杜　超,范馨月,单立平. 基于集成学习算法构建前列腺癌预测模型[J]. 中华医学图书情报杂志,2019,28(12):19-24.

DOI:10.3969/j.issn.1671-3982.2019.12.004

·专题:医疗信息化建设·

基于集成学习算法构建前列腺癌预测模型

杜　超,范馨月,单立平

[摘要]目的:利用机器学习算法建立前列腺癌诊断预测模型,为前列腺癌患者的穿刺术前诊断提供参考。方法:收集 2017 年 1 月-2018 年 12 月中国医科大学附属盛京医院泌尿外科接受前列腺穿刺的 255 例患者的临床信息作为变量,采用 Logistic 多因素分析、信息增益率两种方法筛选研究变量,应用十折交叉验证划分训练集和测试集,采用多种机器学习算法(RF,SVM,Logistic,Naive Bayes)建立前列腺癌诊断模型,收集 2019 年 1-6 月的 75 例患者作为验证集,进一步评估模型性能和临床应用的可能性。结果:应用信息增益率筛选变量所建立的模型性能优于 Logistic 多因素回归分析。在 4 种机器学习算法中,Naive Bayes 算法 AUC 最高,在试验集和验证集上分别为 0.826 和 0.797。RF 算法的 Precision 最高,在试验集和验证集上分别达到 0.839 和 0.791。结论:基于前列腺穿刺患者的多种临床信息,通过机器学习方法建立诊断预测模型具有较高的准确率,能够为前列腺癌的诊断提供一定参考。

[关键词]机器学习算法;前列腺癌;穿刺活检;多因素 Logistic 回归分析

[中图分类号]TP181;R737.25　　[文献标志码]A　　[文章编号]1671-3982(2019)12-0019-06

Integrated learning algorithm-based establishment of prostate cancer prediction model

DU Chao, FAN Xin-yue, SHAN Li-ping

(Affiliated Shengjing Hospital of China Medical University, Shenyang 100004, Liaoning Province, China)

Corresponding author: SHAN Li-ping

[**Abstract**] **Objective** To provide reference for the pre-puncture diagnosis of prostate cancer by establishing a prostate cancer prediction model using the machine learning algorithm. **Methods** The prostate cancer diagnostic model was established using RF, SVM, logistic and Naive Bayes machine learning algorithms with the clinical information of 255 prostate cancer patients (served as an experimental group) who underwent prostate puncture in our hospital from January 2017 to December 2019 served as its variables detected by multivariate logistic regression analysis and information gain rate analysis respectively. The performance and clinical application of the prostate cancer diagnostic model were further evaluated with 75 patients admitted to our hospital from January 2019 to June 2019 served as a validation group. **Results** The performance of the prostate cancer diagnostic model established with the variables detected by information gain rate analysis was better than that established with the variables detected by multivariate logistic regression analysis. The AUC measured by Naive Bayes algorithm was larger than that mea sured by RF, SVM and logistic algorithms, which was 0.826 and 0797 respectively in experimental group and validation group. The accuracy of RF algorithm was higher than that of naive Bayes, SVM and logistic algorithms, which was 0.839 and 0.791 respectively in experimental group and validation group. **Conclusion** The accuracy of prostate cancer diagnostic model established using the machine learning algorithms based on the clinical information of prostate cancer patients is rather high, and can thus provide certain reference for the diagnosis of prostate cancer.

[作者单位]中国医科大学附属盛京医院,辽宁 沈阳　110004

[作者简介]杜　超(1994-),男,辽宁兴城人,在读硕士研究生,研究方向为泌尿系肿瘤。

[通讯作者]单立平(1979-),男,山东烟台人,博士,硕士生导师,副教授,副主任医师,研究方向为泌尿系肿瘤。E-mail:18940259257@163.com

[**Key words**] Machine learning algorithm; Prostate cancer; Puncture biopsy; Multivariate logistic regression analysis

前列腺癌(Prostate Cancer, PCa)是男性最常见的癌症之一,也是全世界男性癌症死亡的第二大原因,2020 年,PCa 相关死亡人数估计为 385 560[1]。为了使前列腺癌患者能获得更好的预后及进一步提高其生活质量,前列腺癌的筛查和诊断已经成为当前研究的重点。

在临床上,前列腺癌需要经过前列腺穿刺活检才能够确诊,而穿刺前最常用的参考指标为前列腺特异性抗原(Prostate Specific Antigen, PSA)[2]。由于 PSA 浓度因受到炎症、射精、导尿操作等一系列非前列腺癌因素的影响而出现一过性升高,导致单纯使用 PSA 无法正确区分前列腺癌和前列腺增生[3]。如果持续采用单一指标诊断模式将造成漏诊或者不必要的活检。

近年来,随着计算机技术的迅猛发展,国内外一些研究开始采用机器学习方法进行前列腺癌的诊断[4-6],但都存在着一定的局限性。预测模型的变量主要为患者的 PSA 水平或 MRI 影像参数,没有综合考虑患者病史、化验及检查等指标,同时也没有将得到的模型在临床中进行验证。为克服前列腺癌单一诊断指标的局限性,本文综合当前的研究现状,收集前列腺癌患者的基本信息(年龄、体重)、病史、症状表现、化验结果及 MRI 检查等指标作为研究变量,通过两种变量筛选方法的比较,确定纳入模型的最佳变量组合,采用多种机器学习方法建立前列腺癌诊断预测模型,并将得到的模型应用于临床,评价模型的准确性,旨在揭示机器学习在前列腺癌诊断中的应用价值,为前列腺癌的早期诊断研究提供新的思路。

1 数据收集与预处理

1.1 数据收集及变量介绍

本文收集 2017 年 1 月-2018 年 12 月于中国医科大学附属盛京医院泌尿外科行超声引导下前列腺穿刺活检术的患者信息,包括患者年龄、血清总 PSA(total Prostate Specific Antigen, tPSA)、游离 PSA(free Prostate Specific Antigen, fPSA)、游离 PSA 百分比((fPSA/tPSA, f/tPSA)、PSA 密度、前列腺体积、碱性磷酸酶(Alkaline Phosphatase, ALP)、血糖、血脂、血压、体重、饮酒、吸烟、核磁共振检查(Magnetic Resonance Imaging, MRI)、尿急尿痛、排尿困难、夜尿频次、血尿共 18 个相关变量用于变量的筛选及建模。部分变量的计算方法及介绍如下:

游离 PSA 百分比(f/tPSA):单纯的 tPSA 升高对前列腺癌的诊断特异性不高。当 tPSA 介于 4-10ng/ml 之间时,因患者的 tPSA 仅轻度升高而加大了诊断的难度;当 f/tPSA<0.16 时,则患者前列腺癌风险增加[2]。

$$f/tPSA = \frac{fPSA}{tPSA} \times 100\% \quad (1)$$

PSA 密度(Prostate Specific Antigen Density, PSAD)表示单位体积内前列腺的 PSA 含量。

$$PSAD = \frac{tPSA}{PV} \quad (2)$$

前列腺体积(Prostate Volume, PV)表示前列腺增生的情况。本文前列腺的左右、前后、上下径由 MRI 测得。

PV = 0.52×左右径(cm)×上下径(cm)×前后径(cm) (3)

碱性磷酸酶(Alkaline Phosphatase, ALP)是广泛分布于人体肝脏、骨骼、肠、肾和胎盘等组织经肝脏向胆外排出的一种酶,临床上测定 ALP 主要用于骨骼疾病的诊断和鉴别诊断,ALP 水平的升高与恶性肿瘤的骨转移相关[7]。

MRI:前列腺核磁共振检查已成为诊断前列腺癌的常规手段,不仅能够发现直肠指诊难以发现的占位性病变,而且具有一定的特异性。由于前列腺癌以前列腺外周带多发[8],因此当磁共振检测出外周带结节时,应警惕前列腺癌的发生。

本文共纳入样本 255 例,其中穿刺结果为前列腺癌患者 85 例,前列腺增生患者 170 例。

1.2 数据预处理

1.2.1 缺失值处理

绝大部分患者的临床信息能够完整收集,但仍有少部分患者的信息是缺失的。我们在纳入数据时,将缺失值大于 10% 的患者排除,在纳入的 255 例患者的缺失数据均小于 10%,使用 SPSS 22.0 中均值填充(序列均值)缺失值的方法,补全所有患者信息。

1.2.2　预测变量筛选

在纳入的 18 个变量中,首先对各个变量的分布情况进行分析,然后采用传统的变量筛选方法即单变量方差分析,筛选出有统计学意义的变量,再对变量进行 Logistic 多元回归分析。进行单变量 t 检验/卡方检验,采用 Weka 软件将所有数值型变量转为标称型变量,并计算各个变量信息增益率,将单变量分析与信息增益率相结合进行变量筛选。

本文采用上述两种方法进行变量筛选,选用最优变量建立前列腺癌诊断预测模型。

1.3　结局变量

纳入的所有患者均行 12 针系统穿刺活检术,以术后病理结果作为患者诊断的"金标准",若病理结果显示为前列腺癌(恶性)则为阳性样本,穿刺结果为前列腺增生(良性)则为阴性样本。本文共纳入 255 例样本,其中阳性样本 85 例,阴性样本 170 例,用于建立模型。

2　构建前列腺癌诊断预测模型

2.1　构造样本集

2.1.1　训练集和测试集

考虑到样本量的限制,本文不再按比例单独划分训练集和测试集,而是采用十折交叉验证方法(10-fold cross-validation)[9]建立模型。所谓十折交叉验证就是每次将数据随机分成 10 份,其中 9 份作为训练集,将余下的 1 份作为测试集。该过程重复进行 10 次,可以有效提高模型的稳定性和泛化能力,防止"过拟合"现象的出现。

2.1.2　验证集

为了更好地评价模型性能,本文引入验证集对模型进行验证。验证集共包含 75 例样本,为 2019 年 1-6 月在中国医科大学附属盛京医院泌尿外科行超声引导下前列腺穿刺活检术的患者(变量纳入和排除标准同上),其中阳性样本 26 例,阴性样本 49 例。

2.2　模型构建方法

2.2.1　集成学习

集成学习(Ensemble Learning)[10]并不是一种单独的机器学习算法,而是将多个单一的分类器组合在一起,使它们共同完成学习任务,可以有效提高基分类器的泛化能力并解决过拟合问题。常见的集成学习方法有 Bagging, Boosting 和 Stacking,本文主要采用 Bagging 方法。

Bagging(Bootstrap aggregating)[11]采用自助采样法(Bootstrap)进行多轮有放回抽样,每轮从原始样本集中抽取 n 个训练样本,共进行 k 轮抽取,得到 k 个训练集;每次使用 1 个训练集进行建模,共得到 k 个模型,每个模型的重要性是相同的。对于分类问题,将 k 个模型结果采用投票的方式获得最终分类结果;对于回归问题,则计算上述 k 个模型的均值作为最后的结果。

2.2.2　朴素贝叶斯

朴素贝叶斯法(Naive Bayes)[12]是基于贝叶斯定理与特征条件独立假设的分类方法。对于一个已知分类的待分类集合(训练样本集)$x=\{a_1,a_2,\cdots,a_m\}$和有类别集合 $C=\{y_1,y_2,\cdots,y_n\}$,统计各类别下各个特征属性的条件概率估计公式为:

$$P=a_{1\cdots m}\mid y_{1\cdots n} \tag{4}$$

因各个特征属性是相互独立的,故得到最终的朴素贝叶斯公式为:

$$P(x_{yi})P(y_i)=P(a_1\mid y_i)P(a_2\mid y_i)\cdots P(a_m\mid y_i)P(y_i)=P(y_i)\prod_{j=1}^{m}P(a_j\mid y_i) \tag{5}$$

朴素贝叶斯对小规模的数据表现良好,对缺失数据不太敏感,算法比较简单,用于本文的数据较为合适。

3　结果与分析

3.1　变量筛选结果及分析

对训练集和验证集的 255 例样本进行单因素分析,可知有 10 个变量有显著性意义。变量分布情况及显著性见表 1。其中前列腺癌组与前列腺增生组的年龄、体重、tPSA、fPSA 、f/tPSA、PSAD、PV、ALP、夜尿频次及 MRI 检查均存在统计学差异。

3.1.1　多因素 Logistic 分析

将上述 10 个指标进行多因素 Logistic 回归分析,经筛选后,年龄、tPSA、游离 PSA 百分比、前列腺体积、体重 5 个指标被纳入(表 2)。其中年龄、tPSA 和体重都是危险因素,tPSA 每提高一个单位水平,患前列腺癌的风险提高 1.067 倍。各个指标的 ROC 曲线如图 1 所示。其中 tPSA 在所有指标中最有诊断意义,游离 PSA 百分比次之。

表 1 前列腺癌与前列腺增生变量分布及差异性比较

变量	前列腺癌($\bar{x}\pm s$)	前列腺增生($\bar{x}\pm s$)	t/χ^2	P
年龄(岁)	69.71±8.50	66.81±8.33	127.140	<0.001
tPSA(ng/ml)	60.25±37.89	16.41±14.15	15.339	<0.001
fPSA(ng/ml)	14.38±16.32	2.50±2.06	9.829	<0.001
游离 PSA 百分比	0.23±0.37	0.15±0.07	12.439	<0.001
PSAD(ng/(mLocm^3))	0.66±0.80	0.29±0.30	11.963	<0.001
PV(cm^3)	52.76±26.61	65.96±39.49	27.061	<0.001
体重(KG)	69.72±9.51	69.99±10.87	106.729	<0.001
夜尿(次)	≤3:19	≤3:59		
	4-5:31	4-5:71	8.726	0.003
	>5:35	>5:40		
ALP	升高:10	升高:3	11.882	0.001
	正常:75	正常:167		
MRI	有结节:77	有结节:93	42.966	<0.001
	无结节:8	无结节:77		

注:t/χ^2 为 t 检验和卡方检验对应的 t 值或 χ^2 值

表 2 多因素 Logistic 分析结果

变量	B	Wald	P	Exp(B)
年龄(岁)	0.084	6.527	0.011	1.088
tPSA(ng/ml)	0.065	20.829	0	1.067
游离 PSA 百分比	3.435	1.172	0.279	31.037
前列腺体积(cm^3)	-0.033	10.809	0.001	0.967
体重(KG)	0.06	5.234	0.022	1.062
常数	-11.393	11.288	0.001	0

注:B 表示系数;Wald 为检验统计量,检验自变量对因变量是否有影响;Exp(B)代表 OR 值

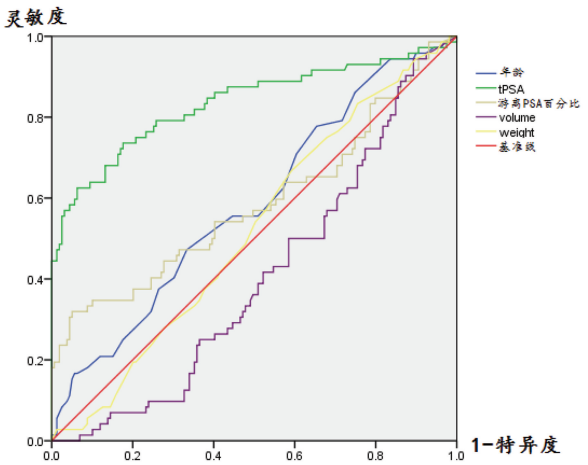

图 1 Logistic 回归各指标 ROC 曲线

3.1.2 信息增益率分析

使用 Weka 中变量选择模块计算各个变量信息增益率,属性评估器(Attribute Evaluator)选择 InfoGainAttributeEval,查找算法(Search Method)选择 Ranker。各变量信息增益率如表 3 所示。其中数值型变量无需变量划分。综合显著性分析和变量重要性分析,将 P<0.05 且信息增益率>0.02 作为纳入标准,年龄、tPSA、fPSA、游离 PSA 百分比、PSAD、前列腺体积、体重、夜尿频次、ALP 及 MRI10 个变量被纳入,而吸烟、饮酒、排尿困难、血尿、血脂、血糖、血压、尿急尿痛等 8 个变量被排除。

表 3 变量类型、划分及重要性排序

数据类型	变量名称	变量划分	信息增益率
数值型	年龄		0.167
	tPSA		0.827
	fPSA		0.627
	游离 PSA 百分比		0.223
	前列腺体积		0.705
	PSAD		0.237
	体重		0.093
分类型	ALP	正常/升高	0.030
	MRI	无/有	0.094
	夜尿频次	≤3 次	
		4-5 次	0.026
		>5 次	

3.2　模型结果分析

为进一步提高模型性能，采用十折交叉验证划分训练集和测试集，运用集成学习的方法选取随机森林、Bagging 集成朴素贝叶斯、支持向量机（Support Vector Machine，SVM）及 Logistic 等基分类器，构建前列腺癌诊断模型。

为验证不同算法及不同变量构建的模型性能，采用 Precision、Recall、F 值及 AUC 共[13]4 个指标对诊断预测模型进行评价与比较的结果，如表 4 所示。

由表 4 可知，4 种算法构建的模型性能因变量筛选方式不同略有差异。在使用信息增益率筛选方式建立的模型中，Naive Bayes 模型 AUC 最高，为 0.826；RF 的 Precision 值最大，达到 0.839；在应用 Logistic 筛选变量建模中，RF 的 AUC 和 Precision 均为最高，分别是 0.743 和 0.823；4 种算法在应用信息增益率筛选变量建立的模型性能均优于应用 Logistic筛选变量的模型性能。

表 4　2 种变量筛选方式、4 种集成学习模型结果比较

筛选方式	信息增益率				Logistic			
算法	AUC	Precision	Recall	F 值	AUC	Precision	Recall	F 值
RF	0.745	0.839	0.800	0.772	0.743	0.823	0.796	0.770
Naive Bayes	0.826	0.784	0.769	0.773	0.698	0.655	0.639	0.645
Logistic	0.740	0.706	0.702	0.704	0.672	0.647	0.612	0.622
SVM	0.761	0.789	0.788	0.773	0.703	0.743	0.749	0.727

3.3　模型验证

应用信息增益率筛选的变量构建的模型性能更佳。为进一步验证模型性能，将模型应用于临床研究，采用相同的纳入和排除标准收集 75 例患者（阳性 26 例，阴性 49 例）作为验证集，采用 4 种算法应用相同的参数进行模型性能评估，结果见表 5。其中 Naive Bayes 算法的 AUC 值最大（AUC = 0.797，Precision = 0.764），RF 的 Precision 最高，而 AUC 值最低（Precision = 0.791，AUC = 0.610）。两种算法的 ROC 曲线及混淆矩阵分别如图 2 和表 6 所示，图 2 分别表示 Naive Bayes 算法 ROC 曲线和 RF 算法 ROC曲线。由表6可知，RF算法对于阴性的预测更准确，49 例阴性样本全部预测正确，可避免不必要的穿刺活检；而 Naive Bayes 算法对于阳性样本的预测效果较好，26 例中有 21 例预测正确，准确率达 80.7%。故在将模型应用于临床时，应该综合考虑多个模型的结果，以达到最好的术前诊断效果。

表 5　4 种算法验证集模型性能

算法	AUC	Precision	Recall	F 值
RF	0.610	0.791	0.693	0.601
Naive Bayes	0.797	0.764	0.720	0.727
Logistic	0.766	0.744	0.720	0.726
SVM	0.766	0.730	0.733	0.708

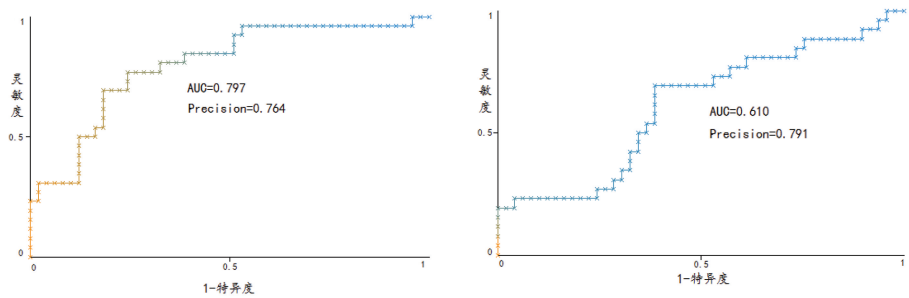

图 2　两种算法 ROC 曲线

表 6 两种算法混淆矩阵

病理结果	预测结局(Naive Bayes)		预测结局(RF)	
	阳性	阴性	阳性	阴性
阳性	21	5	22	4
阴性	16	33	0	49

4 讨论

近年来,我国的前列腺癌发病率逐年升高。在综合考虑患者的基本信息、症状、体征、化验及检查结果后,对穿刺结局进行准确预测能够有效减轻患者身体及经济上的负担。面对庞大的临床信息量,变量筛选方式的选择是研究面临的主要问题之一。本文发现信息增益率筛选出的变量较 Logistics 更为科学合理。Logistics 分析显示,年龄、tPSA 是前列腺的相关危险因素,而前列腺体积则为前列腺癌的保护因素,此结论与近些年的研究结果一致[14]。但 Logistic 分析舍弃了许多有价值的变量,PSA 密度[15]、MRI 检查等重要的参考指标并没有被纳入,容易造成临床医生对患者重要信息的忽视。

信息增益率筛选方式,不仅能够对不同变量的重要程度进行排序,而且能够根据实际情况设计阈值,使实验结果更加贴近临床。PSA 相关指标是重要性最高的几种变量,应在诊断时优先考虑;前列腺体积、体重、MRI 检查在重要性方面次之。虽然夜尿频次、ALP 水平仅对诊断的参考价值较小,但仍然不容忽视。其中,夜尿频次增加是前列腺癌患者的早期临床表现之一,ALP 升高为存在骨转移的重要指标,因此可以间接反应患者是否存在前列腺癌的风险。虽然高血压、高血脂以及糖尿病等代谢综合征的存在会增加前列腺癌风险,但本文中未见统计学差异,有待进一步进行更大样本量的研究。

本文对 Losgistic 多因素分析与机器学习算法的横向对比,证明机器学习算法具有较准确的预测效果。不同机器学习算法间的纵向对比发现,虽然不同算法之间均具有良好的效果,但以 ROC 曲线下面积为标准。朴素贝叶斯的预测效果最好,而以基于精准率与召回率的 F 值为标准,则随机森林效果最佳。除了进行更加全面的对比之外,还对建立的模型进行了临床验证,以较为准确的朴素贝叶斯算法及随机森林算法为例,结果证明两种模型均具有良好的临床应用潜能但随机森林的预测结果更加准确。但是,本文仍存在以下不足:患者临床信息的缺失,未考虑存在患者穿刺结果为假阴性可能、样本量较小等。综上所述,机器学习算法在前列腺癌的诊断中具备较高的准确率,但其临床应用尚待进一步研究。

【参考文献】

[1] Siegel R L, Miller K D, Jemal A, Cancer Statistics, 2017. [J]. CA Cancer J Clin, 2017, 67(1): 7-30.

[2] 黄桂海,李 伟. 前列腺特异性抗原相关参数在前列腺穿刺中的应用及其研究进展[J]. 现代泌尿外科杂志, 2019, 24(1): 72-76.

[3] Cao X L, Gao J P, Han G, *et al*. Relationship between screening by stratifying cases into groups on prostate specific antigen level and the positive rate of transrectal ultrasound guided systematic sextant prostate biopsy[J]. Chinese Journal of Surgery, 2006, 44(6): 372-375.

[4] 彭 涛,肖建明,张仕慧,等. 基于多参数 MRI 及影像组学建立机器学习模型诊断临床显著性前列腺癌[J]. 中国医学影像技术, 2019, 35(10): 1526-1530.

[5] 崔少泽,王杜娟,王苏桐,等. 基于 GMM-RBF 神经网络的前列腺癌诊断方法[J]. 管理科学, 2018, 31(1): 33-47.

[6] 曹文哲,应 俊,张亚慧,等. 基于机器学习算法的前列腺癌诊断模型研究[J]. 中国医疗设备, 2016, 31(4): 30-35.

[7] Heinrich D, Bruland ϕØ, Guise T A, *et al*. Alkaline phosphatase in metastatic castration-resistant prostate cancer: reassessment of an older biomarker. [J]. Future Oncol, 2018, 14(24): 2543-2556.

[8] Jung S I, Jeon H J, Park H S, *et al*. Multiparametric MR imaging of peripheral zone prostate cancer: effect of postbiopsy haemorrhage on cancer detection according to Gleason score and tumour volume [J]. The British Journal of Radiology, 2018, 91(1086): 20180001.

[9] Zhang H, Yang S, Guo L, *et al*. Comparisons of isomiR patterns and classification performance using the rank-based MANOVA and 10-fold cross-validation[J]. Gene, 2015, 569(1): 21-26.

[10] 于 玲,吴铁军. 集成学习:Boosting 算法综述[J]. 模式识别与人工智能, 2004(1): 54-61.

[11] Truong C, Zhang M, Andreae P, *et al*. Bagging and Feature Selection for Classification with Incomplete Data[C]// Evostar. Berlin: Springer International Publishing, 2017: 471-486.

[12] Wolfson J, Bandyopadhyay S, Elidrisi M, *et al*. A Naive Bayes machine learning approach to risk prediction using censored, time-to-event data[J]. Statistics in Medicine, 2015, 34(21): 2941-2957.

[13] 杜 超,范馨月,单立平. 利用随机森林建立输尿管上段结石预后预测模型[J]. 中华医学图书情报杂志, 2019, 28(5): 15-19.

[14] Rao E V, Sridhar P, Rao B V L N, *et al*. Prostate cancer detection: Relationship to prostate size[J]. Urology, 1999, 53(4): 764-768.

[15] Nordström T, Akre O, Aly M, *et al*. Prostate-specific antigen (PSA) density in the diagnostic algorithm of prostate cancer [J]. Prostate cancer and prostatic diseases, 2017, 21(1): 57-63.

[收稿日期:2019-10-25]

[本文编辑:吴方怡]