Advances in Applied Mathematics 应用数学进展, 2021, 10(10), 3261-3268
Published Online October 2021 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2021.1010342

Hans汉斯

基于机器学习的前列腺癌患者分类研究

林中飞，王南芳，李小红

重庆理工大学理学院，重庆

收稿日期：2021年9月8日；录用日期：2021年9月29日；发布日期：2021年10月11日

摘　要

基于机器学习算法，本文对前列腺癌患者构建分类模型，目的是区分出不同患者，从而采取不同的治疗措施。基于随机森林、Adaboost、GradienBoosting、XGBoost、LightBGM和Stacking模型融合算法构建前列腺癌分类预测模型，调整模型参数，并验证模型效能。在单一的机器学习算法中，每一种机器学习算法都能对第一类(前列腺增生)患者进行识别，一部分模型在第二类(前列腺癌)患者和第三类(同时有前列腺癌和前列腺增生)患者中预测错误率较高；Adaboost算法的性能最优，对每一类都能够进行有效识别；Stacking融合算法优于所有单一的机器学习算法，在测试集上的准确率达到了96%。在前列腺癌分类预测模型中，Stacking融合算法效果明显优于单一的机器学习算法。

关键词

融合算法，机器学习，前列腺癌，分类

Research on Prostate Patients Classification Based on Machine Learning

Zhongfei Lin, Nanfang Wang, Xiaohong Li

College of Science, Chongqing University of Technology, Chongqing

Received: Sep. 8th, 2021; accepted: Sep. 29th, 2021; published: Oct. 11th, 2021

Abstract

This paper aims to distinguish different kinds of prostate cancer patients. We constructed a classification model for patients based on machine learning algorithms, so that we can take different treatment measures. Based on Random Forest, Adaboost, GradienBoosting, XGBoost, LightGBM and Stacking fusion algorithm, we constructed a prostate cancer classification model. The parameters were adjusted, and the effectiveness of models was evaluated. In all single machine learning algorithms, they can identify patients in the first category (prostatic hyperplasia). Some models have a high er-

文章引用: 林中飞, 王南芳, 李小红. 基于机器学习的前列腺癌患者分类研究[J]. 应用数学进展, 2021, 10(10): 3261-3268. DOI: 10.12677/aam.2021.1010342

ror rate of classification in the second category (prostate cancer) and the third category (both prostatic hyperplasia and prostate cancer); Adaboost algorithm has the best performance and it can identify each category effectively. The accuracy of Stacking fusion algorithm is up to 96%, better than all single models. Stacking fusion algorithm is better than each single machine learning algorithm significantly in the prostate cancer classification.

Keywords

Fusion Algorithm, Machine Learning, Prostate Cancer, Classification

1. 绪论

对于男性泌尿生殖系统而言，最常见的恶性肿瘤之一非前列腺癌莫属[1]。伴随着生活方式的极大改变以及人口的老龄化，目前，前列腺癌是我国泌尿外科中发病率最高的一种，其平均发病率达到了男性恶性肿瘤的第六位。前列腺癌不仅给患者的生命和家人带来了一种不可磨灭的损失和伤害，也给社会和国家的发展带来了沉重的压力，它的平均死亡率和疾病发生数量都呈逐年递增的变化趋势，术前及时有效地进行诊断对于正确制订治疗计划以及进行预后评估具有非常好的意义[2]。

王逸飞等人[3]选择了 XGBoost 机器学习算法对前列腺癌的预测模型进行构建，同时利用 SHAP 方法解释模型特征的实际意义[4] [5] [6]。在实际生活中，MRI 是临床诊断前列腺癌主要检查方式，同时 T_2WI、T_1WI 增强扫描能够清晰显示肿瘤组织血管，前列腺含有丰富的供血组织，一定程度上影响检出率[7]。动态增强 MRI (dynamic enhanced MRI, DCE-MRI)有助于评估血流灌注及血管通透性，利用薄层扫描技术实现局部病变数据动态采集，提高诊断准确率[8]。法慧[9]使用结构方程模型对前列腺癌与前列腺增生进行鉴别诊断。

目前机器学习算法在数据分析中有很大优势，同时一些集成模型或者融合模型比一般的机器学习算法的性能更好，因此本文考虑将性能较好的算法用于前列腺患者数据中，根据前列腺疾病的一些特征，来预测患者患有前列腺疾病中的哪一种类型，从而给出精准的治疗措施。

2. 数据处理与统计方法

2.1. 数据采集与来源

本文采用的数据“前列腺肿瘤预警数据集”(来源于“国家人口健康科学数据中心数据仓储 PHDA”)。数据集中共包含 129 例前列腺增生患者(标签为 1)、46 例前列腺癌患者(标签为 2)及 25 例同时患有两种疾病的患者(标签为 3)的真实临床信息。

2.2. 数据预处理

删除数据缺失比率大于 90%的七个指标，保留指标 24 个。对于低密度脂蛋白胆固醇、钾、载脂蛋白 A1、载脂蛋白 B 这四个指标，缺失比率大约为 10%，使用默认值进行填补；对于其余指标，使用其平均值进行填补。经过预处理后，共保留特征 24 个，样本 197 例，其中前列腺增生患者 127 例，前列腺癌患

者 45 例，同时患有两种疾病的患者 25 例。

对于提取的不同指标，指标变量之间的量纲和数值的量级不一样，在训练模型之前，将这些数据进行标准化，能够有效地加速指标权重参数的收敛，消除各个指标之间的误差。为减少数据量纲对模型的影响，本文选择了归一化方式进行了预处理。

2.3. 指标选取

通过随机森林、方差阈值过滤法和交互信息过滤法三种方法对数据进行指标提取，利用交集的方法，最后共选取了 11 个指标，即：PSA (游离)、PSA (总)、肌酸激酶同工酶、球蛋白、高密度脂蛋白胆固醇、甘油三酯、血清尿酸、钠、无机磷、钙和肌酸激酶。

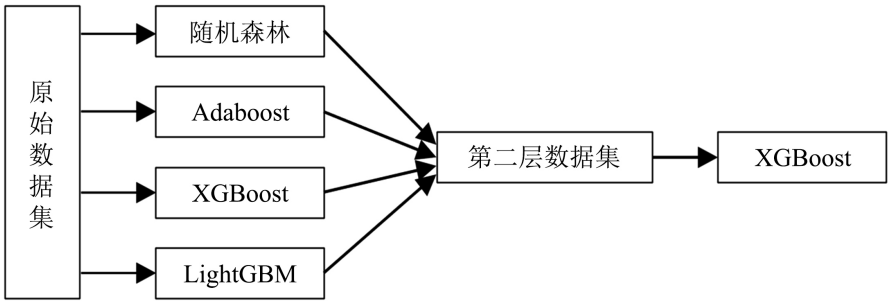

Figure 1. The structure of Stacking classification model
图 1. Stacking 分类模型结构图

2.4. 统计方法

2.4.1. 机器学习模型

本文通过数据分析软件 Python，采用了几种机器学习的方法对前列腺患者数据建立分类模型。最后根据每种机器学习的预测结果建立了 Stacking 分类模型。Stacking 分类模型的结构如图 1 所示。

Stacking 分类模型的效果取决于两个方面：一个是个体学习器的预测效果，通常个体分类器的预测效果越好，模型综合算法的预测效果越好；另一个是个体学习器之间需要有一定的差异性，因为每个模型的主要关注点不同，这样模型综合才能使每个个体学习器充分发挥其优点。

2.4.2. 模型优化及评价

在进行建立模型之前，将样本数据划分训练集、测试集和验证集，确定训练集、验证集和测试集的比例为 6:2.5:1.5。在验证集中利用 5 折交叉验证和网格搜索法确定几种机器学习算法的参数并完成模型训练。利用测试集的混淆矩阵计算 precision (精确率)、recall (召回率)和 f1-Score3 个指标，用于评价分类模型的性能。精确率指的是，对于预测结果而言，分类器判定为正的样本中含有多少对的样本；召回率指的是，对于样本而言，被预测正确的样本占正样本的比例；F1 分数兼顾了两者，是对于模型的准确率和召回比例之间的一种等价调和平均。

每一类评价指标的计算公式如下：

$$\text{precision} = \frac{\text{分类正确的正样本个数}}{\text{分类器判定为正样本的个数}} \tag{1}$$

$$\text{recall} = \frac{\text{分类正确的正样本个数}}{\text{真实正样本的个数}} \tag{2}$$

$$f1\_score = \frac{2 * precision * recall}{precision + recall} \tag{3}$$

3. 统计建模与分析

3.1. 描述性分析

通过对数据集进行预处理之后，先对该数据集中的年龄和患病类型两个字段做一个简单的统计，最小年龄为 40 岁，最大年龄为 89 岁，以十年为区间长度，一共分成五个区间。下面给出分年龄段患病人数的频数表，具体如果见表 1 所示。

Table 1. Frequency table of patients by age
表 1. 分年龄段患病人数的频数表

	人数	患前列腺增生人数	患前列腺癌人数	两者皆患的人数	比例
40~50	1	0	1	0	0.5%
50~60	9	5	4	0	4.6%
60~70	65	40	19	6	33%
70~80	104	70	19	15	52.8%
80~90	18	12	4	2	9.1%

从表 1 可以得知，前列腺患病的人数大多集中在 70~80 岁的年龄段，其次是在 60~70 岁年龄段，这两个两年龄段大约占了整个人数的 86%。甚至可以发现：前列腺患病的年龄段是以 60 岁为分界线，60 岁以上的人数占了 95%。这说明人老了，人的身体机能逐步下降，这时应该更加的注重自身身体的健康情况，疾病早发现治愈的可能性就越高。

再对选取的 11 个指标按照患病类型进行描述性分析，分析各个指标的平均值、最小值和最大值。结果如表 2 所示(表中的 1、2 和 3 分别表示前列腺增生、前列腺癌和同时患有前列腺增生与前列腺癌三类患者)。

Table 2. Descriptive statistical analysis of indicators
表 2. 指标描述性统计分析

	mean			min			max		
	1	2	3	1	2	3	1	2	3
PSA (游离)	1.450	2.513	1.558	0.04	0.03	0.03	14.7	25	13.5
PSA (总)	9.833	25.278	9.724	0.09	0.05	0.05	71.2	300	36.2
球蛋白	40.162	41.173	41.500	30.5	34.3	34.8	49.2	49.1	49.7
高密度脂蛋白胆固醇	1.230	1.173	1.143	0.61	0.62	0.68	2.11	1.87	2.16
肌酸激酶同工酶	11.979	11.298	13.156	2.3	3.8	4.3	61.2	20.4	20.7
无机磷	1.064	1.088	1.092	0.73	0.65	0.71	1.61	1.46	1.37
钙	2.204	2.225	2.251	1.96	1.98	1.99	2.48	2.4	2.51
肌酸激酶	93.472	98.256	93.844	19.9	35.1	13.1	348.7	278.9	177.4
甘油三酯	1.185	1.341	1.314	0.37	0.49	0.59	3.22	5.25	3.13
血清尿酸	329.306	308.622	347.952	119.3	135.8	256.5	669.5	446.7	424.1
钠	142.017	142.513	141.852	131.9	136.3	133.1	147	147	146.5

从表 2 中，PSA (游离)和 PSA (总)这两个指标在第二类别(前列腺癌)表现出较大的差异，平均值和最大值都是第二类别最大，这三个类别的最小值差别不大。球蛋白和高密度脂蛋白胆固醇两个指标，它们的均值在三类别中相差不大；球蛋白的最小值在第一类别(前列腺增生)是低于第二类别和第三类别(前列腺增生与前列腺癌同时患者)，最大值则差不多；高密度脂蛋白胆固醇的最小值差不多，但第二类别的最大值要低于其余两个类别。对于肌酸激酶同工酶指标，第三类别的平均值高于前面两个类别；最小值也各不相同，各具特点；第一个类别的最大值远远大于其余两个类别。对于无机磷、钙、甘油三酯和钠四个指标，在平均值、最小值和最大值上，三个类别的值都只有较小的差距，差距不大。肌酸激酶和血清尿酸在最大值上，三个类别的差距有点大，第一类别大于第二类别，第三类别的值最小；而在最小值上，第二类别的血清尿酸的值大于其余两个类别，第三个类别的肌酸激酶的值远大于其余两个类别。对于其余变量，三个类别的值都只有一点差别。

3.2. 分类模型建立与分析

3.2.1. 机器学习模型建模分析

本文分别利用随机森林、Adaboost、GradientBoosting、XGBoost、LightGBM 模型在训练集上进行建模并进行比较分析。首先，用验证集对于各个模型进行参数调优，结果如表 3 所示。

Table 3. The results of parameter adjustment of each model
表 3. 各个模型调参结果

模型名称	参数名称	调参范围	调参结果
随机森林	n_estimators	[50, 100, 200, 300, 400, 500]	100
	max_depth	[1, 2, 3, 4, 5, 6, 7, 8, 9]	2
	min_samples_leaf	[5, 10, 20, 30, 40, 50]	10
	min_samples_split	[5, 7, 9, 11, 13, 15]	5
Adaboost	n_estimators	[50, 100, 200, 300, 400, 500]	300
	learning_rate	[0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1]	0.9
GradientBoosting	n_estimators	[50, 100, 200, 300, 400, 500]	300
	learning_rate	[0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1]	0.7
	max_depth	[2, 3, 4, 5, 6, 7, 8, 9]	5
XGBoost	n_estimators	[50, 100, 200, 300, 400, 500]	200
	max_depth	[2, 3, 4, 5, 6, 7, 8, 9]	5
	learning_rate	[0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1]	0.5
LightGBM	n_estimators	[50, 100, 200, 300, 400, 500]	100
	num_leaves	[20, 25, 30, 35, 40]	20
	learning_rate	[0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1]	0.1

其次，对各个模型参数调优前后模型进行评价，采用了三个指标即 precision(精确率)、recall(召回率)和 f1_score 对模型调优结果进行比较评价。

结果如表 4 所示(表中的 1、2 和 3 分别表示前列腺增生、前列腺癌和同时患有前列腺增生与前列腺癌三类患者)。

Table 4. Comparison of evaluation indicators before and after adjusting parameters of each model
表 4. 各个模型调参前后评价指标对比

模型名称			precision	recall	f1_score
随机森林	调参前	1	0.67	0.97	0.79
		2	0.75	0.25	0.38
		3	0	0	0
	调参后	1	0.64	1	0.78
		2	0	0	0
		3	0	0	0
Adaboost	调参前	1	0.67	0.91	0.72
		2	0.33	0.25	0.29
		3	0	0	0
	调参后	1	0.67	0.91	0.77
		2	0.40	0.17	0.24
		3	0.5	0.17	0.25
GradientBoosting	调参前	1	0.68	0.84	0.75
		2	0.44	0.33	0.38
		3	1	0.17	0.29
	调参后	1	0.65	0.81	0.72
		2	0.12	0.08	0.10
		3	0	0	0
XGBoost	调参前	1	0.66	0.78	0.71
		2	0.22	0.17	0.19
		3	0	0	0
	调参后	1	0.65	0.81	0.72
		2	0.33	0.25	0.29
		3	0	0	0
LightGBM	调参前	1	0.64	0.78	0.7
		2	0.11	0.08	0.1
		3	0	0	0
	调参后	1	0.67	0.91	0.77
		2	0.20	0.08	0.12
		3	0	0	0

从表 4 中的 precision (精确率)、recall (召回率)和 f1_score 三个指标来看，随机森林进行调优之后的效果还不如参数使用默认值，不调优的模型效果比决策树调优之后的效果要好一些，但都有一个同样的问题，对于第三类，模型不能友好地识别。Adaboost 调参之后的准确率、召回率和 f1_score 有略微提升，对第三类别也能够识别了，但是整体来看，准确率并不高。GradientBoosting 调参之后的准确率、召回率

和 f1_score 没有调参之前的效果好，而且调参之前对于第三类别能够识别，调参之后却不能够识别。XGBoost 进行参数调优之后，准确率、召回率和 f1_score 有略微提升，说明模型效果比较好，但对第三类的识别还是比较困难。LightGBM 模型进行参数调优之后，对于第一类别的准确率、召回率和 f1_score 有略微提升。

根据以上对于各个分类模型的评价结果分析来看，几乎每个模型对于第二类和第三类的识别度都很低，但每个模型的侧重点不一样，因此，接下来考虑将使用 stacking 模型，来尝试进行重新分类，stacking 模型里的基础模型根据以上的分析结果来选择。

3.2.2. Stacking 融合模型建模分析

不难发现随机森林、Adaboost、XGBoost、LightGBM 这四种模型的预测效果好于其他模型。本文为了提高模型的准确率与泛化能力，将上述四种模型进行综合，即利用这几个学习器去预测前列腺患者类型，得到的结果为新的训练集，将其输入 XGBoost，然后去训练一个新的模型，得到最终的预测模型。

经过对第二层的机器学习模型的参数调整后，将个体学习器和 Stacking 模型在测试集上进行预测，把得到的预测结果进行对比，结果如表 5 所示。

Table 5. Comparison of single learners and Stacking model
表 5. 个体学习器与 Stacking 模型对比

		预测为 1	预测 2	预测 3
随机森林	实际为 1	28	3	1
	实际为 2	11	1	0
	实际为 3	6	0	0
Adaboost	实际为 1	26	5	1
	实际为 2	11	1	0
	实际为 3	4	0	2
XGBoost	实际为 1	25	4	3
	实际为 2	11	1	0
	实际为 3	4	0	2
LightGBM	实际为 1	28	2	2
	实际为 2	10	2	0
	实际为 3	4	2	0
stacking 模型	实际为 1	31	1	0
	实际为 2	1	11	0
	实际为 3	0	0	6

由表 5 可见，个体学习器和 Stacking 模型在测试集上的预测准确率有一定的差距，使用 Stacking 模型对于第二类和第三类都能非常好的识别，准确率达到了 96%，高于所有个体分类器。Stacking 模型算法综合了各个体分类器的优点，并充分发挥了集成模型的性能，因此具有更强的泛化能力和更好的预测效果。

4. 结论

本文对数据进行了处理之后，利用随机森林、方差阈值过滤法和互信息过滤法，从 22 个指标值中，选取了 11 个指标作为构建模型的最终指标，这 11 个指标分别为：PSA (游离)、PSA (总)、肌酸激酶同工酶、球蛋白、高密度脂蛋白胆固醇、甘油三酯、血清尿酸、钠、无机磷、钙和肌酸激酶，在三类患者间的差异有统计学意义。

分别对三类患者的 11 个指标进行了描述性统计分析，发现 PSA (游离)和 PSA (总)这两个指标在三类患者身上表现出较大的差异，平均值和最大值都是前列腺癌患者最大。对于球蛋白和高密度脂蛋白胆固醇两个指标，前列腺增生患者的球蛋白的最小值低于前列腺癌患者与同时患有前列腺增生和前列腺癌的患者；前列腺癌患者的高密度脂蛋白胆固醇的最大值要低于其余两类患者。

利用机器学习算法对前列腺增生、前列腺癌以及同时患有前列腺增生和前列腺癌的患者构建分类模型，结果表明随机森林、Adaboost、XGBoost 和 LightGBM 四个模型效果较好，最后构建以 XGBoost 为最终学习器的 Stacking 模型，对三类患者进行了有效分类和预测，最终，Stacking 模型的准确率达到了 96%。

参考文献

[1] Berríos-Torres, S.I., Umscheid, C.A., Bratzler, D.W., *et al*. (2017) Centers for Disease Control and Prevention Guideline for the Prevention of Surgical Site Infection, 2017. *JAMA Surgery*, **152**, 784-791. https://doi.org/10.1001/jamasurg.2017.0904

[2] 方孙福, 周晓燕, 罗朝军, 等. 动态增强 MRI 结合 DWI 对前列腺癌的诊断价值分析[J]. 医学影像学杂志, 2017, 27(1): 186-188.

[3] 王逸飞, 吴欢, 薛万国, 等. 前列腺癌与前列腺增生的分类预测及癌症风险因素分析[J]. 解放军医学院学报, 2021, 42(3): 277-281+305.

[4] Lundberg, S.M., Erion, G.G. and Lee, S.I. (2018) Consistent Individualized Feature Attribution for Tree Ensembles. https://arxiv.org/abs/1802.03888

[5] 中华医学会泌尿外科分会. 前列腺癌诊断治疗指南[J]. 继续医学教育, 2007, 21(6): 30-39.

[6] 张洪侠, 郭贺, 王金霞, 等. 基于 XGBoost 算法的 2 型糖尿病精准预测模型研究[J]. 中国实验诊断学, 2018, 22(3): 408-412.

[7] Jiang, J.X., Xiao, Z.B., Tang, Z.H., Zhong, Y.F. and Qiang, J.W. (2018) Differentiating between Benign and Malignant Sinonasal Lesions Using Dynamic Contrast-Enhanced MRI and Intravoxel Incoherent Motion. *European Journal of Radiology*, **98**, 7-13. https://doi.org/10.1016/j.ejrad.2017.10.028

[8] 王华, 段青, 威晋, 等. 3.0T MRI 动态增强在鉴别前列腺增生及前列腺癌中的价值[J]. 福建医科大学学报, 2010,44(3): 223-226.

[9] 法慧. 基于结构方程模型的前列腺癌与前列腺增生鉴别诊断的影响因素研究[D]: [硕士学位论文]. 长春: 吉林大学, 2020.