·论 著·

极端梯度上升模型在预测临床重症手足口病中的应用价值

赵敬[1]， 冯慧芬[2]， 王斌[1]， 黄平[1]

1. 郑州大学第五附属医院消化内科，河南 郑州 450052；

2. 郑州大学第五附属医院感染科，河南 郑州 450052

摘要：目的 探讨极端梯度上升模型(XGBoost)和 Logistic 回归模型在临床重症手足口病(HFMD)预测中的应用对比。**方法** 回顾性收集郑州市某医院 2017 年 3 月至 11 月期间住院部收治的 HFMD 患儿 872 例的临床资料，其中轻症 488 例，重症 384 例。使用 R3.4.4 软件进行所有资料的分析，分别构建 XGBoost 和 Logistic 回归模型，比较两种模型对重症 HFMD 的预测效果。**结果** 在 XGBoost 模型中，输出变量重要性中前三位分别为：白细胞计数、年龄和心率，其对重症 HFMD 总体预测准确性为 92.4%，ROC 曲线下面积为 0.952(95% *CI*：0.931～0.967)。Logistic 回归模型总体预测准确性为 80.1%，ROC 曲线下面积为 0.848(95% *CI*：0.833～0.866)。模型评估显示 XGBoost模型的预测效果明显优于 Logistic 回归模型。**结论** XGBoost 模型可以用于预测重症 HFMD，相比于传统模型，具有较高的准确性和诊断价值。

关键词： 手足口病，重症；极端梯度上升模型；Logistic 回归模型；预测

中图分类号： R 512.5 **文献标识码：** A **文章编号：** 1674－8182(2019)10－1323－04

Value of extreme gradient boosting model in predicting clinical severe hand-foot-mouth disease

ZHAO Jing*, FENG Hui-fen, WANG Bin, HUANG Ping

* *Department of Gastroenterology, Fifth Affiliated Hospital of Zhengzhou University, Zhengzhou, Henan* 450052, *China*

*Corresponding author: FENG Hui-fen, E-mail: huifen.feng@*163.*com*

Abstract: Objective To compare the application of extreme gradient boosting (XGBoost) model and Logistic regression model in the prediction of severe hand-foot-mouth disease (HFMD). **Methods** Clinical data of 872 children with HFMD admitting to hospital from March 2017 to November 2017 were collected retrospectively, including mild cases (n = 488) and severe cases(n = 384). R 3.4.4 software was used for all data analysis, XGBoost and Logistic regression models were constructed respectively. The predictive effects of two models on severe HFMD were compared. **Results** In XGBoost model, the first three most important output variables were white blood cell count, age and heart rate. The overall accuracy of predicting severe HFMD was 92.4%, and the area under receiver operating characteristic curve(AUC) was 0.952(95% *CI*:0.931－0.967). In Logistic regression model, the overall accuracy of predicting severe HFMD was 80.1%, and AUC was 0.848 (95% *CI*:0.833－0.866). Model evaluation showed that XGBoost model was superior to Logistic regression model in prediction effect. **Conclusion** XGBoost model can be used to predict severe HFMD disease and has a higher precision and diagnostic value compared with traditional model.

Key words: Hand-foot-mouth disease, severe; Extreme gradient boosting model; Logistic regression model; Prediction

Fund program: National Natural Science Foundation of China (81473030); Research Project Overseas of Health System in Henan Province (2015065)

手足口病(hand-foot-mouth disease, HFMD)是由肠道病毒引起的传染病，引发 HFMD 的肠道病毒有 20 多种，以肠道病毒 71 型(EV71)和柯萨奇 A 组 16 型(CoxAl6)多见[1]。HFMD 多发生于 5 岁以下儿童，表现为口痛、厌食、低热及手、足、口腔等部位出现小疱疹或小溃疡[2-3]。重症 HFMD 病情进展迅速，

DOI：10.13429/j.cnki.cjcr.2019.10.005

基金项目： 国家自然科学基金 (81473030)；河南省卫生系统出国研修项目 (2015065)

通讯作者： 冯慧芬，E-mail：huifen.feng@163.com

在发病 1 ~5 d 左右出现脑膜炎、脑炎、脑脊髓炎、肺水肿、循环障碍等,极少数病例病情危重,可致死亡,存活病例可留有后遗症[2,4-5]。而目前尚缺乏有效治疗药物,因此及早识别患儿重症化趋势,可进行临床治疗与干预。

极端梯度上升(extreme gradient boosting, XGBoost)是一种在梯度提升模型(gradient boosting machine, GBM)基础上升级后的新一代集成学习算法[6]。GBM 为 Boosting 家族中经典代表系列,其在数据挖掘领域比较流行,它通过将各个弱分类器加权叠加形成强分类器,从而有效降低误差,达到准确的分类效果。XGBoost 作为一种新型算法,与 GBM 相比,其运算所需时间较短,且精度较高。将其运用于工程领域,可提高诊断的速度和准确率。本研究旨在借助 XGBoost 算法构建用于预测重症 HFMD 的机器学习模型,为后续研究新的模型提供价值参考,以及为临床疾病诊断和重症预警提供更科学的依据。

1 对象与方法

1.1 研究对象 回顾性收集郑州市某医院 2017 年 3 月至 11 月住院部收治的 HFMD 患儿的临床资料。HFMD 的确诊均依据《手足口病诊疗指南(2010 年版)》[7]确定,轻型为手、足、口、臀部皮疹伴或不伴发热;重型为有神经系统受累表现,体征见脑膜刺激征、腱反射减弱或消失。据此将患儿分为轻症组和重症组。

1.2 纳入和排除标准 纳入标准:(1)新确诊的 HFMD 病例;(2)基本住院信息以及各种实验室检查结果完整者;(3)发病时间 < 1 周者。排除标准:(1)处于恢复期的 HFMD;(2)既往有传染病的感染或接触史;(3)存在先天发育缺陷或导致免疫力低下的其他疾病。

1.3 资料提取 临床资料的提取用 EpiData 3.1 软件手动录入完成。相关质控过程,使用双人独立录入后的一致性检验和对数据录入格式进行可靠性检验,对于有争议的,追溯原始病历资料进行校对。最后将数据导出用于分析。经过严格筛选,最终纳入 872 例 HFMD,其中轻症组 488 例,重症组 384 例。使用四格表 χ^2 检验进行组间比较,一般基本信息见表 1。表 1 显示,年龄、病原(EV71)、发病天数、发热时间、心率、血糖、外周血白细胞、中性粒细胞比例等 8 个变量是影响 HFMD 重症与否的因素($P<0.01$, $P<0.05$)。

1.4 模型构建 按照输入格式的要求,将数据按照单热编码方式转化为稀疏矩阵,同时分割为训练样本(70%,轻症组342例,重症组269例)以及测试样本(30%,轻症组 146 例,重症组 115 例)。训练样本用于模型的构建,而测试样本用于对模型性能进行评估。建模均采用训练样本的数据。分别使用‘xgboost’、‘stats’包构建 XGBoost 和 Logistic 回归模型,相关参数采取包中默认的设置。其中 XGBoost 如下:最大树深度为 6,学习速率为 0.3,CPU 线程为 2,迭代次数为 100 次,构建最大树的数目为 100 棵。模型对种树的数量逐步绘制错误率,通常决定构建树的最佳数目,即第 50 棵树后,错误率不再进一步减少,停止迭代;Logistic 参数为:使用二分类、主效应、逐步回归法。在数据挖掘和关联规则学习中,提升图是衡量目标模型(关联规则)在预测或分类病例时具有增强的响应(相对于总体人群)的度量,以随机选择为目标模型。如果目标范围内的反应比整个人口的平均水平好得多,那么目标模式就会做得很好。提升只是这些值的比率:目标响应除以平均响应;学习曲线以不断采集增加的子样本进行训练,并且测量模型性能。测量过程将按照给定的重采样方法和对每个子

表 1 纳入资料的一般信息

项目	轻症组(n=488)		重症组(n=384)		χ^2 值	P 值
	例数	%	例数	%		
性别						
男	314	64.3	246	64.1	0.007	0.931
女	174	35.7	138	35.9		
年龄(岁)						
<3	153	31.4	224	58.3	63.743	0.000
≥3	335	68.6	160	41.7		
病原(EV71)						
阴性	440	90.2	77	20.1	437.655	0.000
阳性	48	9.8	307	79.9		
发病天数(d)						
<3	424	86.9	309	80.5	6.603	0.010
≥3	64	13.1	75	19.5		
发热时间(d)						
<3	432	88.5	321	83.6	4.434	0.035
≥3	56	11.5	63	16.4		
体温(℃)						
<38.5	175	35.9	121	31.5	1.814	0.178
≥38.5	313	64.1	263	68.5		
心率(次/min)						
<130	419	85.9	362	94.3	16.262	0.000
≥130	69	14.1	22	5.7		
嗜睡						
否	157	32.2	114	29.7	0.619	0.431
是	331	67.8	270	70.3		
血糖(mmol/L)						
<8.3	414	84.8	281	73.2	18.056	0.000
≥8.3	74	15.2	103	26.8		
白细胞($\times10^9$/L)						
<10.8	94	19.3	267	69.5	223.841	0.000
≥10.8	394	80.7	117	30.5		
中性粒比率(%)						
<75	86	17.6	95	24.7	6.617	0.010
≥75	402	82.4	289	75.3		

采样的值进行重复。

1.5　统计学方法　所有资料的分析使用 R 3.4.4 软件，对原始资料进行数据清洗及整理等，去除缺失值及异常值的个案，同时对连续性变量进行二分类处理，使用 χ^2 检验对分类变量进行相关检验，检验水准取 $\alpha = 0.05$。在建模环节，构建 XGBoost 和 Logistic 回归模型，其中 Logistic 回归模型作为参照，最后使用 ROC 曲线、提升图及学习曲线对两个模型的性能进行评估。

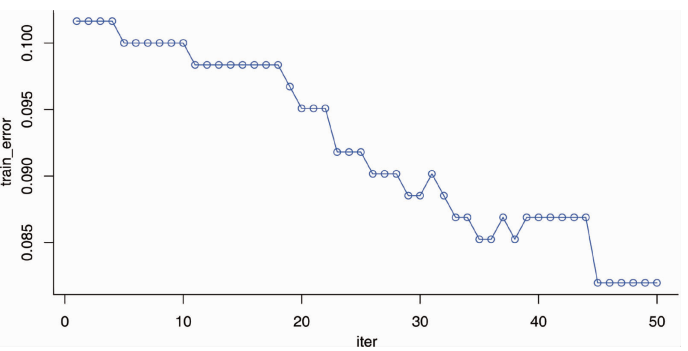
2　结　果

2.1　两种模型的危险因素分析　输出 XGBoost 的迭代错误率图，结果见图 1。根据 *OR* 值排序，Logistic 回归模型（表 2）居前三位的变量为血糖、心率和病原（EV71）阳性；绘制 XGBoost 预测 8 个变量的重要性图，见图 2，其显示，在 XGBoost 模型中，输出变量重要性中前三位分别为白细胞计数、年龄和心率，即它们依次是预测重症 HFMD 三个最具重要性的变量。

2.2　模型总体预测性能和 ROC 曲线　使用检验样本对上述两个模型进行性能检验。计算其对重症 HFMD 总体预测的准确性，XGBoost 为 92.4%，Logistic 为 80.1%。绘制两者的 ROC 曲线图，曲线下面积（AUC）XGBoost 和 Logistic 分别为 0.952（95% *CI*：0.931～0.967），0.848（95% *CI*：0.833～0.866）。两种模型的 AUC 检验均具有统计学意义（$P < 0.05$），提示对重症 HFMD 的预测效果，XGBoost 模型明显优于 Logistic 模型。见图 3。

2.3　模型提升图比较　由图 4 可以看出 XGBoost 模型的提升图与 Logistic 回归模型的提升图无太大差异，但总体上 XGBoost 模型优于 Logistic 回归模型。

2.4　模型学习曲线　由图 5 可以看出 XGBoost 模型的学习曲线明显优于 Logistic 回归模型，对数据的总体预测性能 XGBoost 模型表现更佳。



注：iter 代表迭代次数。

图 1　XGBoost 模型的迭代误差图

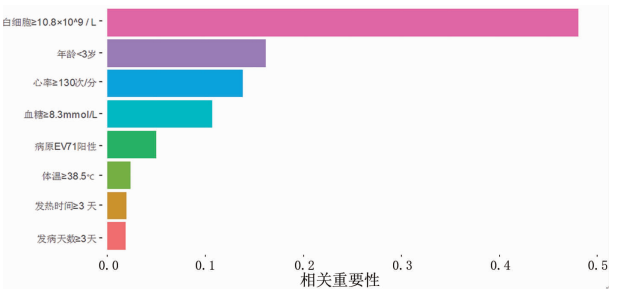


图 2　XGBoost 模型的预测变量重要性

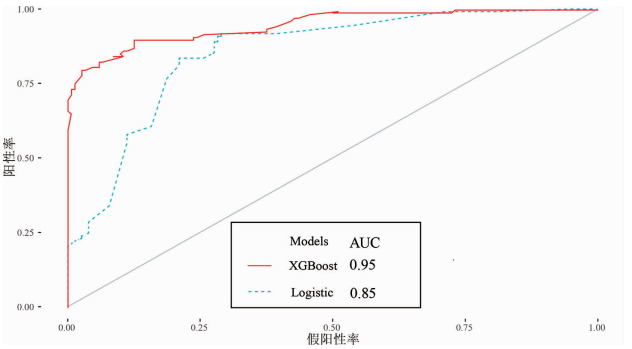


图 3　两种模型的 ROC 曲线图

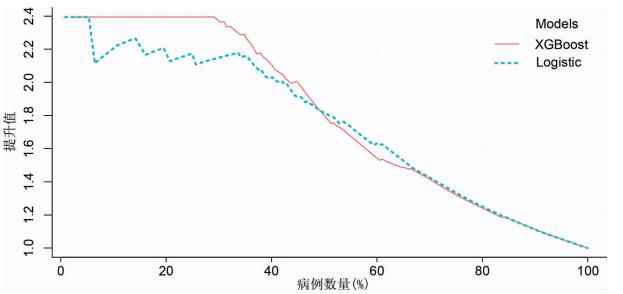
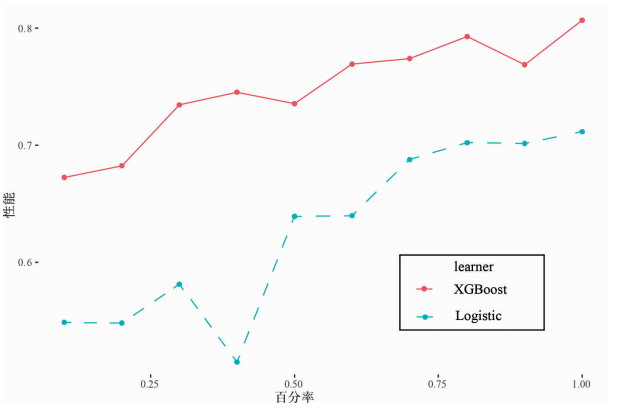


图 4　两种模型的提升图



注：X 轴表示相对子采样大小，Y 轴表示性能。

图 5　两种模型的学习曲线

表 2　HFMD 患儿危险因素的多因素 Logistic 回归分析

自变量	*OR* 值	*P* 值	95% *CI*
年龄 <3 岁	1.021	<0.01	0.523～2.521
病原（EV71）阳性	3.580	0.004	1.529～8.904
发病天数 ≥ 3 d	1.032	0.761	0.480～2.408
发热时间 ≥ 3 d	2.011	0.147	0.821～4.882
心率 <130 次/min	3.023	0.017	1.135～8.904
血糖 ≥ 8.3 mmol/L	9.796	0.001	3.616～28.817
外周血白细胞 $<10.8\times10^9$/L	1.464	0.286	0.680～3.552
中性粒细胞比率 <75%	2.893	0.136	0.738～11.892

3 讨 论

本研究使用 XGBoost 算法构建了一个预测重症 HFMD 的模型,并以 Logistic 回归模型作为参考,通过多种方法对 XGBoost 模型性能进行了对比研究。目前在重症 HFMD 的预测模型方面,既往较多学者都采用传统的 Logistic 回归模型[8-9]。作为经典预测模型,Logistic 回归模型最简单易学,借助于 SPSS 统计软件就可以实现,其在探讨重症 HFMD 危险因素方面有一定价值。但是,随着近年来医学大数据的提出,越来越多的研究将焦点集中在机器学习中,而新的机器学习算法不断被提出,传统算法已经无法应对各种庞大复杂的医学大数据[10]。Logistic 回归模型由于使用广义线性回归方法,因此对于分析非线性问题明显不适合,而医学中很多问题并非简单的线性问题,使得模型拟合存在很大偏差。此外,该传统模型还存在容易过度拟合和对缺失数据敏感等不足。而 XGBoost 算法作为新一代的梯度提升系列的主流算法,与 LightGBM 占据着机器学习算法的领先地位,在国际 KDD Cup、Kaggle 组织的很多数据挖掘竞赛中 XGBoost 屡次夺冠。

相比于传统 Logistic 回归,XGBoost 算法具有以下优点:(1)当 XGBoost 应用于线性分类时,等同于带有 L1 和 L2 正则化的线性回归算法。(2)在分布式算法方面,XGBoost 会把每一维度的特征在一台机器内进行排序,并保存在 Block 结构内;这样,多个计算就可以同时在不同机器内进行,并将最后的结果汇总;因此,加快了 XGBoost 的计算速度。(3)由于特征值只用在了排序方面,所以异常值对 XGBoost 模型的建立影响相对比较小。(4)应用于多特征值方面,由于每次的计算只选择梯度减少最大的特征,因此特征值之间的相关性问题也可以得到解决。基于这些强大的优势,才使得 XGBoost 算法具有更高效和更高精度的预测能力,从而在机器学习方面得到广泛的应用,并活跃在各种国际数据挖掘竞赛舞台[11-12]。

本研究构建的 XGBoost 相比于 Logistic 回归模型,在多种预测性评估指标中,可以看出在总体预测正确率 AUC 中,XGBoost 明显优于 Logistic 回归模型,表明 XGBoost 具有更高的预测准确性。从提升图的对比中,XGBoost 模型的提升图与 Logistic 回归模型的提升图无太大差异,但总体上 XGBoost 模型还是优于 Logistic 回归模型。从学习曲线图的对比中,可以看到 XGBoost 模型的学习曲线明显优于 Logistic 回归模型,对数据的总体预测性能 XGBoost 模型表现更佳。在预测的变量重要性方面,本研究得出的结果为白细胞最重要,这与 Zhang 等[13]通过使用 GBM 算法的预测结果一致;在其他变量方面,也与既往学者的研究结果较为符合[14-15]。

综上所述,本研究采用 XGBoost 分类算法与 Logistic 回归算法对重症 HFMD 的预测进行了对比,结果发现 XGBoost 模型可用于预测重症 HFMD,相比于传统模型,具有较高的精准性和诊断价值。

参考文献

[1] Xing WJ, Liao QH, 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.

[2] 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.

[3] Liu BY, Luo L, Yan SY, et al. Clinical features for mild hand, foot and mouth disease in China [J]. PLoS One, 2015, 10(8): e0135503.

[4] 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.

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

[6] Sheridan RP, Wang WM, Liaw A, et al. Extreme gradient boosting as a method for quantitative structure-activity relationships[J]. J Chem Inf Model, 2016, 56(12): 2353-2360.

[7] 中华人民共和国卫生部. 手足口病诊疗指南(2010 版)[J]. 中国实用乡村医生杂志, 2012, 19(19): 9-11.

[8] 刘丹, 苏豪浩, 王建红, 等. 手足口病重症病例的流行特征及危险因素[J]. 实用医学杂志, 2013, 29(6): 995-998.

[9] 陈璐, 任静朝, 段广才, 等. 手足口病重症化影响因素分析及预测模型的建立[J]. 中华实用儿科临床杂志, 2017, 32(6): 469.

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

[11] Dong H, Xu X, Wang L, et al. Gaofen-3 PolSAR Image Classification via XGBoost and Polarimetric Spatial Information [J]. Sensors (Basel), 2018, 18(2): E611.

[12] Torlay L, Perrone-Bertolotti M, Thomas E, et al. Machine learning-XGBoost analysis of language networks to classify patients with epilepsy[J]. Brain Inform, 2017, 4(3): 159-169.

[13] 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.

[14] Chew SP, Chong SL, Barbier S, et al. Risk factors for severe hand foot mouth disease in Singapore: a case control study[J]. BMC Infect Dis, 2015, 15: 486.

[15] Fang YR, Wang SP, Zhang LJ, et al. Risk factors of severe hand, foot and mouth disease: a meta-analysis[J]. Scand J Infect Dis, 2014, 46(7): 515-522.

收稿日期:2019-01-02 **修回日期**:2019-02-25 **编辑**:石嘉莹