DOI: 10.3969/j.issn.1005-8982.2018.23.010
文章编号： 1005-8982（2018）23-0048-05

新进展研究·论著

数据挖掘与模型构建在预测重症手足口病中的应用 *

黄平[1]，冯慧芬[2]，王斌[1]，赵敬[1]，易佳音[1]
（郑州大学第五附属医院 1. 消化内科，2. 感染科，河南 郑州 450052）

摘要：目的 探讨数据挖掘与模型构建在预测重症手足口病方面的价值。**方法** 回顾性分析郑州大学第五附属医院 2016 年 6 月 -2017 年 10 月收治的 838 例手足口病患儿的临床资料，使用 SPSS Statistics 23.0 统计软件进行数据的预处理和分析，使用 SPSS Modeler 18.0 软件进行模型构建和评估。根据总体精确性对所有算法进行筛选，选取最优算法，配置模型参数，输出分类树模型，评估模型的预测性能。**结果** 经过自动分类器筛选，最终确定 C&R 算法最佳。模型共纳入 3 个解释变量：易惊、呕吐及肢体抖动。使用错分矩阵计算后，模型的预测正确率为 91.17%，敏感性为 84.36%，特异性为 96.25%。ROC 曲线下面积为 0.903[（95% CI：0.878，0.927），P =0.000]。**结论** 决策树模型在预测手足口病方面有一定的优势，模型预测精确度较高，对临床疾病诊疗有一定的辅助价值。

关键词： 重症手足口病；数据挖掘；模型构建；决策树
中图分类号： R512.5　　**文献标识码：** A

Application of data mining and model construction in prediction of severe hand-foot-mouth disease*

Ping Huang[1], Hui-fen Feng[2], Bin Wang[1], Jing Zhao[1], Jia-yin Yi[1]
(1. Department of Gastroenterology, 2. Department of Infectious Diseases, the Fifth Affiliated Hospital of Zhengzhou University, Zhengzhou, Henan 450052, China)

Abstract: Objective To explore the value of data mining and model construction in predicting severe hand-foot-mouth disease (HFMD). **Methods** A retrospective analysis was performed on the clinical data of 838 children with HFMD treated in the Fifth Affiliated Hospital of Zhengzhou University from June 2016 to October 2017. SPSS Statistics 23.0 was used for data preprocessing and statistical analysis, while SPSS Modeler 18.0 was used for modeling and evaluation. The model parameters were configured to output classification tree model and assess predictive performance when the optimal algorithm was screened from all algorithms based on overall accuracy. **Results** C&R algorithm was finally determined to have better accuracy by the automatic classifier screening. The model included three explanatory variables: shock, vomiting and limb shaking. The prediction accuracy of the model was 91.17%, with the sensibility of 84.36% and the specificity of 96.25%. The area under the ROC curve was 0.903 (95% CI: 0.878, 0.927) (P < 0.05). **Conclusions** Decision tree model has some advantages in the prediction of hand, foot and mouth disease and has high prediction accuracy. The model has a supplementary value in clinical diagnosis and treatment of the disease.

Keywords: severe hand-foot-mouth disease; data mining; model establishing; decision tree

收稿日期：2018-02-13
* 基金项目：国家自然科学基金（No：81473030）；河南省医学科技攻关普通项目（No：201403130）；河南省卫生系统出国研修项目（No：2015065）
[通信作者] 冯慧芬，E-mail：huifen.feng@163.com；Tel：13938587058

手足口病（hand-foot-mouth disease, HFMD）是一种由肠道病毒引起的，以婴幼儿发病为主的急性传染性疾病[1]。尽管多数患儿表现为症状轻微，但也有少部分患儿因各种严重的神经系统、呼吸系统并发症而导致后遗症，甚至死亡，因而如何早期识别重症患者成为临床医生面临的重要难题[2]。数据挖掘是一种从大量的数据中，通过数理模式来探索隐藏数据中未知规律的过程。本研究通过数据挖掘的思想，构建决策树模型，从复杂的临床资料中找出较佳的预测指标，从而为临床医生 HFMD 诊断治疗提供一种辅助决策手段。

1　资料与方法

1.1　一般资料

选取 2016 年 6 月 -2017 年 10 月于郑州大学第五附属医院收治的 HFMD 患儿 838 例。其中，男性 513 例，女性 325 例；年龄 3 个月 ~ 4 岁，平均（2.3 ± 1.1）岁；平均住院时间（4.5 ± 0.8）d。所有患儿经病原学确诊。根据《手足口病诊疗指南（2010 年版）》中的诊断标准，将所有患儿分成轻症组 480 例和重症组 358 例[3]。轻症组：仅表现为手、足、口及臀部的皮疹，伴或不伴发热；重症组：出现神经系统受累的表现，如头痛、呕吐；精神差、嗜睡、易惊、谵妄及惊厥；肢体抖动，肌阵挛、眼球震颤、共济失调及眼球运动障碍；无力或急性弛缓性麻痹；体征可见脑膜刺激征，腱反射减弱或消失。收集患儿信息资料，初步制定预选的分析变量，包括性别、年龄、发热时间、最高体温、易惊、肢体抖动、抽搐、寒战、嗜睡及呕吐等。

1.2　方法

由专人负责设计问卷调查表，通过交叉核对，使用 EpiData 3.1 软件进行原始数据的录入。通过一系列数据整理，包括去除缺失、异常及重复个案等，最后生成一份完整的数据。对所有预测变量进行二分类处理，并赋值为 0 或 1。其中连续性变量处理后分别为：年龄 <3 岁，发热时间 ≥ 3 d，体温 ≥ 38.5 ℃，白细胞 ≥ 10.8×10^9/L，中性粒细胞比率 ≥ 75%，血糖 ≥ 8.3 mmol/L。满足上述条件的均赋值为 1，不满足上述条件的赋值为 0。其余变量按照是否存在相应症状，将是赋值为 1，否赋值为 0；性别男赋值为 1，女赋值为 0；居住地农村赋值为 1，城市赋值为 0。最后将处理后的数据，进行统计学分析。

1.3　统计学方法

数据分析采用 SPSS Statistics 23.0 统计软件，模型构建和评估采用 SPSS Modeler 18.0 软件。Modeler 软件在决策树构建模块提供了多种算法，包括随机树、分类和回归（classification and regression, C&R）树、C5.0、χ^2 自动交互检测法（chi-squared automatic interaction detector, CHAID）及高效统计树（quick unbiased efficient statistical tree, QUEST）等。所有算法的基本操作相同，即将数据分隔成多个子组来实现最佳分类或预测，但因输入和目标（输出）字段的类型是连续型变量或分类变量而有区别。其中 C&R 和 CHAID 的输入和目标字段可以是连续或分类变量，而 QUEST 和 C5.0 要求目标字段必须是分类变量。根据以上原理，在 Modeler 软件中，先选用自动分类器，对上述常见算法进行建模，最后根据总体精确性对所有算法进行筛选，选取最优算法，配置模型参数，输出分类树模型，评估模型的预测正确率，输出模型的累计收益图，评估模型拟合效果，同时绘制受试者工作特征（receiver operating characteristic, ROC）曲线，评估模型的诊断性能。

2　结果

2.1　模型构建及患者临床资料比较

经过自动分类器筛选，最终确定 C&R 算法最佳。C&R 树是个组合，包括分类树和回归树，目标变量为分类变量时使用分类树，以 Gini 系数来确认分割点，为连续型变量时则使用回归树，以方差来确认分割点。决策树包括种树和修建 2 个环节，模型参数设置包括构建单个决策树，最大树深度为 5 层，修建树以防止过度拟合。中止规则为父分支使用最小记录数 2%，子分支为 1%。两组患儿居住地、发热时间、体温、白细胞、中性比率、血糖、精神差、嗜睡、易惊、肢体抖动、呕吐、寒战及咽部疱疹比较，差异有统计学意义（P <0.05）。见表 1 和图 1 ~ 3。

2.2　模型结果

C&R 算法构建的决策树共包括 3 层 7 个节点，其中终末节点共有 4 个。模型共纳入 3 个解释变量：易惊、呕吐及肢体抖动。决策树生成原理第一步为训练样本集生成决策树的过程，第二步为决策树的剪枝过程，以新的测试数据为对象进行模型的修建过程，即总样本被分为训练和测试 2 个数据集，图中节点 0 的

表 1　患者的临床资料比较

组别	男 / 女 / 例	年龄 / 岁		居住地 / 例		发热时间 /d		体温 /℃		白细胞 / × 10^9/L	
		<3	≥ 3	农村	城市	<3	≥ 3	<38.5	≥ 38.5	<10.8	≥ 10.8
轻症组（n =480）	289/191	53	427	379	101	445	35	262	218	350	130
重症组（n =358）	224/134	39	319	338	20	303	55	151	207	208	150
χ^2 值	0.482	0.005		39.647		13.936		12.624		20.232	
P 值	0.488	0.946		0.000		0.000		0.000		0.000	

组别	中性比率 /%		血糖 /（mmol/L）		斑疹 / 例		丘疹 / 例		疱疹 / 例		精神差 / 例	
	<75	≥ 75	<8.3	≥ 8.3	是	否	是	否	是	否	是	否
轻症组（n =480）	460	20	418	62	421	59	254	226	99	381	174	306
重症组（n =358）	331	27	256	102	311	47	201	157	69	289	211	147
χ^2 值	4.413		31.602		0.130		0.861		0.234		42.503	
P 值	0.036		0.000		0.718		0.353		0.629		0.000	

组别	嗜睡 / 例		易惊 / 例		肢体抖动 / 例		呕吐 / 例		抽搐 / 例		寒战 / 例		咽部疱疹 / 例	
	是	否	是	否	是	否	是	否	是	否	是	否	是	否
轻症组（n =480）	0	480	10	470	112	368	89	391	5	475	38	442	37	443
重症组（n =358）	5	353	101	257	252	106	107	251	8	350	58	300	68	290
χ^2 值	6.744		121.830		184.821		14.734		1.911		13.875		23.832	
P 值	0.009		0.000		0.000		0.000		0.167		0.000		0.000	

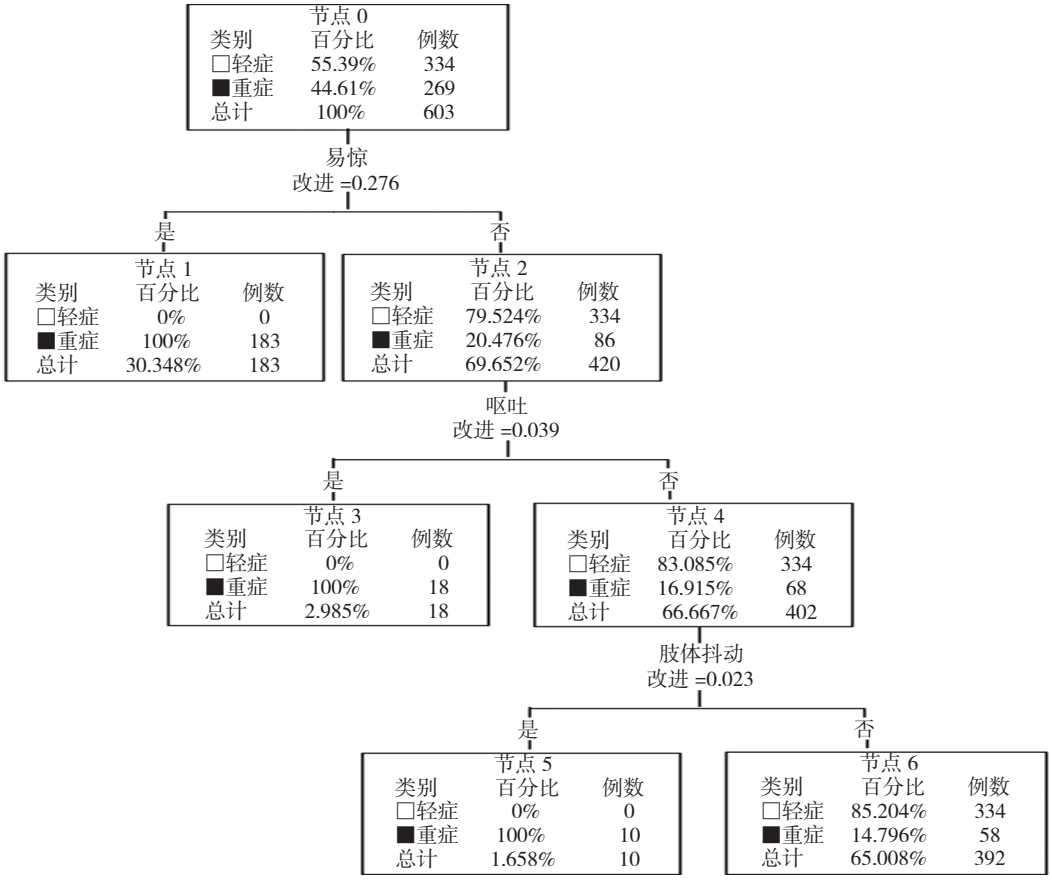


图 1　C&R 算法决策树

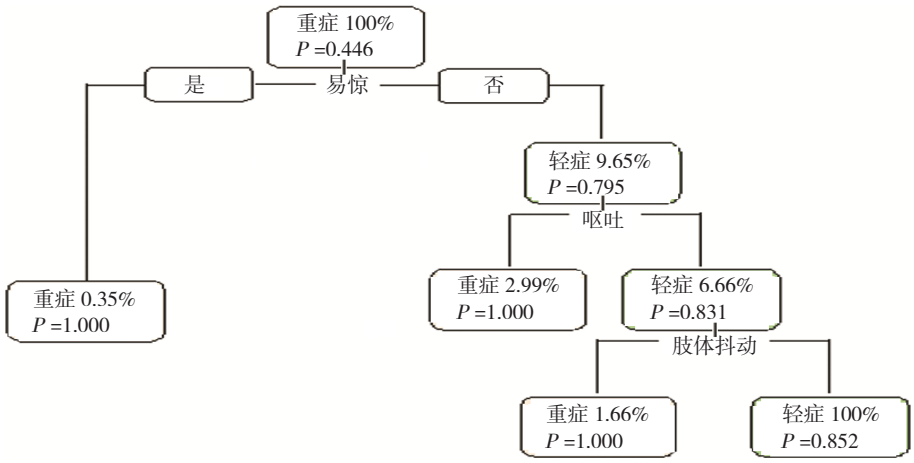
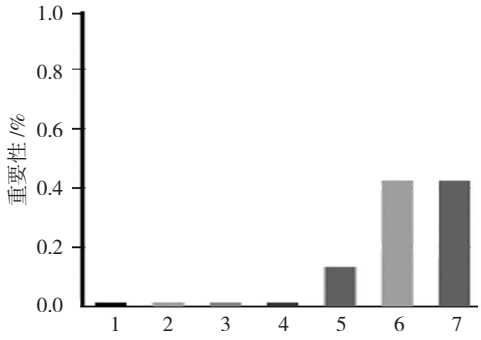


图 2　简易决策树



1：发热时间 >3 d；2：抽搐；3：中性比率 >75%；4：嗜睡；5：呕吐；6：易惊；7：肢体抖动

图 3　预测变量的重要性

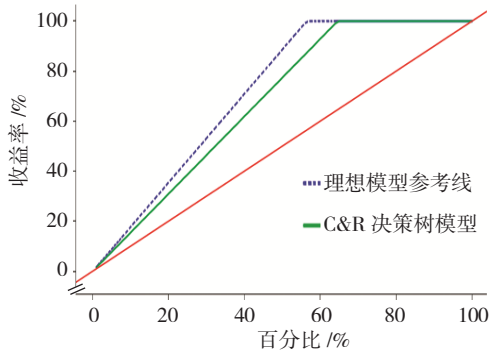


图 4　C&R 决策树模型的累计收益率

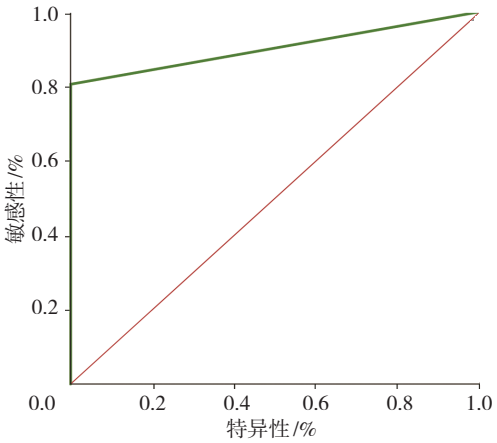


图 5　C&R 决策树模型的 ROC 曲线图

总数 603 例即为训练样本大小，而图中未显示的测试样本量为 235 例。Gini 系数作为分割点，它代表了目标变量组间的差异程度，其系数越小，组间差异越大。从根节点出发，计算每个节点的 Gini 系数，然后再计算 1 个系数的变化量，代表了异质性的下降，反应到决策树的图形上，显示为改进等于系数变化量。决策树从上往下分支，可以看到改进越来越小。从生成的预测变量重要性图中可以看出，HFMD 的分组与肢体抖动、易惊以及呕吐相关，而与其他变量则关系不大，再次验证了决策树模型的纳入变量选择。见图 1~3。

2.3　模型拟合效果

为了评价决策树模型的整体拟合效果，绘制模型累计收益图，在前期快速达到较高点后，快速趋于平稳，而本研究构建的决策树模型可以看到距离理想模型参考线较为接近；根据模型的预测结果绘制 ROC 曲线图，其曲线下面积为 0.903[(95%CI：0.878，0.927)，*P* =0.000]。模型的预测的准确性为 91.17%，敏感性为 84.36%，特异性为 96.25%，见图 4、5。

3　讨论

近年来，机器学习、人工智能等数据挖掘领域新兴技术蓬勃发展，随着大数据时代的到来，对庞大数据的处理分析也变得更为复杂和关键[4-5]。借助统计建模等领域与大数据端口的联结，可以通过分析数据提供决策向导，及时应对复杂的变化。决策树作为建模的一种算法，在处理分类问题方面具有精确性高、

输出结果容易理解等优势[6]。

目前，国内关于手足口病方面的临床研究，较为常见的是使用 Logistic 回归模型来筛选变量，预测病情[7-10]。Logistic 回归属于一种参数统计，其主要用于解决探讨危险因素以及预测发生概率等问题[11]。它属于一种线性模型，在分析主效应方面占优，但是无法处理各种变量的交互效应以及模型过度拟合的问题。而决策树属于非参数统计，可以很好地弥补 Logistic 回归模型的不足[12]。在处理交互相应方面较佳，能有效地避免模型的过度拟合，提升模型预测精确性的同时，提高适用广度[13]。此外，分类树模型可以很好地处理缺失值的情况，通过优化的算法，使得模型在实际使用中更高效便捷。本研究通过回顾性分析临床收集的患者资料，建立决策树模型，从众多待分析变量中筛选出预测变量，最后对模型进行评估，模型预测准确性为 91.17%，提示模型拟合效果较好。隋美丽等[14]的研究显示通过决策树筛选出精神差、手足抖动、易惊及热峰≥ 39℃共 4 个解释变量，预测准确性为 95.5%。ZHANG 等[15]的临床研究通过更高级的迭代决策树算法，构建的模型预测准确性为 92.3%。可以看出决策树模型在预测 HFMD 方面有一定的优势。本研究尚有一定不足，由于样本量偏小以及患者搜集时存在局限性，研究的人群能否很好地代表整体特征，以及模型的适用范围仍有待验证。任何模型都有其优势和不足，由于实际数据的复杂多样性，一种模型很难完全胜任，往往需要多种模型协调联合，通过优势互补，发挥功能。

综上所述，本研究提供了一种新的思路，通过决策树模型共纳入 3 个解释变量：易惊、呕吐及肢体抖动，模型预测精确度较高，对临床疾病诊疗有一定的辅助价值。后续仍需更多研究的深入开展，以挖掘出更佳算法，构建更优模型应用于临床，为重症手足口病的诊疗及预防做出更大的贡献。

参 考 文 献：

[1] FUJIMOTO T. Hand-foot-and-mouth disease, aseptic meningitis, and encephalitis caused by enterovirus[J]. Brain Nerve, 2018, 70(2): 121-131.

[2] WU X, HU S, KWAKU A B, et al. Spatio-temporal clustering analysis and its determinants of hand, foot and mouth disease in Hunan, China, 2009-2015[J]. BMC Infect Dis, 2017, 17(1): 645.

[3] 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.

[4] OBERMEYER Z, EMANUEL E J. Predicting the future-big data, machine learning, and clinical medicine[J]. N Engl J Med, 2016, 375(13): 1216-1219.

[5] 司家瑞 . 浅谈机器学习在医学大数据中的应用 [J]. 科技展望 , 2016, 26(23): 304.

[6] 程斐斐 , 王子牛 , 侯立铎 . 决策树算法在 Weka 平台上的数据挖掘应用 [J]. 微型电脑应用 , 2015, 31(6): 63-65.

[7] 单宝英 . 不同病原所致手足口病临床特征分析 [J]. 中国现代医学杂志 , 2016, 26(23): 119-122.

[8] 郑亚明 , 常昭瑞 , 姜黎黎 , 等 . 手足口病重症病例分析 : 基于全国手足口病监测试点数据 [J]. 中华流行病学杂志 , 2017, 38(6): 759-762.

[9] 徐树红 , 李青 , 顾胜利 , 等 . 贵州省 415 例重症手足口病临床分析 [J]. 中国现代医学杂志 , 2015, 25(4): 44-47.

[10] 朱韩武 , 谭徽 , 李成华 , 等 . 2010 ~ 2014 年郴州市手足口病重症和死亡病例的流行病学及病原学特征研究 [J]. 中国现代医学杂志 , 2015, 25(35): 83-87.

[11] 光琳 , 宗序平 . Logistic 模型的统计诊断 [J]. 江南大学学报 (自然科学版), 2012, 11(1): 113-117.

[12] 李泓波 , 白劲波 , 杨高明 , 等 . 决策树技术研究综述 [J]. 电脑知识与技术 , 2015, 11(24): 1-4.

[13] 薛允莲 . logistic 回归结合决策树技术在冠心病患者住院费用组合分析中的应用 [J]. 中国卫生统计 , 2015, 32(6): 988-989.

[14] 隋美丽 , 申远方 , 黄学勇 , 等 . 分类树模型在重症手足口病风险预测中的应用 [J]. 郑州大学学报 (医学版), 2015, 50(1): 20-25.

[15] ZHANG B, WAN X, OUYANG F S, et al. Machine Learning Algorithms for Risk Prediction of Severe Hand-Foot-Mouth Disease in Children[J]. Sci Rep, 2017, 7(1): 5368.

（李科 编辑）