·论 著·

基于信息熵的决策树在慢性胃炎中医辨证中的应用

徐 蕾[1], 贺 佳[1*], 孟 虹[1], 王忆勤[2], 贺宪民[1], 范思昌[1], 郎庆波[2]
(1. 第二军医大学卫生勤务学系卫生统计学教研室, 上海 200433; 2. 上海中医药大学基础医学院, 上海 200032)

[摘要] **目的**: 探讨基于信息熵的决策树在慢性胃炎中医辨证分型中的应用。**方法**: 采用 bootstrap 方法对 406 例样本进行扩增以满足数据挖掘对样本量的要求, 采用基于信息熵的决策树 C4.5 算法建立中医辨证模型。**结果**: 决策树 C4.5 算法筛选出对中医辨证分型有意义的 26 个因素并对其重要性进行排序; 产生清楚易懂可用于分类的决策规则; 建立辨证模型, 模型分类符合率为: 训练集 83.60%, 验证集 80.67%, 测试集 81.25%; 模型区分各类证型的灵敏度和特异度也较高。**结论**: 决策树 C4.5 算法建立的模型效果较好, 可应用于慢性胃炎中医证型的鉴别诊断。

[关键词] 信息熵; 决策树; 中医; 数据挖掘
[中图分类号] R 259.733; R 311 [文献标识码] A [文章编号] 0258-879X(2004)09-1009-04

Application of decision tree based on entropy in traditional Chinese medicine symptom analysis of chronic gastritis

XU Lei[1], HE Jia[1*], MENG Hong[1], WANG Yi-Qin[2], HE Xian-Min[1], FAN Si-Chang[1], LANG Qing-Bo[2] (1. Department of Health Statistics, Faculty of Health Services, Second Military Medical University, Shanghai 200433, China; 2. School of Preclinical Medicine, Shanghai University of Traditional Chinese Medicine, Shanghai 200032)

[**ABSTRACT**] **Objective**: To explore the application of decision tree based on entropy in traditional Chinese medicine (TCM) symptom analysis chronic gastritis. **Methods**: Bootstrap methods were used to multiply 406 cases for data mining, and models for TCM symptom were built using decision tree of C4.5. **Results**: Twenty-six important factors were selected and ranked according to the importance; readable diagnostic rules were produced and a model was built, of which the correctly classified rate of training set, validation set and test set were 83.60%, 80.67% and 81.25%, respectively. The sensitivity and specificity of the model to differentiate the TCM symptom was good. **Conclusion**: The model built by C4.5 is satisfactory and can be used for differentiating diagnosis of chronic gastritis with TCM symptom.

[**KEY WORDS**] entropy; decision tree; traditional Chinese medicine; data mining
[Acad J Sec Mil Med Univ, 2004, 25(9): 1009-1012]

* 中医对于慢性胃炎有很好的疗效, 但是如何从已积累的临床资料中提炼出蕴含的信息, 并结合已有的中医辨证准则[1]确定慢性胃炎患者的中医证型, 以达到辨证施治目的便成为中医研究者颇为关注的问题。本文尝试将数据挖掘决策树技术用于慢性胃炎中医证型的分类, 以考察这一方法用来解决此类问题的效果。

1 资料和方法

1.1 资料来源 本研究资料来自与上海中医药大学的合作研究项目"慢性胃炎中医证型分析"的 406 例调查资料, 共计 88 项指标。其中证型指标是本次研究的目标变量, 根据中医药大学课题组主要研究成员(由 4 所附属医院的 4 名主任医师和临床诊断学教研室 5 名副高以上的专家组成)的建议, 将证型分为 5 类: 第 1 类脾胃湿热、第 2 类湿浊中阻、第 3 类脾胃气虚或脾胃虚寒、第 4 类肝胃气滞或肝胃郁火、第 5 类湿浊中阻兼脾胃气虚。该资料有 4 个明显特征: (1)指标多, 指标间关系复杂; (2)88 项指标中存在计量、两分类、多分类、等级 4 种量度; (3)部分指标(年龄、体质量等)存在缺失值; (4)应变量(即本次分析的目标变量)是多分类名义型指标。

1.2 方法 本研究采用数据挖掘技术中的决策树方法, 构建以信息熵减少为特征的决策树分类模型, 整个过程在 SAS 8.2 Enterprise Miner(以下简称 SAS/EM)中实现。为了满足数据挖掘对数据量的要求, 研究在充分利用原有样本信息的基础上, 采用 bootstrap 方法对原始数据进行扩增至 2 000 例病例。通过对扩增前后数据不同类型变量分布情况进行考察, 发现数据吻合情况良好。所以, 将扩增病例作为本次研究数据进行决策树模型的训练。

* [基金项目] 国家中医药管理局实验室建设项目
[作者简介] 徐 蕾(1978-), 女(汉族), 硕士生
E-mail: lotusxulei@hotmail.com
* Corresponding author. E-mail: hejia@smmu.edu.cn

© 1994-2007 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

在 SAS/EM 软件中, 首先将 2 000 例病例引入, 然后分配给训练集、验证集、测试集各 50%、30%、20% 的样本量, 3 个数据集起的作用分别是训练模型、纠正模型、评价模型, 最后进行决策树参数的设置。SAS/EM 软件中可近似实现 C4.5、CART 和 CHAID 3 种决策树算法, 本研究采用基于信息熵的 C4.5 算法, 即以信息熵的减少 (entropy reduction) 作为树分裂准则, 其基本原理同于 Quinlan 的 ID3 算法[2]。按照软件中算法的要求, 将决策树分支的个数定义为 8 个 (资料中多分类名义变量的最多分类数), 为防止树生长"繁茂"而拟合过多噪声, 通过定义树深 6 层以及节点分裂时最少需要 10 个病例等手段对树进行预修剪, 而模型的评价采用正确分类百分比 (proportion correctly classifid) 指标, 选择子树时以具有最优评价值 (best assessment value) 且子树规模最小为准。一系列参数设置完毕之后, 便开始训练决策树。

2　结　果

2.1　筛选变量　决策树不同于神经网络建模方法的一个很明显的区别就是能筛选出对预测目标变量分类有重要意义的变量。本研究共筛选出 26 个对慢性胃炎中医辨证分型有意义的变量, 具体见表 1。

表 1　重要变量排序列表

Tab 1　Rank list of important variables

Rank	Variable	Label	Importance
1	X67	Greasy fur	1
2	X61	Thin fur	0.832 0
3	X52	Complexion	0.802 8
4	X11	Property of stomach pain	0.700 7
5	X36	Weakness	0.649 4
6	X34	Degree of stomach pain	0.387 4
7	X45	Halitosis	0.380 3
8	X69	Scattering of tongue covering	0.378 9
9	X10	Abnormal number of stool	0.364 6
10	X47	Stomach distention	0.352 9
11	X9	Hiccough	0.348 1
12	X27	Tongue color	0.340 6
13	X35	Cold and heat	0.316 2
14	X24	Belching	0.310 8
15	X49	Sensation of defecation	0.305 9
16	X15	Stagnation of *qi* and blood stasis	0.299 7
17	X13	Deficiency and excess	0.293 0
18	X55	Thirst	0.292 7
19	X44	Insomnia	0.243 2
20	X12	Dreaming often	0.240 7
21	X48	Feses quality	0.234 1
22	X40	Palpitation	0.201 1
23	X23	Nausea and vomiting	0.198 6
24	X50	Quantity of urine	0.188 7
25	X18	Anorexia	0.179 2
26	X16	Gastric upset	0.163 8

2.2　产生的规则　所谓规则, 就是从决策树根节点到叶节点对应的路径, 每 1 个节点处 (包括内部节点) 都对应着 1 个筛选出的重要变量。本研究决策树共有 126 个叶节点, 对应 126 条分类规则。下面列出的是决策树根节点到记号为 203 叶节点 (叶节点的序号与叶子在整棵决策树的位置有关) 对应的规则:

IF　X27　EQUALS　1
AND　X35　EQUALS　2
AND　X52　EQUALS　3
AND　X11　IS　ONE OF: 1、9
AND　X61　EQUALS　1
AND　X67　EQUALS　1
THEN
NODE : 203
N : 20
5 : 80.0%
4 : 10.0%
3 : 0.0%
2 : 10.0%
1 : 0.0%

2.3　正确分类率变化　随着决策树的增长即叶子数目的增加, 训练集 (TRAIN) 和验证集 (VALID) 的正确分类率不断地升高, 验证集主要用来校正训练集的拟合, 防止异常值、噪声等的影响造成拟合过度。当决策树叶子数目增长到约 105 左右时, 两者的正确分类率基本相当。以后, 验证集的正确分类率开始下降, 训练集的正确分类率继续上升, 两者距离逐渐增大, 但均衡来看两者平均的正确分类率仍在上升, 当叶子数目增长到 126 个时, 达到了最高点。此时, 决策树剪枝完成, 模型构建完毕。

2.4　决策树评价　所建慢性胃炎中医辨证模型可以采用正确分类百分比指标或误分率指标进行评价。误分率 (misclassification rate), 即被分错的例数占全部例数的百分比, 3 个不同样本量数据集的误分率分别为: 训练集 16.40%, 验证集 19.33%, 测试集 18.75%, 较小的误分率说明预测模型的性能较好。也可利用 SAS/EM 中专门的模型评价方法对所建模型进行评价, 结果一致。

2.5　模型检验　为了检验模型预测实际资料的效能, 本研究以原始的 406 例慢性胃炎患者的中医证型作为待预测对象, 从而利用所建决策树模型实现分类预测, 实际结果见表 2。经过计算, 得到模型在不同中医证型上的判断指标, 见表 3。

© 1994-2007 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

表 2　预测后分类与实际分类的对应(仅列出部分结果)

Tab 2　Accordance of predictive and actual classification (only partly)

NUM	F- HYZX	I- HYZX	P- HYZX5	P- HYZX4	P- HYZX3	P- HYZX2	P- HYZX1
1	1	1	1. 000 0	0. 000 0	0. 000 0	0. 000 0	0. 000 0
2	4	4	0. 333 3	0. 666 7	0. 000 0	0. 000 0	0. 000 0
3	4	4	0. 000 0	1. 000 0	0. 000 0	0. 000 0	0. 000 0
4	4	4	0. 000 0	1. 000 0	0. 000 0	0. 000 0	0. 000 0
5	3	4	0. 000 0	0. 857 1	0. 142 9	0. 000 0	0. 000 0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
402	2	2	0. 142 9	0. 000 0	0. 000 0	0. 714 3	0. 142 9
403	1	2	0. 000 0	0. 000 0	0. 000 0	0. 571 4	0. 428 6
404	2	2	0. 142 9	0. 000 0	0. 000 0	0. 714 3	0. 142 9
405	3	3	0. 000 0	0. 000 0	1. 000 0	0. 000 0	0. 000 0
406	3	3	0. 000 0	0. 000 0	1. 000 0	0. 000 0	0. 000 0

NUM: Original case number; F- HYZX: Actual classification; I- HYZX: Predictive classification; P- HYZX5: Probability of classification 5; P- HYZX4: Probability of classification 4; P- HYZX3: Probability of classification 3; P- HYZX2: Probability of classification 2; P- HYZX1: Probability of classification 1

表 3　模型的评价指标

Tab 3　Evaluation indices of model

Classification	Sensitivity	Specificity	Mis-diag-nosis rate	Miss-diag-nosis rate	Youden's index
1	83. 78	89. 83	10. 17	16. 22	73. 61
2	77. 64	96. 45	3. 55	22. 36	74. 09
3	85. 71	96. 43	3. 57	14. 29	82. 14
4	87. 84	96. 39	3. 61	12. 16	84. 23
5	74. 70	95. 67	4. 33	25. 30	70. 37

通过比较发现, 模型进行中医辨证分型的特异度比灵敏度高, 即误诊率较低, 漏诊率较高。综合灵敏度与特异度来看, 5 类证型之间诊断准确率有差别, 第 4 和第 3 类较高, 而其他 3 类较低。中医证型的量化与标准化问题是中医研究的热点, 量化与标准化不仅与客观的技术标准有关, 还与中医研究人员自身对于证型诊断标准的把握度有很大的关系[3]。因此判断得到的证型有很大的主观性, 这使得研究所得到的资料不一定完全反映真实情况。但从总体看, 所建模型灵敏度与特异度均较高, 可以运用模型进行预测。

3　讨　论

决策树筛选出的变量按照相对于第 1 个变量的重要性进行排序, 第 1 个变量的重要性记为 1。某变量的重要性与采用该变量作为分裂变量进行分裂时使目标分类不确定性减少的程度相关, 重要性大, 则目标分类不确定性减少的程度大; 反之亦然。本研究决策树共筛选出 26 个变量, 主要是苔质腻、苔质薄、面色、胃脘疼痛性质、大便便次异常、胃脘胀、口气等, 与临床实际情况基本相符。

决策树建立的规则是指导决策树分类的重要依据, 未知中医证型的病例按照规则归属到已知的证型中, 从而完成证型的预测。规则以"IF…THEN…"语句形式陈述, 最末 1 个 IF 对应的变量位于决策树的根节点。前述规则实例的含义为: 如果某人苔质(X67)腻、苔质(X61)薄、胃脘疼痛性质(X11)为冷痛或其他痛、面色(X52)萎黄、口气(X35)较重、面近时气秽触鼻、无呃逆(X27)发生, 那么他有 80% 的可能发生湿浊中阻兼脾胃气虚, 10% 的可能发生肝胃气滞或脾胃虚寒, 10% 的可能发生湿浊中阻。这样的结果可能有些绝对, 因为模型不可避免地会模拟噪声, 但结果本身无疑为临床判断提供了重要的参考价值。

决策树 C4. 5 算法的前身 ID3 算法只适用于应变量为两分类的资料, 发展到 C4. 5 算法时, 已经能够处理应变量为多分类名义型变量的资料。而且, C4. 5 算法能有效地处理缺失值而勿需对缺失值进行填充, 避免了因剔除缺失病例而造成的数据量减少或填充方法不当而引起的数据失真。运用决策树 CHAID 算法建模时, 如果自变量有连续性变量, 需将其首先进行分段以数量化。CART 算法是典型的两分类树, 更适合处理应变量为两分类的资料。所以本研究最终确定以 C4. 5 算法建立决策树模型。决

© 1994-2007 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

策树建模有着其他方法无法比拟的优点, 可以清晰地显示出哪些变量对于预测比较重要; 可以生成容易理解的规则; 可以同时处理连续性、两分类和多分类、有序性自变量; 计算量相对来说不大[4]。但是, C4.5 算法也有其缺点, 它在分裂时有偏向于取值较多的变量的毛病[2], 即多分类变量较二分类变量更易于被选入模型, 因本资料二分类变量很少, 故可忽略算法的缺陷。

[参 考 文 献]

[1] 燕忠生, 李恒谟, 周运茂, 等. 114 例初诊为慢性胃炎患者的临床资料分析[J]. 甘肃中医学院学报, 1999, 16(4): 27-29.

[2] 史忠植 主编. 知识发现[M]. 北京: 清华大学出版社, 2002. 21-45.

[3] 高月求, 王灵台. 慢性乙型肝炎中医证型研究探讨[J]. 中国中医基础医学杂志, 2003, 9(8): 600-601.
Gao YQ, Wang LT. Discussion and study on Chinese medicine syndrome types of chronic hepatitis B[J]. *Zhongguo Zhongyi Jichu Yixue Zazhi* (*Chin J Basic Med Trad Chinese Med*), 2003, 9(8): 600-601.

[4] 刘　昆, 刘业政. 基于决策树的医疗数据分析[J]. 计算机工程, 2002, 28(2): 41-43.
Liu K, Liu YZ. Analysis of medical treatment data bases on decision tree[J]. *Comput Engineer*, 2002, 28(2): 41-43.

[收稿日期] 2004-01-27　　[修回日期] 2004-04-19
[本文编辑] 尹　茶

·短篇报道·

桂林地区南蛇藤总生物碱的含量测定

Determination of general alkaloids from *Celastrus orbiculatus* **Thunb. in Guilin area**

张应辉, 阳丽华, 李传枚(解放军第 181 医院药剂科, 桂林 541002)

[关键词] 南蛇藤; 总生物碱; 含量; 测定
[中图分类号] R 282.71　　[文献标识码] B　　[文章编号] 0258-879X(2004)09-1012-01

南蛇藤(*Celastrus orbiculatus* Thunb.)为卫矛科植物, 能祛风胜湿, 行气散血, 具有显著的抗炎作用[1]。我院自 20 世纪 70 年代开始研究应用南蛇藤(全株, 灌阳产)[2]治疗各种类风湿性关节炎数千例, 总有效率达 93%。

1　材料和方法

1.1　材料和试剂　分别采自灌阳、龙胜、兴安三地的生药全株, 经本院李传枚副主任药师鉴定为卫矛科植物南蛇藤(*Celastrus orbiculatus* Thunb.)。所用试剂均为分析纯。

1.2　方法[3]　将南蛇藤生药的根、茎和枝叶分别切片, 晒干, 粉碎, 于 80℃干燥至恒重备用。分别取粉末 20 g, 加氨水 6 ml, 拌匀, 用索氏提取器加 180 ml 氯仿提取 4 h, 将提取液移至分液漏斗, 用盐酸(0.5 mol/L)振摇 4 次(20、15、15、10 ml), 合并酸液, 用 15 ml 氯仿洗涤, 收集酸液, 过滤, 滤液加氨水至 pH 值为 9, 加氯化钠饱和, 用氯仿振摇提取 4 次(40、30、20、20 ml), 合并氯仿液, 用饱和氯化钠溶液洗涤 3 次, 每次 5 ml, 合并洗液, 再用 10 ml 氯仿振摇提取, 合并, 置水浴上蒸干, 于 105℃干燥至质量恒定, 准确称质量。

2　结果和讨论

灌阳、龙胜、兴安三地产南蛇藤中总生物碱含量分别为(0.292±0.004)%、(0.283±0.002)%、(0.288±0.003)%, 无显著差别。南蛇藤的全根、全茎中的总生物碱含量显著高于枝叶部分, 分别为(0.510±0.013)%、(0.340±0.018)%、(0.022±0.002)% ($P<0.01$)。结果表明, 桂林周边地区的灌阳、龙胜、兴安三地产的南蛇藤总生物碱含量无明显差异, 可替代使用, 这有利于保护灌阳当地的生药资源。

南蛇藤根、茎总生物碱含量高于枝叶部分($P<0.01$), 但根、茎两部分总生物碱含量差别不大, 可以考虑用茎取代以往的全株入药。

[参 考 文 献]

[1] 江苏新医学院 编. 中药大辞典(下册)[M]. 上海: 上海人民出版社, 1997. 1563.

[2] 唐庆年, 褶炯华. 南蛇藤治疗风湿、类风湿性关节炎 500 例探讨[J]. 实用中医药杂志, 1998, 14(3): 20.

[3] 郭远强, 李　铣. 南蛇藤属植物化学成分研究进展[J]. 沈阳药科大学学报, 2003, 20(3): 226.

[收稿日期] 2004-04-26　　[修回日期] 2004-07-16
[本文编辑] 尹　茶

[作者简介] 张应辉(1970-), 男(汉族), 主管药师

© 1994-2007 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net