·论　　著·

决策树在居民就诊影响因素研究中的应用*

刘海霞，钟晓妮，周燕荣，田考聪△
（重庆医科大学卫生统计教研室　400016）

摘　要：目的　了解影响重庆地区居民就诊服务的主要影响因素，满足更多居民卫生服务需求，提高卫生服务利用率。**方法**　针对重庆地区不同人群的影响因素，采取不同的卫生政策，构建影响居民就诊率的决策树模型。**结果**　调查的 11 570 名居民中，合计就诊人次为 2 447 人次，平均就诊次数 2.1 次，两周就诊率为 21.15%（城市为 12.58%、农村为 29.19%），高于全国平均水平，而各年龄段的就诊率呈现中间低两端高的趋势，各年龄段就诊率差异有统计学意义（$P<0.05$）；从决策树模型来看，此决策树共有 17 个节点，对应 17 条分类规则，树的根节点为职业类型，此变量对就诊率的影响最大，职业类型、年龄、居民类型、参保情况以及家庭年收入对居民就诊影响较大，且所选出的影响因素对不同人群的影响不同。**结论**　重庆地区居民就诊卫生服务利用相对较高，且不同人群的影响因素不同，因此，在制订卫生服务规划时应针对不同人群提出相应的卫生政策。

关键词：决策树；卫生服务利用；危险因素；决策
doi：10.3969/j.issn.1671-8348.2011.09.004　　**文献标识码：**A　　**文章编号：**1671-8348(2011)09-0840-03

Application of decision tree in study of factors affecting residential medical treatment service*

Liu Haixia, *Zhong Xiaoni*, *Zhou Yanrong*, *Tian Kaocong*△
(*Department of Health Statistios*, *Chongqing Medical University*, *Chongqing* 400016, *China*)

Abstract: Objective　To better know main factors affecting the treatment service of residents to meet the demands of health service of more residents and improve, health service the utilization. **Methods**　Aiming at the different affecting factors of different crowds in Chongqing area and adopting different health polices, the decision tree model affecting the rate of residential seeking medical care was constracted. **Results**　Of 11 570 residents receiving investigation, there were 2 447 person seeing the doctors in total, 2.1 times on average, the rate of 2-week seeking medical care was 21.15% (12.58% in city and 29.19% in rural areas), which was higher than national average. However, the seeking medical care rate for each age section showed the tendency that the middle part was lower than both ends. There were statistical differences for treatment rate of each age section. As far as decision tree model was concerned, there were 17 nodal points in the decision tree, corresponding to 17 classified rules separately. And the nodal points were professional types, indicating that the variant affected most on treatment rate, further more, the selected factors were different for different groups. **Conclusion**　The utilization for residents in Chongqing on health service treatment is relatively high, occupational type, age, resident types, insurance as well as family income have great impact on the treatment of residents, besides, influencing factors are different in different groups. Therefore, while formulating health service treatment plan, corresponding health policies should be put forward aiming at different groups.

Key words: decision trees; attendance rate; health service; risk factors; decision making

决策树（decision tree）是一种主要解决实际应用中分类问题的数据挖掘方法，具有可视效果好、分类率高、容易理解的特点[1]，在包括销售、电子商务、金融、生物医学、电信与客户关系管理等各领域中都有应用[2]。通过训练样本集建立目标变量关于各输入变量的分类预测模型，全面实现输入变量和目标变量不同取值下的数据分组，进而用于对新数据对象的分类和预测[3-4]。一个决策树由一系列节点和分支组成，而节点和子节点之间形成分支，节点代表着决策过程中所考虑的属性，而不同属性值形成不同分支，在决策树的叶节点得到结论，且从根节点到叶节点的每一条路径对应着一条决策规则[5]。目前决策树主要有 4 个研究方向[6]：(1)决策树技术与其他技术的结合应用；(2)寻找新的构建决策树的方法；(3)寻找更好的简化决策树的方法；(4)研究产生决策树的训练和检验数据的大小及特性与决策树特性之间的关系。本研究主要是通过构建决策树模型，发现影响重庆地区居民就诊卫生服务利用的主要影响因素，针对不同人群的影响因素，采取针对性的卫生政策，满足居民卫生服务需求，提高卫生服务利用率。

1　资料与方法

1.1　一般资料　资料来源于第 4 次国家卫生服务调查——重庆西部扩点地区调查数据，本次调查根据全国第 4 次卫生服务调查的政策指导与要求，采用多阶段整群随机抽样的方法，调查了 3 970 户 11 570 名居民的家庭健康状况，具体了解该地区居民健康状况及卫生服务需要、需求与利用情况。

1.2　方法　目前比较流行的决策树算法主要有 C4.5、分类与回归树（classification and regression tree，CART）[7] 和 χ^2 自动交互探测（chi-squared automatic interaction detector，CHAID）等[8]，这些算法主要是根据数据的特点建立相应的函数从而尽可能地正确分类所有的观察[9-10]。其中，C4.5 是目前最有影

*　**基金项目：**第 4 次国家卫生服务调查资助项目（卫办综函[2008]180 号）。　△　**通讯作者**，Tel：13983107373；E-mail：tkc5155@163.com。

响力的算法，是 ID3 的改进算法[11]，允许输入变量的类型可以是两分类、多分类名义型和区间型，目标变量可以是两分类或多分类名义型；也允许输入变量的类型可以是名义型、有序型，目标变量可以是名义型或区间型；还允许输入变量的类型可以是名义型、区间型，如果是有序型，则可以当作区间型变量处理，目标变量可以是两分类、多分类名义型，区间型和有序型。将是否就诊(i15)作为目标变量，可能影响患者就诊的因素性别(m5)、民族(m6)、年龄(x2)、婚姻状况(m9)、文化程度(m10)、职业类型(m12)、就业状况(m11)、医疗保险情况(m13)、居民类型(urb_rur)、家庭人口数(h1)、家庭收入(h12)、自感病情(i2)和是否患慢性病(m18)作为自变量，构建决策树模型。其中，将家庭收入和年龄两个自变量进行离散化，家庭收入按四分位数间距分为低、中、高等收入，年龄分为 8 个等级(0～4、>4～14、>14～24 、>24～34、>34～44 、>44～54、>54～64、>64)。

1.3　统计学处理　采用 SAS8.1 和 SPSS17.0 统计软件分析。本文所用调查数据由课题组成员进行核查并录入建立 EPI 数据库，通过 SAS8.1 将各个数据库和所需变量进行整理，最后将 SAS 文件导入 SPSS17.0 数据库。以 $P<0.05$ 为差异有统计学意义。

2　结　　果

2.1　居民就诊情况　本次调查共 3 970 户 11 570 名居民，其中农村和城市各 1 985 户，分别为 5 968、5 602 名居民，男性占 49.2%，女性占 50.7%。所调查的居民中合计两周就诊者为 2 447人次，平均就诊次数 2.1 次，两周就诊率为 21.15%(城市为 12.58%、农村为 29.19%)，其中男女就诊率分别为 18.41%、23.84%，差异有统计学意义($P<0.05$)；与 2003 年相比就诊率虽有所提高，但是两周患病未就诊所占比例却达到 56.2%，与 2003 年相比反而有上升趋势。两周患病治疗者选择的治疗方式：有 29.6%(其中 96.4%采用口服药物治疗)选择自我治疗，只有不到一半的患者选择两周内就诊，且大部分选择门诊卫生室/站和社区卫生服务中心就诊，分别占39.1%、28.8%。

2.2　居民就诊影响因素决策树分析　就诊卫生服务利用影响因素的决策树模型，是根据数据特点选择 CART 树增长法，过程中进行树的修剪以自动控制树的过增长。通过构建决策树模型得出，树的根节点依据职业类型划分，树状图的其他节点还包括居民类型、年龄、医疗保险、文化程度和家庭人口数，共 5 层、17 个节点，见图 1。

模型构建过程中，对各解释变量的重要性进行了排序，重要性大，说明增加该变量进入决策树时，整个系统不确定程度减少得多，各解释变量重要性排序见表 1。模型效果通过误分率(误分率是被分错的例数占全部例数的比例)来评估，训练集和验证集的误分率分别为 0.01 和 0.017。

表 1　就诊卫生服务利用各解释变量的重要性排序

排序	变量标签	自变量	重要性	标准化重要性
1	职业类型	m12	0.136	1.00
2	年龄	x2	0.114	0.842
3	居民类型	urb_rur	0.068	0.503
4	是否患慢性病	m18	0.063	0.464

表 1(续)　就诊卫生服务利用各解释变量的重要性排序

排序	变量标签	自变量	重要性	标准化重要性
5	文化程度	m10	0.058	0.431
6	就业状况	m11	0.046	0.342
7	家庭人口数	h1	0.039	0.289
8	医疗保险情况	m13	0.035	0.256
9	婚姻状况	m9	0.033	0.246
10	自感病情	i2	0.018	0.135
11	家庭年收入	h12	0.012	0.880
12	性别	m5	0.004	0.310
13	民族	m6	0.003	0.210

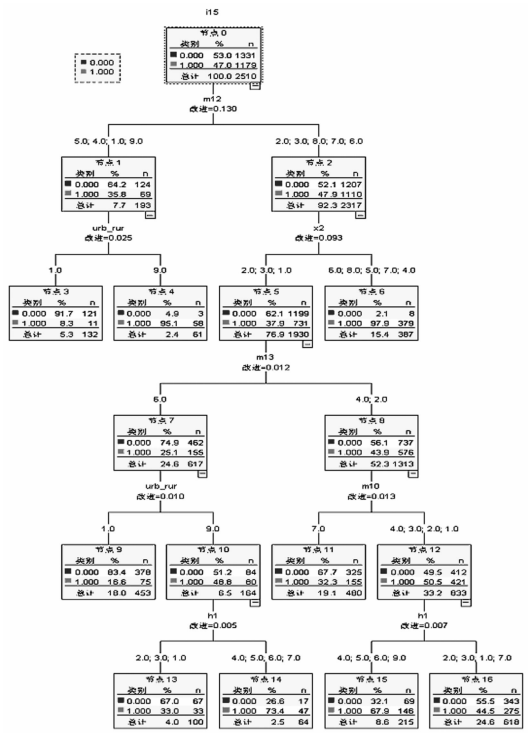

图 1　居民就诊卫生服务利用决策树模型

3　讨　　论

通过分析，模型的根节点为“职业类型”，说明该解释变量对就诊卫生服务利用的影响很大，从决策树模型可以看出，所选出的影响因素对不同人群的影响不同，如年龄对农民、非农户产业工人和一般技术人员等职业的人群影响较大，而居民类型对机关、企事业单位管理者、商人则影响较大。建议卫生部门针对不同的职业人群制定相应的卫生法规，同时加强农民、工人等人群的健康教育，提倡健康的生活方式和工作习惯。从年龄这个分支来看，各年龄层呈现中间低两端高的就诊现状，针对这一现状，建议卫生部门加强老年人和婴幼儿的健康保健管理，提倡健康的生活方式，控制慢性病的发病率。从居民类型这个分支来看，农村的就诊率相对较高，这与新型农村合作医疗的开展有关，相对的，城市就诊率则较低，因此政府应该完善城镇居民医疗保险制度和加强城市基层医疗机构建设，改善基层医疗服务条件[12]，提高城市社区卫生服务质量，建立社区卫生服务网，以步行 15 min 距离作为社区卫生中心配置依据，

并完善对社区卫生服务中心的综合管理[13]，充分发挥社区卫生服务站的职能，提高居民就诊卫生服务需求与利用率。

重庆地区居民就诊率为 21.15%，高于全国平均水平[14]，一方面说明重庆地区居民就诊率高，同时也说明重庆地区居民卫生服务需求量较大。但是该地区应该就诊而未就诊的患者中，46.6%的患者是因为经济困难，说明看不起病的现象依然存在，因此卫生部门在制定卫生政策时，首先应该保证居民看得起病；同时，从该地区居民就诊机构选择的流向来看，卫生室和乡镇卫生院/社区卫生服务中心等基层医疗机构的利用率相对较高，因此，应该加强基层医疗机构建设，合理分配卫生资源，比如实施和完善乡村一体化管理，这对于改善和加强村卫生室和乡镇卫生院的功能、改善农村居民就医难和提高农村居民的卫生服务需求和利用都发挥着重要的作用；相对应的，在城市加强社区卫生服务中心建设[15]，对于优化配置卫生资源、方便群众就医和提高卫生服务利用率具有重大意义。

参考文献：

[1] 徐蕾，贺佳，孟虹，等. 决策树技术及其在医学中的应用[J]. 数理医药学杂志，2004，17(2)：161-164.

[2] 李成. 数据挖掘技术的应用探析[J]. 内江科技，2008，29(6)：46.

[3] 靳淑敏，张翠肖，孙珊珊. 决策树技术及其在药物治疗中的应用[J]. 科技情报开发与经济，2008，22(18)：164-166.

[4] Dong M，Kothari R. Look-ahead based fuzzy decision tree induction[J]. IEEE Transactions on Fuzzy Systems，2002，9(3)：461，468.

[5] 王玉珍. 基于数据挖掘的决策树方法分析[J]. 电脑开发与应用，2007，20(5)：64-66.

[6] Jiawei Han，Micheline Kamber. 数据挖掘概念与技术[M]. 范明，孟小峰，译. 北京：机械工业出版社，2001：6-9.

[7] Breiman L，Friedman JH，Qlshen RA，et al. Classification and regression trees：modern applied statistics with S-plus[M]. 2nd ed. California：Wadsworth international group，1984：6-9.

[8] Jordan MI. Learning in graphical models[M]. Cambridge (Massachusetts)：MIT Press，1998：7-8.

[9] 中国人民大学统计学系数据挖掘中心. 数据挖掘中的决策树技术及其应用[J]. 统计与信息论坛，2002(2)：4-10.

[10] 但小容，陈轩恕，刘飞，等. 数据挖掘中决策树分类算法的研究与改进[J]. 软件导刊，2009，9(8)：41-43.

[11] Quinlan JR. Induction of decision tree[J]. Machine Learning，1986，1(1)：81-106.

[12] 李林. 门诊就诊影响因素调查[J]. 中华医院管理杂志，2000，8(16)：499-450.

[13] 刘朝杰，David Legge. 中国城市社区卫生服务政策分析[J]. 中国全科医学，2007，10(19)：1579-1583.

[14] 中华人民共和国卫生部卫生统计信息中心. 国家卫生服务研究——第 4 次卫生服务调查报告[M]. 北京：中国协和医科大学出版社，2009：37.

[15] 梁万年，饶克勤，常文虎，等. 卫生事业管理学：社区卫生服务管理[M]. 北京：人民卫生出版社，2003：294-308.

（收稿日期：2010-09-10 修回日期：2011-01-10）

（上接第 839 页）

[2] 姜若萍，傅民魁. 安氏Ⅱ类 1 分类错殆患者亲子间相似性的个体研究[J]. 中华口腔医学杂志，2001，36(2)：143-145.

[3] 姜若萍，傅民魁. 安氏Ⅱ类 1 分类错殆的遗传特征初探[J]. 现代口腔医学杂志，2001，15(5)：368-370.

[4] 刘继光，李晓光，王曦，等. 成人安氏Ⅱ^1 与Ⅱ^2 类错殆颅面特征对比研究[J]. 黑龙江医药科学，2008，31(3)：48-49.

[5] 傅民魁. 口腔正畸学[M]. 5 版. 北京：人民卫生出版社，2007：1.

[6] Basdra EK，Kiokpasoglou M，Stelliziig A. The Class Ⅱ Division 2 craniofacial type is associated with numerous congenital tooth anomalies[J]. Eur J Orthod，2000(5)：529-535.

[7] Walkow TM，Peck S. Dental archwidthin class Ⅱ Division 2 deepbite malocclusion[J]. Am J Orthod Dentofac Orthop，2002，122(6)：608-613.

[8] 范春香，吴丽萍. 安氏Ⅱ类 2 分类上切牙内倾的相关因素研究[J]. 口腔医学，2008，28(4)：3-6.

[9] 谢荣敏，秦朴，杜跃华. Angle′Ⅱ类 2 分类错畸形牙列指数的测量分析[J]. 重庆医科大学学报，2009，34(3)：368-370.

[10] 张晓歌，杨帆，陈琳，等. 安氏Ⅱ类一家系分析[J]. 华西口腔医学杂志，2010，7(2)：219-220.

[11] 姚宁，吴燕平，顾永佳. 安氏Ⅱ^2 类青少年不拔牙矫治前后的软硬组织变化[J]. 口腔医学，2009，29(7)：367-368.

[12] 冯驭驰. 安氏Ⅱ类 2 分类青少年不拔牙矫治前后的硬组织变化[J]. 口腔正畸学，2008，15(1)：30-33.

[13] Falconer DS. The inheritance of liability to certain disease estimated from the incidence among relatives[J]. Ann Hum Genet，1965，29(1)：51-76.

[14] 究匡正，宋岩，匡艳. 煎饼主食地区错殆畸形及内倾性深覆殆调查研究[J]. 广东牙病防治，2010，18(1)：33-35.

[15] Fernando DS，Mackay TFC. Introduction to Quantitative Genetics[M]. 4th ed. London：Longman，1999：40-45.

[16] Cui JJ，Li WuL，Mei LX. The study on the PAX9 related with oligodontia[J]. Int J Stomatol，2008，35(1)：38-40.

（收稿日期：2010-08-10 修回日期：2010-10-15）