的准确信息也较少)。填补一个特征时，其它特征用中位数暂时代替，每完成一次预测，就将预测值放回到原本的特征矩阵中。

方案二：KNN 填补缺失值

将样本数据做归一化处理之后，进行 k 近邻填补，用相邻样本的中位数填补缺失值。

方案三：不做处理(将缺失值替换为−1)

方案四：分特征处理缺失值

缺失值比例大于 10%的特征不做处理(将缺失值替换为−1)；缺失值比例小于 10%特征用中位数填充。

2.4. 异常值剔除

根据拉以达准则，剔除所有特征中标准差在$(-3\delta,3\delta)$之外的所有样本数据[3]。

2.5. 相关性检验

相关系数 r 的绝对值大小表示两个指标之间的相关强度，为了定量分析两个指标之间的关系，可以通过计算相关系数来进行分析。相关系数 r 的计算公式如下：其中 x, y 为任意两个指标。

$$r=\frac{\sum_{i=1}^{850}\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)}{\sqrt{\sum_{i=1}^{850}\left(x_i-\bar{x}\right)^2}\sqrt{\sum_{i=1}^{850}\left(y_i-\bar{y}\right)^2}} \tag{2}$$

相关系数 r 的绝对值大小表示两个指标之间的相关强度，具体关系如下表 1：

Table 1. Correlation intensity table
表 1. 相关强度表

$0.7<\lvert r\rvert<1$	$0.4<\lvert r\rvert<0.7$	$0.2<\lvert r\rvert<0.4$	$\lvert r\rvert<0.2$
高度相关	中等相关	低度相关	极低相关

根据上述相关系数公式，得到相关性最高的 10 个指标间的相关性强度(见图 2)。从图 2 中可以看出，除对角线是自相关外，大部分指标之间都极低相关，个别低度相关。

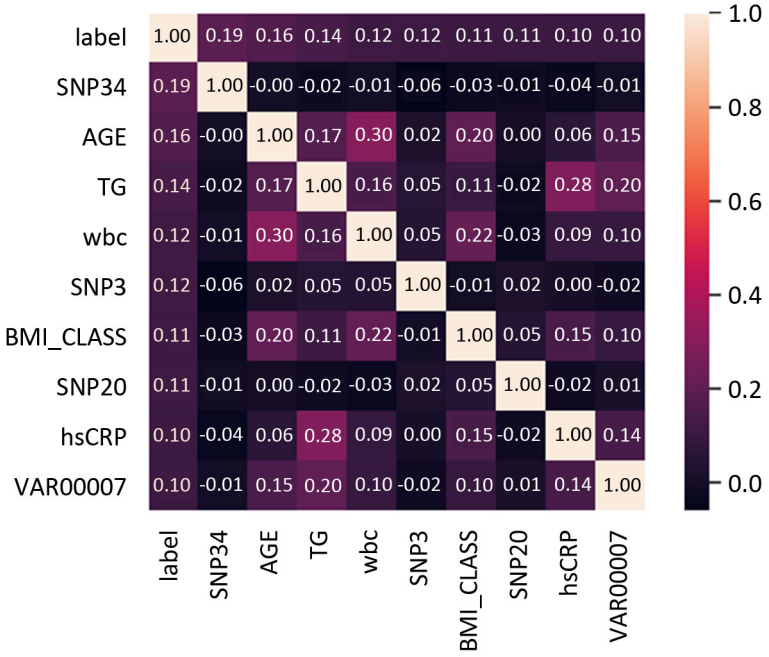

Figure 2. Correlation coefficient graph between some indicators
图 2. 部分指标间相关系数图

3. 特征选择

3.1. 去掉方差较小的特征

方差阈值是特征选择的一个简单方法，即去掉那些方差没有达到阈值的特征[4]。

3.2. 单变量特征选择

单变量特征提取的原理是分别计算每个特征的某个统计指标，根据该指标来选取特征。本文选择根据卡方统计量排名前 K 个的特征。通过卡方检验得到的特征之间是最可能独立的随机变量，因此这些特征的区分度很高。

4. 建模调参

4.1. KNN 算法

KNN 算法的核心思想为若一个样本在特征空间中的 K 个最相邻的样本中的大多数属于某一个类别，则该样本也属于这个类别，并具有这个类别上样本的特性。在训练集中数据和标签已知的情况下，输入测试数据，将测试数据的特征与训练集中对应的特征进行比较，找到训练集中与之最为相似的前 K 个数据，则该测试数据对应的类别就是 K 个数据中出现次数最多的那个分类。

4.1.1. KNN 算法流程[5]

① 计算测试数据与各个训练数据之间的距离；
② 按照距离的递增关系进行排序；
③ 选取距离最小的 K 个点；
④ 确定前 K 个点所在类别的出现频率；
⑤ 返回前 K 个点中出现频率最高的类别作为测试数据的预测分类。

4.1.2. KNN 具体参数设定

① K 值的选择

K 值的选择会对算法的结果产生重大影响。K 值较小意味着只有与输入实例较近的训练实例才会对预测结果起作用，但容易发生过拟合；如果 K 值较大，优点是可以减少学习的估计误差，但缺点是学习的近似误差增大，这时与输入实例较远的训练实例也会对预测起作用，使预测发生错误。本文采用网络搜索和 4 折交叉验证的方式，按从小到大的方式选择出较合适的 K 值。

② 距离度量的方式

有很多距离度量的方式，本问采用最常归的欧式距离，即对于两个 n 维向量 x 和 y，两者的欧式距离定义为：

$$D(x,y)=\sqrt{(x_1-y_1)^2+(x_2-y_2)^2+\cdots+(x_n-y_n)^2} \tag{3}$$

③ 分类决策规则

该算法中的分类决策规则往往是多数表决，即由输入实例的 K 个最临近的训练实例中的多数类别决定输入实例的类别。

4.1.3. 防止过拟合措施

由于 K 值较小容易发生过拟合，因此，我们尽量控制 K 值的大小，使 K 值尽量偏大一些。通过网络搜索与 4 折交叉验证，我们最终选择 K 为 9。

4.1.4. 结果分析

针对缺失值的不同处理方案，分别得到结果见表 2：

Table 2. F1 value table
表 2. F1 数值表

方案类别	方案一	方案二	方案三	方案四
F1	0.641	0.663	0.675	0.686

4.2. 多层感知器分类器[6]

多层感知器分类器(MLPC)是基于前馈人工神经网络(ANN)的分类器，由多个节点层组成。每个层完全连接到网络中的下一层。输入层中的节点表示输入数据。所有其他节点，通过输入与节点的权重 w 和偏置 b 的线性组合，并应用激活函数，将输入映射到输出。对于具有 K + 1 层的 MLPC，这可以用矩阵形式写成如下：

$$y(x)=f_k\left(\cdots f_2\left(w_2^T f_1\left(w_1^T x+b_1\right)+b_2\right)\cdots+b_k\right) \tag{4}$$

4.2.1. MLPC 算法流程

① 网络初始化：根据系统的输入确定网络的输入层节点的个数，隐含层节点的个数，输出层节点的个数，输入层、隐含层和输出层神经元之间的连接权值 w_{ij}， w_{jk}。初始化隐含层阈值、输出层阈值，给定学习速率和激励函数；

② 隐含层输出计算：根据权值和阈值，计算隐含层输出 H；

③ 输出层输出计算：根据隐含层输出 H，连接权值和阈值，计算 MLPC 神经网络预测输出 O；

④ 误差计算：根据网络预测输出和期望输出，计算网络预测误差；

⑤ 权值更新：根据网络预测误差更新连接权重；

⑥ 阈值更新：根据误差更新节点阈值；

⑦ 判断算法迭代是否结束，若没有结束则返回步骤②。

4.2.2. 防止过拟合措施

控制神经网络的复杂度的方法有很多种，如：隐层的个数、每个隐层中的单元个数与正则化(alpha)。神经网络的调参首先应创建一个大到足以过拟合的网络，确保这个网络可以对任务进行学习，然后通过缩小网络或者增大 alpha 来增强正则化，从而提高泛化性能。

4.2.3. 结果分析

针对缺失值的不同处理方案，分别得到结果见表 3 (此时 F1 为多次输出取平均值)：

Table 3. F1 value table
表 3. F1 数值表

方案类别	方案一	方案二	方案三	方案四
F1	0.760	0.781	0.797	0.803

4.3. 随机森林

随机森林指的是利用多棵树对样本进行训练并预测的一种分类器，是一种重要的基于 Bagging 的集成学习方法[7]。

4.3.1. RF 算法流程

① 给定训练集 S，测试集 T，特征维数 F。确定参数：使用到的 CART 的数量 t，每棵树的深度 d，每个节点使用到的特征数量 f，终止条件：节点上最少样本数 s，节点上最少的信息增益 m。

对于第 $1-t$ 棵树，$i=1-t$：

② 从 S 中有放回的抽取大小和 S 一样的训练集 S (i)，作为根节点的样本，从根节点开始训练。

③ 如果当前节点上达到终止条件，则设置当前节点为叶子节点，由于是分类问题，该叶子节点的预测输出为当前节点样本集合中数量最多的那一类 c (j)，概率 p 为 c (j)占当前样本集的比例。然后继续训练其他节点。如果当前节点没有达到终止条件，则从 F 维特征中无放回的随机选取 f 维特征。利用这 f 维特征，寻找分类效果最好的一维特征 k 及其阈值 th，当前节点上样本第 k 维特征小于 th 的样本被划分到左节点，其余的被划分到右节点。继续训练其他节点。

④ 重复②③直到所有节点都训练过了或者被标记为叶子节点。

⑤ 重复②③④直到所有 CART 都被训练过。

4.3.2. 防止过拟合措施

由于随机性的引入，使得随机森林不容易过拟合，同时，也使得随机森林有很好的抗噪声能力。max_features 决定每棵树的随机性大小，调为较小值时可以再降低过拟合风险。除此之外，还可调整 max_depth 参数进行预剪枝[7]。

4.3.3. 结果分析

针对缺失值的不同处理方案，分别得到结果见表 4 (此时 F1 为多次输出取平均值)：

Table 4. F1 value table
表 4. F1 数值表

方案类别	方案一	方案二	方案三	方案四
F1	0.748	0.772	0.799	0.810

4.4. 梯度提升树

梯度提升采用连续的方式构造树，每棵树都试图纠正前一棵树的错误。背后的主要思想是合并许多简单的模型。

4.4.1. GBDT 算法流程

① 在训练数据集所在的输入空间中，递归的将每个区域划分为两个子区域并决定每个子区域上的输出值，构建二叉决策树，选择最优切分变量与切分点 s，求解 j

$$\min_{j,s}\left[\min_{c_1}\sum_{x_i\in R_1(j,s)}\left(y_i-c_1\right)^2+\min_{c_2}\sum_{x_i\in R_2(j,s)}\left(y_i-c_2\right)^2\right]$$

遍历变量 j，对固定的切分变量 j 扫描切分点 s，选择使得上式达到最小值的对 (j,s)。

② 用选定的对 (j,s) 划分区域并决定相应的输出值：

$$R_1(j,s)=x\left|x^{(j)}\le s, R_2(j,s)=x\right|x^{(j)}>s$$

$$\hat{c}_m=\frac{1}{N}\sum_{x_1\in R_m(j,s)}y_i, x\in R_m, m=1,2 \tag{5}$$

③ 继续对两个子区域调用步骤①和②，直至满足停止条件。

④ 将输入空间划分为 M 个区域 $R_1,R_2,\cdots,R_M$，生成决策树。

4.4.2. 防止过拟合措施

由于增大 learning_rate 或 n_estimators 都会增加模型的复杂度，所以，降低树的最大深度和学习率都能降低过拟合。

4.4.3. 结果分析

针对缺失值的不同处理方案，分别得到结果见表 5：

Table 5. F1 value table
表 5. F1 数值表

方案类别	方案一	方案二	方案三	方案四
F1	0.787	0.809	0.817	0.805

4.5. 支持向量机

支持向量机它在解决小样本、非线性及高维模式识别中表现出许多特有的优势，支持向量机方法是根据有限的样本信息在模型的复杂性(即对特定训练样本的学习精度)和学习能力(即无错误地识别任意样本的能力)之间寻求最佳折中，以期获得最好的推广能力。

4.5.1. SVM 算法原理

SVM 方法是通过一个非线性映射 p，把样本空间映射到一个高维乃至无穷维的特征空间中(Hilbert

空间)，使得在原来的样本空间中非线性可分的问题转化为在特征空间中的线性可分的问题。简单地说，就是升维和线性化。升维，就是把样本向高维空间做映射，一般情况下这会增加计算的复杂性，甚至会引起“维数灾难”，因而人们很少问津。但是作为分类等问题来说，很可能在低维样本空间无法线性处理的样本集，在高维特征空间中却可以通过一个线性超平面实现线性划分。一般的升维都会带来计算的复杂化，SVM 方法巧妙地解决了这个难题：应用核函数的展开定理，就不需要知道非线性映射的显式表达式。由于是在高维特征空间中建立线性学习机，所以与线性模型相比，不但几乎不增加计算的复杂性，而且在某种程度上避免了“维数灾难”[5]。

4.5.2. 防止过拟合措施

解决过拟合的办法是为 SVM 引入了松弛变量 ξ，将 SVM 公式的约束条件改为：

$$y_i\left(w^T x_i + b\right) \geq 1 - \xi_i, i = 1, \cdots, n \tag{6}$$

同时，降低惩罚因子 C 的参数值能有效避免过拟合。

4.5.3. 结果分析

针对缺失值的不同处理方案，分别得到结果见表 6 (此时 F1 为多次输出取平均值)：

Table 6. F1 value table
表 6. F1 数值表

方案类别	方案一	方案二	方案三	方案四
F1	0.776	0.805	0.805	0.808

4.6. 朴素贝叶斯

4.6.1. 朴素贝叶斯算法原理[6]

朴素贝叶斯法是基于贝叶斯定理与特征条件独立假设的分类方法，基于一个简单的假定，即给定目标值时属性之间相互条件独立。对于给定的训练数据集，首先基于特征条件独立假设学习输入/输出的联合概率分布；然后基于此模型，对给定的输入 x，利用贝叶斯定理求出后验概率最大的输出 y。通过求解后验概率，并依据后验概率的值来进行分类。

4.6.2. 欠拟合原因分析

由于需要知道先验概率，且先验概率很多时候取决于假设，假设的模型可以有很多种，因此在某些时候会由于假设的先验模型的原因导致预测效果不佳。并且，我们是通过先验和数据来决定后验的概率从而决定分类，所以分类决策存在一定的错误率。除此之外，朴素贝叶斯对输入数据的表达形式很敏感，因此，本文方案中将用−1 替换缺失值的策略改为用 100 替换。

4.6.3. 结果分析

针对缺失值的不同处理方案，分别得到结果见表 7 (此时 F1 为多次输出取平均值)：

Table 7. F1 value table
表 7. F1 数值表

方案类别	方案一	方案二	方案三	方案四
F1	0.432	0.411	0.659	0.511

4.7. 缺失值处理方案分析与算法选择

4.7.1. 缺失值处理方案分析

针对不同机器学习算法，已求出每种缺失值处理方案相对应的 F1 值。从图 3 中可以定性地观察出每种处理方案的好坏。

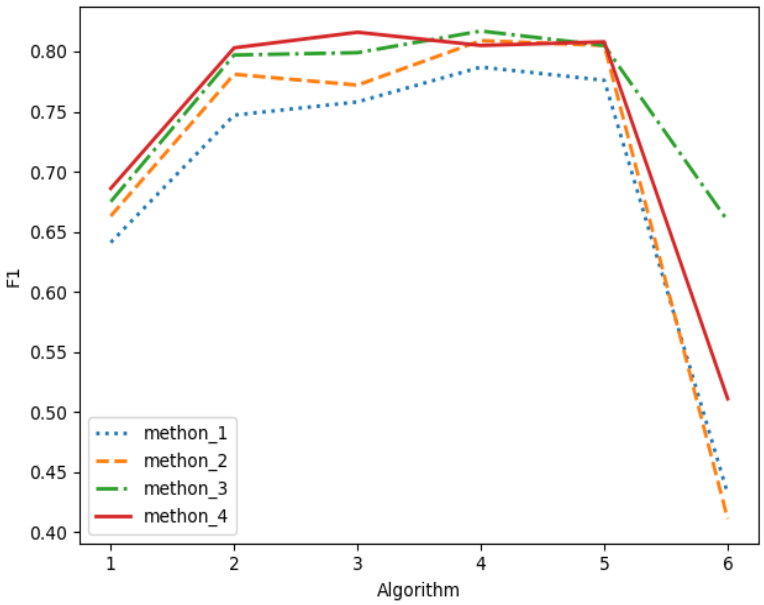

Figure 3. F1 value discount graph of algorithm under four processing schemes
图 3. 四种处理方案下算法 F1 值折现图

为定量分析出不同处理方案的优劣性，并且，考虑到朴素贝叶斯算法的拟合效果极差，下面我们将分别求出不同算法的 F1 均值与去掉贝叶斯算法后的 F1 均值。见表 8。

从表中数据，我们可以得出结论：方案四处理缺失值效果最好，其次是方案三，方案一处理缺失值的效果最差。即：分特征处理 > 不做处理 > KNN 填充 > 随机森林填充。

Table 8. F1 value table
表 8. F1 数值表

方案类别	方案一	方案二	方案三	方案四
F1	0.690	0.707	0.759	0.738
去掉 Beyes 后 F1 均值	0.742	0.766	0.779	0.784

4.7.2. 算法选择

利用方案四中处理缺失值的方法进行缺失值处理，然后，绘制出每种机器学习算法的 ROC 曲线与 AUC 示意图(图 4) [8]。朴素贝叶斯算法拟合效果极差，在这里我们直接去除该算法，用余下的五种算法继续做优化。

4.8. Stacking 集成算法

当训练数据很多时，一种更强大的策略是使用“学习法”，即通过另一个学习器来进行结合。Stacking 是学习法的典型代表。这里我们把个体学习器称为初级学习器，用于结合的学习器称为次级学习器。

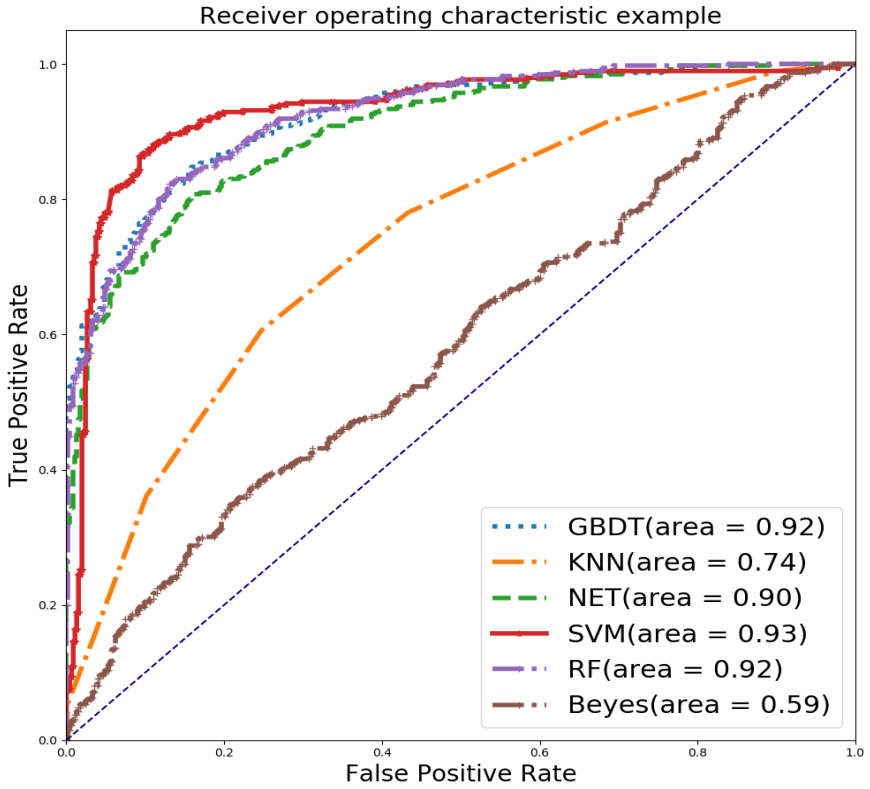

Figure4. Curve of ROC-AUC
图 4. ROC-AUC 曲线图

4.8.1. Stacking 算法原理[9]

① 对于 Model1，将训练集 D 分为 k 份，对于每一份，用剩余数据集训练模型，然后预测出这一份的结果。

② 重复上面步骤，直到每一份都预测出来。得到次级模型的训练集。

③ 得到 k 份测试集，平均后得到次级模型的测试集。

④ 对于 Model2、Model3…重复以上情况，得到 M 维数据。

⑤ 选定次级模型，进行训练预测。本文最后一层用的是逻辑回归模型。

4.8.2. 算法集成

将上述筛选出的五种机器学习算法作为初级学习器，选择逻辑回归算法作为次级处理器进行 stacking 集成学习(见图 5)。

这里，为避免过拟合，提高模型的泛化能力，我们给 LR 的参数 C 一个较小数值，调整完逻辑回归模型参数后就最终得到了基于妊娠期糖尿病患病风险预测问题的 Stacking 集成学习算法。多次预测求得 F1 平均值为 0.833，预测精度为 0.858，AUC 值为 0.93。该模型对应的 ROC 曲线如图 6。各种算法对应的 F1 值见表 9。

Table 9. F1 value table
表 9. F1 数值表

算法	KNN	MLPC	RF	GBDT	SVM	Stacking
F1	0.686	0.803	0.816	0.805	0.808	0.833

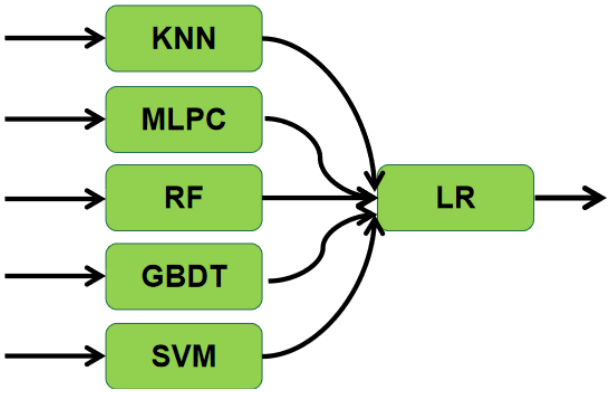

Figure 5. Stacking integrated learning diagram
图 5. Stacking 集成学习示意图

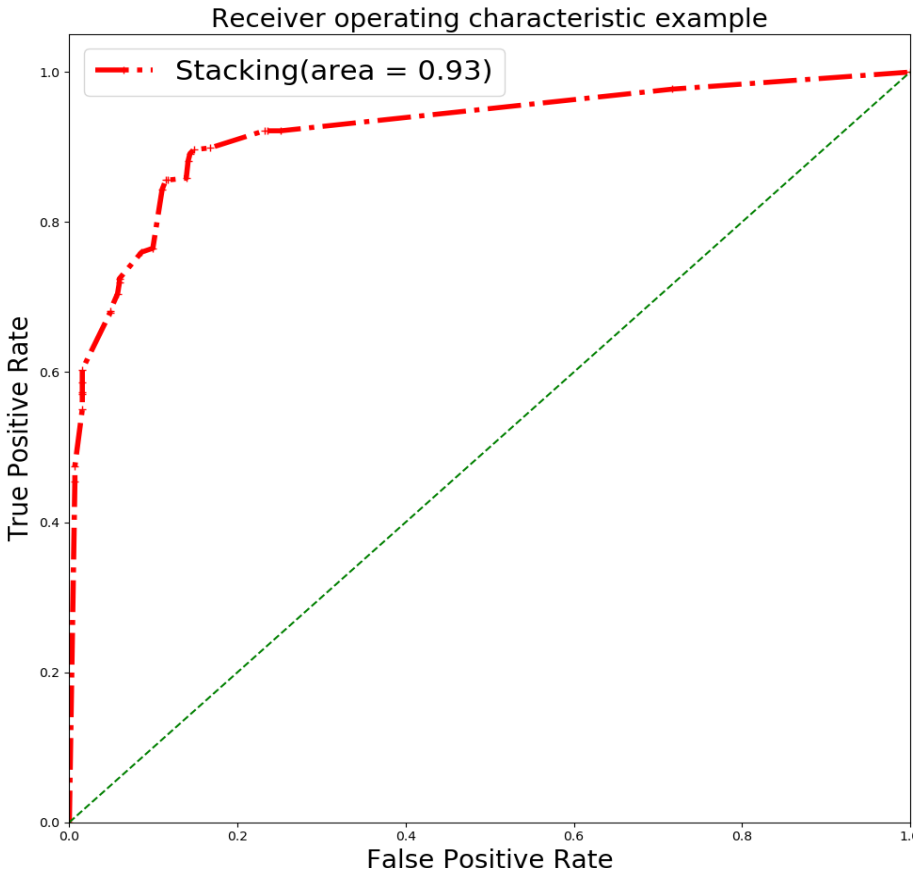

Figure 6. ROC-AUC curve of stacking
图 6. Stacking 的 ROC-AUC 曲线图

5. 结论

1) 通过对比分析四种不同的缺失值处理方案，得到了适合于解决该问题数据缺失值的处理方案，即用分特征的方案处理，将缺失值比例大于 10%的特征缺失值替换为−1，缺失值比例小于 10%特征用中位数填充，得到的处理效果最好。

2) 本文以倾向于提高模型的泛化能力为目标，对于每一种机器学习算法，均根据算法本身的特点分

别采用了适合于各自算法的避免过拟合的方法。并且通过对图像和数据的分析筛选出了能较好预测出该问题结果的算法模型，为 Stacking 集成学习算法做准备。

3) 以 KNN、MLPC、GBDT、随机森林、SVM 等算法作为初级学习器，以逻辑回归算法作为次级处理器的 Stacking 集成学习算法模型通过适当调整逻辑回归模型的参数，最终使得基于妊娠期糖尿病预测问题的 Stacking 集成学习算法模型在保证模型具有较强泛化能力的条件下进一步提升了糖尿病患病预测结果的可靠性。

参考文献

[1] 学习建模-个人信用风险评估模型实例[EB/OL]. https://www.zhihu.com/tardis/sogou/art/37355703, 2018-06-20.

[2] 张良均, 王璐, 谭立云, 等. Python 数据分析与挖掘实战[M]. 北京: 机械工业出版社, 2016: 23.

[3] 动脉网蛋壳研究院. 大数据 + 医疗: 科学时代的思维与决策开本[M]. 北京: 机械工业出版社, 2019: 21.

[4] 酒卷隆治, 里洋平. 数据分析实战[M]. 北京: 民邮电出版社, 2017.

[5] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016: 73.

[6] 李航. 统计学习方法[M]. 北京: 清华大学出版社, 2012: 35.

[7] Harrington, P. 机器学习实战[M]. 北京: 人民邮电出版社, 2013: 15.

[8] 萨扬·穆霍帕迪亚. Python 高级数据分析[M]. 北京: 机械工业出版社, 2019: 23.

[9] stacking 算法原理及代码[EB/OL]. https://www.cnblogs.com/dudumiaomiao/p/9692935.html, 2018-09-23.