大数据背景下医院门诊挂号预约爽约行为预测研究

朱　光[1],邓弘林[2]

(1.沧州市中心医院宣传策划部,河北　沧州　061000;
2.中山大学管理学院,广东　广州　510000)

摘要:目的　了解目前医院预约诊疗服务中患者爽约行为的现状,探讨和鉴别患者爽约的关键特征,运用这些特征建立机器学习算法模型预测未来患者爽约行为。**方法**　挖掘 2018 年河北省某大型三甲医院预约大数据,首先用 Stata 采取传统 Logistic 回归找出患者爽约的显著因子,再将数据划分为训练集和预测集,采用 SVM、决策树、随机森林和 BP 神经网络等不同模型学习训练患者爽约行为和特征,检验每种算法对患者爽约预测的准确率。**结果**　目前医院患者预约爽约率为 16.16%,Logistic 回归分析显示年龄、性别、预约时间和预约科室是爽约行为的关键性特征;使用这些特征进行机器学习和预测能取得较好效果,SVM、决策树、随机森林和 BP 神经网络各个算法准确率均超过 75%,其中 SVM 和 BP 神经网络准确率最高,是该特定情境下的最优算法。**结论**　我国大型三甲医院预约诊疗服务有待进一步加强,在大数据时代的背景下,机器学习方法可为医院预测并降低爽约率提供强有力支持。

关键词:预约诊疗;爽约率;机器学习;大数据

中图分类号:R197.3　　**文献标识码:**B　　DOI:10.3969/j.issn.1006-1959.2020.22.004

文章编号:1006-1959(2020)22-0013-04

An Investigation of Predicting Patient Missing Appointment Behavior Under the Big Data Background

ZHU Guang[1],DENG Hong-lin[2]

(1.Department of Publicity and Planning,Cangzhou Central Hospital, Cangzhou 061000,Hebei,China;
2.School of Business,Sun Yat-sen University,Guangzhou 510000,Guangdong,China)

Abstract:Objectives　To understand the current situation of patients'missing-appointment behavior in the appointment service of hospitals; to explore and identify the key features of patients'missing appointment.Use these features to build a machine learning algorithm model to predict future patient missing-appointment behavior. **Methods**　Mining the big data of appointments in a large tertiary hospital in Hebei Province in 2018. First, Stata adopts traditional Logistic regression to find the significant factors of patients' appointments, and then divides the data into training sets and prediction sets, using SVM, decision tree, random forest and BP Different models, such as neural networks, learn and train patients' absentee behavior and characteristics, and test the accuracy of each algorithm in predicting patient absenteeism. **Results**　The current appointment rate of hospital patients is 16.16%. Logistic regression analysis shows that age, gender, appointment time and appointment department are the key features of appointment cancellation behavior; using these features for machine learning and prediction can achieve better results, SVM, decision tree accuracy of each algorithm of random forest and BP neural network exceeds 75%. Among them, SVM and BP neural network have the highest accuracy, which is the best algorithm in this specific situation. **Conclusion**　The appointment diagnosis and treatment services of my country's large tertiary hospitals need to be further strengthened. In the context of the era of big data, machine learning methods can provide strong support for hospitals to predict and reduce the rate of missing-appointment.

Key words: Appointment of diagnosis and treatment;Missing-appointmentrate;Machine learning;Big data

国家提出的互联网+等战略，表明以互联网和大数据为核心的科技和社会变革已形成了推动国家医疗发展的新浪潮。在此环境下,全国范围内推广电子渠道预约挂号(如微信平台)的医院越来越多。然而,许多医院反映预约挂号存在着爽约率较高的局面,造成了医疗资源的浪费,更在一定程度上扰乱了医院的诊疗秩序[1]。因此,如果可以从患者预约时输入的多维信息(如性别、年龄、预约时间、预约科室等多个变量)鉴别其中的关键因素,并利用这些因素对该预约的爽约概率进行预测,可大大提高预约诊疗服务的管理效率[2]。以往研究大多从现象表面出发,以单一维度对爽约行为进行定性探讨,无法准确地预测每个预约的爽约概率。本研究从大数据驱动的角度出发,构建不同的机器学习算法模型对爽约行为进行预测和识别，通过比较预测准确率选择最优算法,为医疗资源配置优化提供可行性建议。

1 资料与方法

1.1 资料来源　本研究选取华北地区某大型三甲医院 2018 年全年通过电子渠道(包括微信平台、医院官网)共 94651 例预约。

1.2 研究变量和测量工具　本研究的因变量为二元变量,即患者赴约(标记为 1)或爽约(标记为 0),自变量包括患者年龄、性别(类别变量,男性=1,女性=0)、预约就诊时间与订单时间差、预约科室(类别变量,共 41 个科室类别)、医生预约名额上限(即预约医生每天可供预约名额)。

1.3 机器学习算法　采用 Stata 进行逻辑回归分析,考察和筛选患者爽约行为的关键性特征。然后运用 Python 中的 sklearn 库对筛选后的关键性特征进行建模、分析和预测。运用支持向量机(SVM,RBF 核函数)、决策树(C4.5 算法)、随机森林和 BP 神经网络机器学习模型对患者爽约行为的特征进行学习和

作者简介:朱光(1991.10-),女,黑龙江大庆人,本科,经济师,主要从事医院人力资源、医疗管理及市场营销

通讯作者:邓弘林(1987.7-),男,广东湛江人,博士,助理教授,主要从事电子商务、医疗大数据分析及机器学习研究

预测，比较不同算法的准确率，探讨当前情境下的最优算法[3]。

在大数据分析的算法中，SVM 是目前最为广泛应用的为二进制分类而设计的算法。本文的实证场景为患者“是”或“否”爽约，符合 SVM 的研究情境。利用核函数（最常用的为 RBF 核函数）机制构造一个最优的超平面，从而使负数据集和正数据集之间的间隔最大 [4]。

决策树是（DT）一个有监督分类与回归算法，其中每个内部节点表示一个属性上的判断，每个分支代表一个判断结果的输出，最后每个叶节点代表一种分类结果[5]。在机器学习里的决策树主要优点是在克服传统方法的缺点的同时，利用逻辑模型对数据进行分类，具有更高的精度。最常见的决策树类型为 C4.5 算法（以信息增益率为分枝方式）。然而，决策树容易过度拟合，此时随机森林（random forest）很好地缓解了这个问题。随机森林是决策树的集合，其结果被聚合为一个最终结果。随机森林算法能限制过拟合的问题并且不会因为偏差而大大增加误差。

BP 神经网络是一种按照误差逆向传播算法训练的多层前馈神经网络，是目前应用最广泛的神经网络，其主要利用链式规则的梯度来优化算法，特点是其迭代、递归和有效的计算权值更新的方法，以改进网络，直到能够执行训练任务为止[6]。

1.4 统计学方法 采用 Stata 进行逻辑回归分析，考察和筛选患者爽约行为的关键性特征。然后运用 Python 中的 sklearn 库对筛选后的关键性特征进行建模、分析和预测。

2 结果

2.1 患者预约爽约行为分析 进行预约的 94,651 例患者的平均年龄为 37.04 岁，以女性较多，通常患者提前一天半进行预约，该医院共有 41 个可供预约的科室类别，每个医生平均可接受预约的名额约为 23 个，较为充足，在预约实例中，爽约率达到 16.16%，有 15,300 例，见表 1。

2.2 患者爽约行为的关键性特征分析 构建方程：Logit（患者是否爽约）=α + β_1 年龄+ β_2 性别+ β_3 预约就诊时间与订单时间差+ β_4 预约科室+ β_5 医生预约名额上限 + μiLogistic 回归模型用以预测事件发生或不发生概率。预测值最大时趋向 1，最小时趋向 0，即如果通过模型计算出来的概率大于 0.5，则预测该患者会爽约。在上式中，β_i（i = 1, 2, 3, 4, 5）为自变量的相关系数，α 为常数项，μ 为残差。通过 Stata 软件进行 Logistic 回归分析的结果显示，①患者的爽约行为与年龄呈正相关，患者年龄每增加 1 岁，其爽约的可能性便上升约 0.22%；②患者的爽约行为与性别显著相关，其中女性更容易爽约；③患者的爽约行为与预约时间显著负相关，越提早预约的患者越不容易爽约；④不同科室的爽约概率也不同，其中外科的爽约率最高，达到 55.56%，皮肤科的爽约率也超过 33%，爽约率最低的为产科，约 11.28%；⑤医生的预约名额与患者是否爽约没有显著关系，见表 2。

表 1 患者预约爽约行为的描述性分析[*n*（%）]

变量		*n*	平均值	最小值	最大值
年龄		94651	37.04	1	117
性别	男	29233	/	/	/
	女	65418	/	/	/
预约就诊时间与当前时间差（d）		94651	1.43	0	29
预约科室		94651	/	/	/
医生预约名额上限		94651	22.90	5	85
是否爽约	赴约	79351	/	/	/
	爽约	15300	/	/	/

表 2 患者爽约行为 Logistic 回归统计结果

是否爽约	Coef.	*OR*	Std. Err.	*Z*	*P*>|z|	95% *CI*
年龄	0.002192	1.002195	0.000619	3.550	0.0000	1.000983~1.003408
性别	-0.05837	0.943302	0.01984	-2.780	0.0060	0.905208~0.983000
预约就诊时间与当前时间差	-0.11423	0.892055	0.004044	-25.200	0.0000	0.884165~0.900016
预约科室	-0.67947	0.54465	0.06912	-6.332	0.0320	0.436488~0.729048
医生预约名额上限	0.001216	1.001217	0.002198	0.550	0.5800	0.996917~1.005535
常数项	2.157676	8.65101	0.581891	32.080	0.0000	7.582503~9.870088

2.3 基于机器学习的患者爽约行为预测　针对数据驱动的患者爽约行为预测，本研究采用 SVM、C4.5 决策树、随机森林和 BP 神经网络对数据进行分析。主要使用 Python 语言的进行建模、训练和预测。首先指定机器将 118627 条数据划分为训练集和测试集，其中训练集占 70%原始数据。在 SVM 分析中，首先验证 RBF 核函数下的预测准确率，gamma 值设置为 1/4（即 0.25），惩罚系数 C 设置为 1。在决策树模型中采用 C4.5 算法，不指定最大深度和最大子叶节点以提高准确率，而随机森林算法的最大迭代次数设置为 100。BP 神经网络模型则设置 3 层隐藏层，每层 50 个神经元，即 hidden_layer_sizes=(50,50,50)，最大迭代次数为 200 次。各项测试结果显示，机器学习的方法能获得较高的预测准确率（均超过 70%）；在各种大数据建模分析方法中，最适合预测患者爽约行为的算法是 BP 神经网络算法和采用 RBF 核函数的 SVM 算法，预测患者是否爽约的准确率均达到 83.80%，其次是随机森林算法，准确率为 79.80%。见表 3。

表 3　患者爽约行为机器学习分析和预测结果

项目	Precision	Recall	f1-score	Support
SVM（RBF 核函数）：SVC(kernel='rbf', gamma=0.25, decision_function_shape='ovo', C=1)				
0（爽约）	0.5657	0.0121	0.4677	4612
1（赴约）	0.8390	0.9982	0.9117	23784
预测准确率	83.80%	83.80%		34684
决策树（C4.5）：DTC(criterion='entropy',random_state=0)				
0（爽约）	0.1996	0.2147	0.2068	4612
1（赴约）	0.8454	0.8330	0.8392	23784
预测准确率	73.26%	73.26%		28396
随机森林：RFC(n_estimators=100, random_state=10)				
0（爽约）	0.2397	0.1121	0.1528	4612
1（赴约）	0.8439	0.9310	0.8854	23784
预测准确率	79.80%	79.80%		28396
BP 神经网络：MLPClassifier(random_state=0, hidden_layer_sizes=(50, 50, 50), max_iter=200)				
0（爽约）	0.5821	0.0085	0.0167	4612
1（赴约）	0.8386	0.9988	0.9117	23784
预测准确率	83.80%	83.80%		28396

3 讨论

预约诊疗能合理、有效、公平地分配和利用医疗资源。对改进患者就医秩序，缩短患者等待时间，提高诊疗效率发挥了重要作用[7]。因此我国各医疗机构充分发挥多种手段和渠道开展预约诊疗。本文的实证研究和分析有以下三点发现。①患者预约爽约率总体上比较高，预约就诊服务仍需改善：预约就诊服务能为医院管理高效地、有计划地分配资源，当中最常见也最难以解决的问题就是患者爽约，过高的爽约率使得预约就诊服务失去其促进资源分配的意义[8]。本研究发现，目前医院一年内预约次数已达将近 10 万人次，大部分科室也已经开始推广预约就诊服务，但是预约爽约率较为偏高，与以往研究的爽约率相仿，证明就降低患者爽约行为而言，目前预约就诊服务尚未得到明显改善。因此，本文的研究，先从患者预约的多维信息中检测其中的关键因素，再通过机器学习预测患者是否爽约的策略，可为医院改善预约诊疗服务提供思路。②患者的社会人口学特征、预约时间和预约科室对爽约行为的影响较大：本研究结果表明，决定患者是否爽约的关键因素包括了患者本身的特征、提前多久预约以及预约的科室。其中，在预约的患者当中，女性患者占了大多数，不过，女性患者相比男性患者而言也更容易出现爽约行为，同时高龄患者也比年轻患者爽约的概率高，而越早进行预约的患者更不容易爽约，提早预约说明患者对该次诊疗更为重视，因此赴约的可能性更高。科室之间的爽约率也相差甚远，这可通过科室诊疗特点进行解释[9]。例如，妇产科的患者爽约的比例最低，主要因为妇产科患者的复诊率较高而且有相对固定医生，接诊医生会给予明确的复诊时间，因此患者通常会依照医生指引预约特定的时间复诊。③在大数据背景下通过机器学习预测患者爽约行为：尽管大数据机器学习的方法越来越流行，但能否以及如何应用到医院预约诊疗服务管理尚未有定论，本文立足于预测患者爽约行为的特定情境，发现总体而言机器学习的方法在大数据环境下能有效地预测患者该次预约爽约的可能性（多个算法的预测准确

（下转第 21 页）

(上接第 15 页)

率均超过 70%),然而同时,本文发现了选择合适的算法才是最关键的,不同算法得出的准预测确率也有区别。

4 对策建议

各医疗机构要高度重视预约爽约问题，根据本研究，提出改进和提高医院预约诊疗服务管理具体建议：①将日常预警和跟踪作为预约诊疗服务的中心工作。可根据爽约率高的人群进行 VIP 管理,如事前提醒,事后服务跟踪,就医过程无障碍化等。②需对预约诊疗的人群进行现场满意度测评和定期回访，根据患者的意见和建议改善医疗服务。③对爽约率较高的科室和人群进行个案分析,找出具体原因,提出对应策略,提高患者的认同感。④患者就医过程是对医院多项服务的综合体验,医院需根据跟踪的意见,改善全院服务,因此研究患者爽约行为也是发现医院服务缺陷的一个重要途径。

总之,患者爽约率高是目前医院推行预约诊疗服务的亟待解决的一个瓶颈,深入研究患者爽约行为的特征和要素,结合大数据时代下的先进技术,从而采取相应的管理措施降低爽约率,提高管理质量,是进一步推行预约诊疗服务和提高医院工作效率的关键所在。

参考文献：

[1]喻锎.探索精准预约服务建立通畅就医流程--武汉市第一医院预约诊疗服务实践和探索 [J]. 中国医院管理,2019,459(10):2-3.

[2]Kogan S,Moskowitz TJ,Niessner M.Fake News:Evidence from Financial Markets[D].SSRN Electronic Journal,2018.

[3]黄洛.医院门诊预约挂号爽约的现状及对策[J].现代医院,2019,19(4):63-66.

[4]陈默,蔡苗,黄阿红,等.基于 K-means 聚类与支持向量机的大病患者住院费用影响因素与控制策略研究[J].中国医院管理,2019,39(5):45-47.

[5]吴越,徐丛剑,程子桐,等.二值响应模型与决策树在门诊失约行为研究中的应用[J].中国医院管理,2018,38(10):36-38.

[6]易焱琪,鞠水,家晓艳,等.浅析 BP 神经网络技术在医院信息系统中的应用[J].科技创新与生产力,2017(2):107-109.

[7]周奇.医院门诊预约系统的优化分析:基于国内某大型医院的研究[D].中国科学技术大学,2017.

[8]周萍,冯笑,赵岭,等.医院预约挂号爽约现象的调查分析[J].中医药管理杂志,2018,26(18):29-31.

[9]刘玉琦,郝晓刚,马亚飞.某三级医院预约挂号爽约情况及其原因调查[J].武警医学,2018,29(2):117-119.

收稿日期:2020-07-10;修回日期:2020-08-19
编辑/成森