DOI:10.12138/j.issn.1671－9638.20193918

·论著·

基于机器学习的 CatBoost 模型在预测重症手足口病中的应用

王　斌[1],冯慧芬[1],王　芳[2],秦新华[2],黄　平[1],党德建[3],赵　敬[1],易佳音[1]
(1. 郑州大学第五附属医院消化内科,河南 郑州　450052;2. 郑州大学附属儿童医院感染科,河南 郑州　450051;3. 郑州大学第五附属医院感染控制科,河南 郑州　450052)

[摘　要] **目的**　通过机器学习算法,探究 CatBoost 模型在预测重症手足口病(HFMD)中的应用价值。**方法**　收集郑州市某医院 2014 年 1 月—2017 年 6 月住院部诊治的 2 983 例 HFMD 患儿,使用 R 3.4.3 软件进行数据分析,构建 CatBoost 模型和其他普通模型,评估 CatBoost 模型的预测性能。**结果**　最终构建的 CatBoost 模型,预测正确率可达 87.6%,人工神经网络模型位居第二(83.8%),其他(决策树、支持向量机、logistic 回归、贝叶斯网络)模型预测正确率＜80%。CatBoost 算法模型 ROC 曲线下面积、灵敏度、特异度均高(分别为 0.866、80.80%、92.33%),其中居前 3 位的预测变量依次为呕吐、肢体抖动和病原学结果。**结论**　CatBoost 模型可以用于预测重症 HFMD,相比于其他传统算法,具有较高的预测正确率和诊断价值。

[关　键　词] 手足口病;重症手足口病;机器学习;CatBoost;预测
[中图分类号] R512.5

Application of CatBoost model based on machine learning in predicting severe hand-foot-mouth disease

WANG Bin[1], *FENG Hui-fen*[1], *WANG Fang*[2], *QIN Xin-hua*[2], *HUANG Ping*[1], *DANG De-jian*[3], *ZHAO Jing*[1], *YI Jia-yin*[1] (1. *Department of Gastroenterology, The Fifth Affiliated Hospital of Zhengzhou University, Zhengzhou* 450052, *China*; 2. *Department of Infectious Disease, Children's Hospital Affiliated to Zhengzhou University, Zhengzhou* 450051, *China*; 3. *Department of Healthcare-associated Infection Control, The Fifth Affiliated Hospital of Zhengzhou University, Zhengzhou* 450052, *China*)

[**Abstract**] **Objective**　To explore the value of CatBoost model in predicting severe hand-foot-mouth disease (HFMD) by the machine learning algorithm. **Methods**　A total of 2 983 children with HFMD diagnosed and treated in a hospital in Zhengzhou from January 2014 to June 2017 were collected, data were analyzed with R 3.4.3 software, CatBoost model and other common models were constructed, prediction performance of CatBoost model was evaluated. **Results**　The predictive accuracy of the finally constructed CatBoost model was 87.6%, artificial neural network model ranked second (83.8%), other models (decision tree, support vector machine, logistic regression, Bayesian network) had predictive accuracy less than 80%. The area under receiver operating characteristic (ROC) curve, sensitivity, and specificity of CatBoost algorithm model were all high (0.866, 80.80% and 92.33% respectively), the top three predictive variables were vomiting, limb jitter, and pathogenic results. **Conclusion**　CatBoost model can be used to predict severe HFMD, which has higher accuracy and diagnostic value than other traditional algorithms.

[**Key words**] hand-foot-mouth disease; severe hand-foot-mouth disease; machine learning; CatBoost; prediction

[收稿日期]　2018－05－21
[基金项目]　国家自然科学基金(81473030);河南省医学科技攻关普通项目(201403130);河南省卫生系统出国研修项目(2015065)
[作者简介]　王斌(1993－),男(汉族),甘肃省定西市人,硕士研究生,主要从事消化、感染及传染病研究。
[通信作者]　冯慧芬　　E-mail:huifen.feng@163.com

手足口病(hand-foot-mouth disease, HFMD)是由肠道病毒引起的一种常见的儿童传染病[1],以肠道病毒 71 型(EV-A71)和柯萨奇 A 组 16 型(CV-A16)感染多见。近年报道的 CV-A6 和 CV-A10 亚型也逐渐上升为重要病原体,与 HFMD 的散发和世界范围大流行相关,导致发生神经系统并发症和死亡的患儿数目增加[2]。CV-A6 HFMD 于 2008 年首次在欧洲芬兰暴发,现大多数集中在亚太地区,如中国、印度、新加坡、日本等[3-4]。该病具有自限性,大多数患儿仅表现出轻微的症状,如发热及伴随躯体相应部位的出疹。但少数患儿病情容易进展为重症,出现严重的并发症[5],如肺水肿、病毒性脑炎等[6],导致不良预后。2012 年我国流行病学调查[7]数据显示,月龄为 12～23 个月的患儿发病率和病死率最高,心肺或神经系统的并发症发生率为 1.1%,病死率为 3%,其中超过 90%的死亡病例与 EV-A71 型有关。因此,及早识别患儿重症化趋势,可及时进行临床治疗与干预。

机器学习是一门涉及多领域的交叉学科,如统计学、人工智能、概率论、数据挖掘等多种领域[8-9]。机器学习的算法种类繁多,按照学习方式可以分为非监督、半监督、监督式及强化学习;按照算法类似性又可以分为决策树、回归、聚类、人工神经网络及集成算法[10]。集成算法是一种非常强大的算法,包括 Boosting 技术(用于提升模型的正确率)、Bagging 技术(提高模型的稳定性)等。Boosting 技术作为一种高级算法,属于一种嵌套建模技术,包括建模和投票两个阶段,在建模时,通过多次迭代建立多个模型,通过投票阶段筛选最佳模型,最终将一组预测正确率较低模型组合变成一个整体正确率较高的模型[11]。很多机器学习库的代码质量比较差,需做大量的调优工作,而 CatBoost 只需少量调试,就可以实现良好的性能。本研究拟借助于机器学习算法,通过回顾性分析临床病例资料,探究 CatBoost 模型在预测重症 HMFD 的应用价值,同时通过与传统算法比较,评估该模型的预测性能,为后续研究提供更多的参考依据。

1 对象与方法

1.1 研究对象 收集郑州市某医院 2014 年 1 月—2017 年 6 月住院部诊治的 HFMD 患儿病例。所有病例的确诊均以《手足口病诊疗指南(2010 年版)》[12]为参考标准,将患儿分为轻症组、重症组,其中重症组以 3 期(心肺功能衰竭前期)为结局指标。

1.2 纳入及排除标准 纳入标准:(1)初次诊断的 HFMD 患儿;(2)患儿除 HFMD 疾病外,无其他基础疾病,一般情况尚好;(3)患儿病例信息及检验结果等资料完整。排除标准:(1)HFMD 恢复期的患儿;(2)HFMD 发病之前,已经合并心肺等其他并发症的患儿;(3)存在免疫力缺陷等先天性疾病的患儿。

1.3 资料收集 按照预先制定的表格,由专人负责对符合纳入标准的患儿病例资料进行记录及整理,包括基本住院信息(性别、年龄、入院和出院日期、居住地)、发病及查体情况(主诉、发热时间、最高体温、皮疹、呕吐、嗜睡、抽搐、肢体抖动、入院和出院诊断)、血常规、生化、免疫(S100 蛋白、白介素、免疫球蛋白、淋巴亚群)、病原学(抗体检查与病毒鉴定)及其他指标(降钙素原、细菌毒血定量、真菌葡聚糖)等。用于纳入模型的主要变量包括发病的年龄、发热、病程,以及重要的神经系统体征和血常规感染方面的指标等。使用 EpiData 3.1 软件,由两位研究者独立完成数据的录入,最后进行数据的一致性和可靠性检验。对有异议的数据,追溯原始资料进行取舍分析,保证资料的完整及真实性。

1.4 统计学分析 将 EpiData 3.1 软件导出数据进行后续分析,所有数据的处理使用 R 3.4.3 软件完成,主要用到的 R 包有'mlr'、'nnet'、'rpart'、'e1071'、'catboost'、'stats'、'pcalg'、'ggplot2'、'caret'等,其中'catboost'的相关官方使用介绍详见以下网址:https://tech.yandex.com/catboost/。使用 R 中相关函数进行数据的整理和统计学分析,对数据的完整性及整体分布情况进行分析,最后剔除缺失值、偏离值及其他异常值,经过数据的预处理,最终保留高质量的变量和数据。将数据划分为 70%的训练样本和 30%的测试样本,其中训练样本用于模型构建,测试样本用于模型评估。对模型进行逐一构建,最后评价模型的预测性能,输出总体预测正确率等,利用 ROC 曲线结果判断各模型的诊断价值。

2 结果

2.1 一般信息 对数据资料进行严格的筛选,最终纳入 2 983 例 HFMD 病例,其中轻症组 1 759 例,重症组 1 224 例。纳入资料的一般信息,见表 1。

表 1　HFMD 患儿纳入资料的一般信息[例(%)]

Table 1　General information of included data of children with HFMD(No. of cases[%])

项目	轻症（n = 1 759 例）	重症（n = 1 224 例）	χ^2	P
性别			1.059	0.303
男	1 119(63.6)	756(61.8)		
女	640(36.4)	468(38.2)		
年龄(岁)			22.257	<0.001
<1	403(22.9)	195(15.9)		
1～	746(42.4)	578(47.2)		
2～	352(20.0)	260(21.3)		
≥3	258(14.7)	191(15.6)		
发病时间(d)			13.279	<0.001
<3	1 310(74.5)	837(68.4)		
≥3	449(25.5)	387(31.6)		
居住地			30.307	<0.001
城市	1 053(59.9)	621(50.7)		
农村	664(37.7)	584(47.7)		
城乡结合部	42(2.4)	19(1.6)		
病原学结果			81.665	<0.001
EV-A71	282(16.0)	333(27.2)		
CV-A16	413(23.5)	338(27.6)		
其他阳性	1 000(56.9)	529(43.2)		
阴性	64(3.6)	24(2.0)		
体温(℃)			26.023	<0.001
37.5～	690(39.2)	387(31.6)		
38.5～	880(50.0)	646(52.8)		
≥39.5	189(10.8)	191(15.6)		
发热时间(d)			42.039	<0.001
<3	1 247(70.9)	728(59.5)		
≥3	512(29.1)	496(40.5)		
心率(次/分)			11.023	0.004
<130	1 321(75.1)	944(77.1)		
130～	432(24.6)	265(21.7)		
≥150	6(0.3)	15(1.2)		
白细胞($\times10^9$/L)			7.504	0.006
<10.8	1 168(66.4)	763(62.3)		
≥10.8	591(33.6)	461(37.7)		
中性粒细胞百分比(%)			9.201	0.002
<75	1 638(93.1)	1 102(90.0)		
≥75	121(6.9)	122(10.0)		
肢体抖动			557.415	<0.001
是	6(0.3)	355(29.0)		
否	1 753(99.7)	869(71.0)		
呕吐			297.731	<0.001
是	13(0.7)	220(18.0)		
否	1 746(99.3)	1 004(82.0)		
嗜睡			74.148	<0.001
是	1(0.1)	53(4.3)		
否	1 758(99.9)	1 171(95.7)		
抽搐			4.506	0.034
是	97(5.5)	91(7.4)		
否	1 662(94.5)	1 133(92.6)		
降钙素原(ng/mL)			19.492	<0.001
<0.1	1 001(56.9)	795(65.0)		
≥0.1	758(43.1)	429(35.0)		

2.2　模型构建　根据 R 包提供的相应模型函数功能，对 CatBoost 和几种普通常见模型分别建模，其中模型的参数配置按照函数提供的默认设置。共构建 CatBoost 模型、决策树模型、人工神经网络模型、支持向量机模型、贝叶斯网络模型及 logistic 回归模型 6 类模型。

2.3　模型预测性能评价　输出每个模型的总体预测正确率，其中 CatBoost 模型的预测正确率最高(87.6%)，人工神经网络模型位居第二(83.8%)，见图 1。根 CatBoost 算法模型，输出预测变量重要性图，其中居前 3 位的变量依次为呕吐、肢体抖动和病原学结果。见图 2。

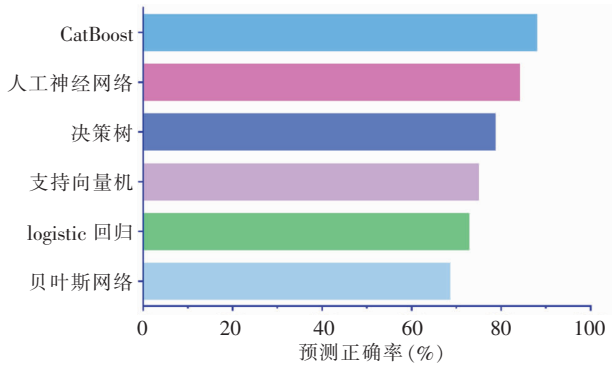

图 1　6 个模型诊断 HFMD 总体预测正确率的比较

Figure 1　Comparison of the overall prediction accuracy of six models for diagnosing HFMD

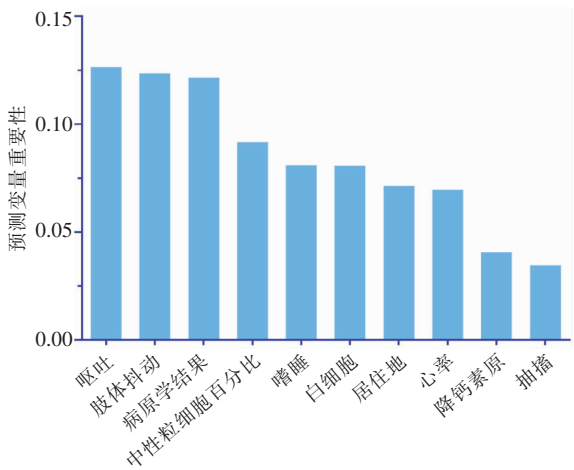

图 2　CatBoost 模型的预测变量重要性图

Figure 2　Importance of predictive variables in CatBoost model

为评估模型是否存在过度拟合，输出分类器校准图，对于 HFMD 重症组进行校准，结果显示，图中校准线距离理想参考线(图中虚线)较接近，表明模型拟合性能稳健，见图 3。输出每个模型所对应的诊断性能指标，结果显示 CatBoost 算法模型

ROC 曲线下面积、灵敏度、特异度均高(分别为 0.866、80.80%、92.33%)。见表 2。

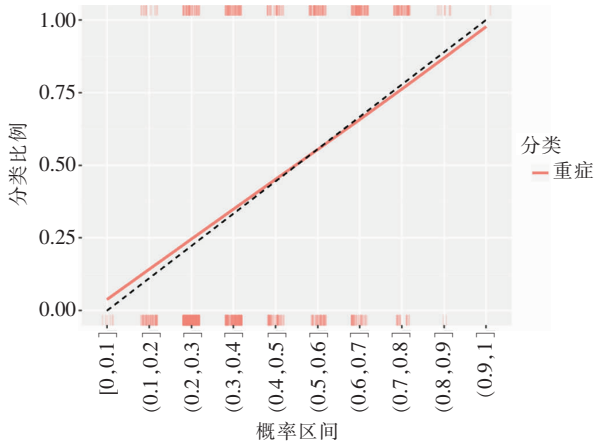

图 3　CatBoost 模型的分类器校准图(HFMD 重症组)

Figure 3　Classifier calibration plot for CatBoost model(severe HFMD group)

表 2　各模型 ROC 曲线分析相关诊断性能指标

Table 2　Diagnostic performance indicators related to ROC curve analysis for each model

模型类别	灵敏度(%)	特异度(%)	区域下面积	95%可信区间	
				下限	上限
CatBoost	80.80	92.33	0.866	0.851	0.880
人工神经网络	75.49	89.65	0.826	0.809	0.842
决策树	63.15	88.80	0.749	0.730	0.768
支持向量机	54.08	92.33	0.732	0.713	0.751
贝叶斯网络	56.54	89.88	0.732	0.713	0.751
logistic 回归	53.19	92.67	0.729	0.710	0.749

3　讨论

CatBoost 是由俄罗斯 Yandex 的研究人员和工程师开发的一种基于决策树梯度提升的新型集成算法，于 2017 年 7 月正式对外宣布开源，此前，在 Boosting 家族中，两大主流算法是 XGBoost 和 lightGBM，而据官方测评，该家族中新添的 CatBoost 成员模型性能超越上述两大算法。本研究通过机器学习算法，对 HFMD 的临床资料数据进行建模，与其他普通分类预测模型比较，发现 CatBoost 算法模型在预测重症 HFMD 中的总体正确率最高。目前，预测重症 HFMD 较多的研究使用 logistic 回归模型[13-16]。作为经典模型方法，logistic 回归属于一种广义线性回归模型，也可以用于分类预测，但主要适用于简单线性的二分类问题，在处理非线性问题方面存在不足，同时，由于其回归方程的构建过分依赖于训练样本，也存在过度拟合问题，导致最终构建的模型在检验样本的预测方面不理想。

在众多的机器学习模型中，处理分类(字符串)变量时经常会面临机器无法识别而报错的问题，因此需要将其用数字格式进行转换。常用的几种预处理方法有标签编码、独热编码等。而 CatBoost 可以直接使用分类特征，并具有可扩展性，可以轻松地与 Google 的 TensorFlow 和 Apple 的 Core ML 等深度学习框架相整合，以及处理各种数据类型，从而帮助解决当今企业面临的各种问题。最重要的是，CatBoost 提供了同类算法中的最佳正确率。相比于其他模型，该模型具有以下优势：(1)性能方面，能与任何领先的机器学习算法进行竞争；(2)自动处理分类特征，CatBoost 使用有关分类特征及分类和数字特征组合的各种统计数据，将分类值转换为数字，无需任何明确的预处理就可以将类别转换为数字；(3)稳定性，减少了对广泛的超参数调整的需求，降低了过拟合的机会，从而导致更广义的模型；(4)易于使用，提供了与 scikit 集成的 Python 界面，以及 R 和命令行界面。

ROC 曲线结果显示，CatBoost 算法具有较好的灵敏度和特异度，其曲线下面积也大于 logistic 回归模型，诊断性能较高。Zhang 等[17]在构建重症 HFMD 预测模型时，通过 R 统计软件使用了梯度提升树(gradient boosting tree，GBT)模型构建的决策树。GBT 是 Boosting 家族的一种算法，而 XGBoost 和 lightGBM 是在此基础上的新一代升级算法，其构建的 GBT 模型预测正确率可达到 92.3%，ROC 曲线下面积为 0.985，充分说明了新的集成算法的优势，支持本研究的结果。

CatBoost 模型筛选的预测变量居前 3 位的依次为呕吐、肢体抖动和病原学结果。Peng 等[18]纳入 14 项研究的荟萃分析结果显示，肢体抖动和呕吐为重症 HFMD 合并神经源性肺水肿的独立危险因素。在病原学结果方面，目前研究认为 EV-A71 型与重症 HFMD 存在相关性，Nadel 等[19]研究认为，EV-A71 型可以侵犯呼吸和神经系统，引起一系列并发症，如肺水肿、脑炎等。Cox 等[20]研究进一步揭示了 EV-A71 型的感染涉及宿主－病毒相互作用的分子机制，EV-A71 病毒利用自身病毒蛋白破坏宿主免疫，进而发生免疫逃逸。Luo 等[21]研究则通过大型流行病学调查发现，EV-A71 型的感染是 HFMD 进展为重症和危重症的主要原因，其占比分别为 65.75%和 88.78%。谭艳芳等[22]研究亦强调了

EV-A71 阳性是重症 HFMD 的危险因素，与本研究得出的病原学结果为重要预测变量结果一致。

近年来，利用机器学习算法预测疾病，以及辅助临床医生诊断等方面的研究逐渐成为热点。可以预想，借助于机器学习及大数据分析平台，未来国家在传染病的预测监管方面可以实现相应决策的动态调整。

本研究虽然纳入了较多样本，以及通过一系列数据处理流程提高数据的质量，保证分析结果的稳定性，但仍然存在一些局限性：(1)受收集病例资料来源限制，本研究虽然纳入了许多预测变量进行筛选，但仍不够全面，可能存在未纳入的潜在预测变量。(2)由于 CatBoost 目前仅在编程 Python 软件中提供了完整的功能库，而本研究仅使用了 R 软件，并未使用 Python 软件。R 比较局限于科研工作者进行数据科学分析，而 Python 涵盖了很大范围的用户群体，主要包括程序员和科研人员等，在计算机方面庞大的编程功能中，数据分析只是一个分支。由于目前该算法的 R 包中仅提供了基础的模型构建功能，而未完整地进行相应模块的编译和封装，因此尚无法实现更多数据可视化的高级功能。

综上所述，本研究通过机器学习算法发现，CatBoost 模型可以用于预测重症 HFMD，相比于其他传统算法，具有较高的预测正确率和诊断价值，后续更多功能的开发仍然需要借助于主流编程 Python 软件以及更多研究的深入开展。

[参 考 文 献]

[1] Wang Y, Feng Z, Yang Y, et al. Hand, foot, and mouth disease in China: patterns of spread and transmissibility[J]. Epidemiology, 2011, 22(6): 781 - 792.

[2] Aswathyraj S, Arunkumar G, Alidjinou EK, et al. Hand, foot and mouth disease (HFMD): emerging epidemiology and the need for a vaccine strategy[J]. Med Microbiol Immunol, 2016, 205(5): 397 - 407.

[3] Osterback R, Vuorinen T, Linna M, et al. Coxsackievirus A6 and hand, foot, and mouth disease, Finland[J]. Emerg Infect Dis, 2009, 15(9): 1485 - 1488.

[4] Lu QB, Zhang XA, Wo Y, et al. Circulation of Coxsackievirus A10 and A6 in hand-foot-mouth disease in China, 2009 - 2011[J]. PLoS One, 2012, 7(12): e52073.

[5] Ooi MH, Wong SC, Lewthwaite P, et al. Clinical features, diagnosis, and management of enterovirus 71[J]. Lancet Neurol, 2010, 9(11): 1097 - 1105.

[6] Liu Z, Wang S, Yang R, et al. A case-control study of risk factors for severe hand-foot-mouth disease in Yuxi, China, 2010 - 2012[J]. Virol Sin, 2014, 29(2): 123 - 125.

[7] Xing W, Liao Q, Viboud C, et al. Hand, foot, and mouth disease in China, 2008 - 12: an epidemiological study[J]. Lancet Infect Dis, 2014, 14(4): 308 - 318.

[8] Baştanlar Y, Ozuysal M. Introduction to machine learning[J]. Methods Mol Biol, 2014, 1107: 105 - 128.

[9] Deo RC. Machine learning in medicine[J]. Circulation, 2015, 132(20): 1920 - 1930.

[10] 林巧. 数据挖掘中决策树算法的探讨[J]. 伊犁师范学院学报(自然科学版), 2007, 9(3): 36 - 38.

[11] 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, 17(1): 52 - 59.

[12] 中华人民共和国卫生部. 手足口病诊疗指南(2010 年版)[J]. 国际呼吸杂志, 2010, 30(24): 1473 - 1475.

[13] 吕云磊, 朱凤才, 杨小平. 常州市金坛区托幼机构手足口病影响因素的流行病学调查[J]. 中华疾病控制杂志, 2016, 20(10): 1011 - 1013.

[14] 周喜桃, 肖鹏程, 曾莉怡, 等. 手足口病住院患儿的病原学和临床特征[J]. 中国感染控制杂志, 2017, 16(11): 1069 - 1073.

[15] 周永东, 颜云盈. 2008—2012 年南宁市 7 岁以下儿童手足口病流行病特征分析[J]. 现代预防医学, 2015, 42(18): 3276 - 3279.

[16] 陈庆会. 肠道病毒 71 型致手足口病合并脑炎 81 例临床分析[J]. 中国感染控制杂志, 2011, 10(5): 354 - 356.

[17] Zhang B, Wan X, Ouyang FS, et al. Machine learning algorithms for risk prediction of severe hand-foot-mouth disease in children[J]. Sci Rep, 2017, 7(1): 5368.

[18] Peng L, Luo R, Jiang Z. Risk factors for neurogenic pulmonary edema in patients with severe hand, foot, and mouth disease: A meta-analysis[J]. Int J Infect Dis, 2017, 65: 37 - 43.

[19] Nadel S. Hand, foot, mouth, brainstem, and heart disease resulting from enterovirus 71[J]. Crit Care Med, 2013, 41(7): 1821 - 1822.

[20] Cox JA, Hiscox JA, Solomon T, et al. Immunopathogenesis and virus-host interactions of enterovirus 71 in patients with hand, foot and mouth disease[J]. Front Microbiol, 2017, 8: 2249.

[21] Luo KW, Gao LD, Hu SX, et al. Hand, foot, and mouth disease in hunan province, China, 2009 - 2014: epidemiology and death risk factors[J]. PLoS One, 2016, 11(11): e0167269.

[22] 谭艳芳, 魏婷婷, 欧阳文献, 等. 重症 EV-A71 型手足口病炎症因子的临床意义[J]. 中国感染控制杂志, 2017, 16(12): 1156 - 1160.

(本文编辑:左双燕)

本文引用格式:王斌,冯慧芬,王芳,等. 基于机器学习的 CatBoost 模型在预测重症手足口病中的应用[J]. 中国感染控制杂志,2019,18(1):12 - 16. DOI:10.12138/j.issn.1671 - 9638.20193918

Cite this article as: WANG Bin, FENG Hui-fang, WANG Fang, et al. Application of CatBoost model based on machine learning in predicting severe hand-foot-mouth disease[J]. Chin J Infect Control, 2019,18(1): 12 - 16. DOI: 10.12138/j.issn.1671 - 9638.20193918