

DOI: 10.13288/j.11-2166/r.2018.03.006

思路与方法

基于贝叶斯网络构建慢性阻塞性肺疾病急性加重期痰热壅肺证症状与证候间关联模式

王至婉^{1,2,3}, 李建生^{2,3,*}, 陈楚湘⁵, 章亮⁵, 王翔⁶, 李海霞^{2,4}

(1. 河南中医药大学第一附属医院, 河南省郑州市人民路 19 号, 450000; 2. 呼吸疾病诊疗与新药研发河南省协同创新中心; 3. 河南省中医药防治呼吸病重点实验室; 4. 河南中医药大学; 5. 中国人民解放军信息工程大学; 6. 郑州航空工业管理学院)

摘要 目的 基于贝叶斯网络探讨慢性阻塞性肺疾病急性加重期症状对证候的贡献度, 构建症状与痰热壅肺证的关联模式。方法 采用全国九家省级中医院为期 2 年所调查的 1768 例患者数据建立痰热壅肺证数据库。数据预处理后导入贝叶斯网络。以痰热壅肺证为目标变量, 以决策树运行结果筛选出的变量为输入变量, 运用 IBM SPSS Modeler 14.2 软件实现贝叶斯网络算法, 选用评估及分析对 TAN 及马尔科夫毯两种模型所获得的结果进行比较分析。结果 痰热壅肺证因果关系如下: 1) 痰色黄、痰白黏、脉滑数、脉数、大便秘结、口渴喜冷饮; 2) 痰色黄、痰白黏、脉滑数、舌苔黄腻; 3) 痰色黄、痰白黏、脉滑数、舌苔黄、舌苔黄腻; 4) 痰色黄、痰白黏、脉滑数、舌质红。上述 4 条中, 共有的症状为痰色黄、痰白黏、脉滑数, 其中, 痰色黄对痰热壅肺证的贡献度最大, 为 0.72; 其次痰白黏, 为 0.61。马尔科夫毯模型显示, 口渴喜冷饮、大便秘结、舌苔黄腻对痰热壅肺证的贡献度均 > 0.7。结论 痰热壅肺证与症状间有 4 条因果关系, 痰色黄、口渴喜冷饮、大便秘结、舌苔黄腻对痰热壅肺证的贡献度较大。

关键词 慢性阻塞性肺疾病; 痰热壅肺证; 贝叶斯网络; 证候; 症状

贝叶斯网络是一种研究不确定问题的决策方法, 可以从数据中发现变量间的因果关系, 并用概率定量表示这些因果关系的强度; 此外, 它还可根据模型的概率集合, 运用相加规则、相乘规则和贝叶斯定理等概率计算规则计算任意感兴趣的概率, 从而帮助做出诊断^[1]。由于其良好的知识表达形式以及处理不确定问题的能力在医疗诊断领域也得到了越来越广泛的运用, 在中医证候领域也得到了初步应用。我们用贝叶斯网络以慢性阻塞性肺疾病(简称“慢阻肺”)急性加重期痰热壅肺证为例探讨症状对证候的贡献度, 构建症状与证候的关联模式。

1 资料与方法

1.1 临床资料

1.1.1 一般情况 以 2006 年 12 月至 2008 年 12 月开展的全国九家省级中医院调查所得的 1768 份

患者资料为数据来源, 其中男 919 例, 女 849 例。肺功能^[2] I 级者 645 例, II 级者 812 例, III 级者 311 例。≥60 岁者 1275 例, <60 岁者 493 例。城镇居民 1120 例, 农村居民 648 例。吸烟者 1034 例, 吸烟时间 8~59 年, 每天吸 ≥20 支者 564 例, <20 支者 656 例; 不吸烟者 734 例(包括戒烟者 289 例)。合并肺间质纤维化 75 例, 陈旧性肺结核 97 例, 高血压病 231 例, 冠心病 99 例, 肺源性心脏病 307 例, 中风 65 例, 糖尿病 261 例。

1.1.2 诊断标准 西医诊断标准参照 2002 年中华医学会呼吸病学分会颁布的《慢性阻塞性肺疾病诊治指南》中诊断标准^[3]。中医辨证标准依据中华中医药学会内科分会肺系病专业委员会 2012 年颁布的《慢性阻塞性肺疾病中医证候诊断标准(2011 版)》^[4]。

1.1.3 纳入及排除标准 纳入符合上述诊断标准及中医辨证标准且为急性加重期; 年龄 18~80 岁。排除神志不清、痴呆、各种精神病患者。

基金项目: 国家自然科学基金(81173201)

* 通讯作者: li_js8@163.com, (0371) 65676568

1.2 数据仓库的建立

数据的录入与修改由数据管理员负责。使用 Epidata 3.0 版软件, 将所收集资料进行数据审核、整理后, 由两个数据管理员独立进行双录入, 建立慢阻肺急性加重期临床调查研究数据库。数据预处理: 数据清洗、删除冗余数据、数据转换、处理缺失值、降维等。数据库中既有二分类变量又有等级变量。二分类变量如舌象、脉象, 按有、无变量赋值为 0、1; 等级变量如痰色黄、痰白黏, 按无、轻、中、重、极重变量依次赋值为 1、2、3、4、5。再将等级变量转化为二分类变量, 1 (无) 转化为 0, 2、3、4、5 转化为 1。

1.3 模型的建立

运用 IBM SPSS Modeler 14.2 软件实现贝叶斯网络算法。数据预处理后导入贝叶斯网络。以痰热壅肺证为目标变量, 以决策树运行结果筛选出的变量为输入变量, 选用评估及分析对 TAN 及马尔科夫毯两种模型进行比较分析。

将原始数据库经过数据预处理和数据转换等建立基于贝叶斯网络的数据挖掘仓库; 设计和实现贝叶斯网络算法, 探索包含以上算法程序的数据挖掘模型; 从数据仓库中随机选择不同病例样本, 进行参数调整及算法修改功能; 建立贝叶斯网络数据挖掘模型。

贝叶斯网络包括有向无环图 (directed acyclic graph, DAG) 和条件概率表 (conditional probability graph, CPG)。有向无环图是一个有向无环的图形结构, 其中每个节点代表一个变量, 节点之间的有向弧段反映了变量间的依赖关系, 指向节点 X 的所有节点称为 X 的父节点。条件概率表则给出了每个节点以其父节点为条件时的所有可能的条件概率^[4-5]。基于慢阻肺急性加重期证候临床研究数据库, 建立痰热壅肺证数据库。类型选用标志, 过滤后导入贝叶斯网络。以痰热壅肺证为目标变量, 以决策树运行结果筛选出的 11 个变量 (舌苔黄腻、痰色黄、口渴喜冷饮、舌苔黄、脉滑数、大便秘结、舌质红、痰白黏、痰多、咯痰不爽、脉数) 为输入变量, 所有变量均为二分类, 以 0, 1 表示。结构类型选用 TAN 模型及马尔科夫毯模型, 参数学习方法选用“对小单元格计数的贝叶斯调整”, 模式选用专家, 选用“包括特征选择预处理步骤”, 缺失值选用“仅使用完整记录”, 显著性水平为 0.01, 最大条件集大小为 5, 最大输入量为 15。选用评估及分析将对两种模型进行比较分析。

2 结果

2.1 TAN 模型

结构类型 TAN 的运行结果显示, 上述 11 个变量减为 9 个, 痰多、咯痰不爽被筛除掉。所得贝叶斯网络最左侧节点代表的是输出节点也就是痰热壅肺证, 它是其余节点的父节点。节点颜色深浅不一, 代表的是输入变量的重要性, 颜色越深就越重要。得出痰热壅肺证因果关系如下: 痰色黄、痰白黏、脉滑数、脉数、大便秘结、口渴喜冷饮; 痰色黄、痰白黏、脉滑数、舌苔黄腻; 痰色黄、痰白黏、脉滑数、舌苔黄、舌苔黄腻; 痰色黄、痰白黏、脉滑数、舌质红。

2.2 TAN 模型节点与父节点的条件概率

输出变量为父节点。贝叶斯网络用条件概率表达事件变量间关系的强度, 一个节点在其父节点的不同取值组合条件下, 取不同属性值的概率, 就构成了该节点的条件概率。条件概率表明了节点对父节点的贡献度。表 1 示, 痰色黄的父节点是痰热壅肺证, 条件概率是 0.72。其他节点均有两个父节点, 条件概率最高的是舌苔黄, 条件概率是 0.88, 其父节点为舌苔黄腻、痰热壅肺证; 其次条件概率大于 0.7 的有: 脉数、舌质红、口渴喜冷饮, 其父节点分别为痰热壅肺证、脉滑数, 痰热壅肺证、脉滑数, 痰热壅肺证、大便秘结。

表 1 慢性阻塞性肺疾病痰热壅肺证 TAN 模型条件概率

节点	父节点	条件概率 = 1
痰色黄	痰热壅肺证	0.72
痰白黏	痰色黄	0.34
	无痰色黄	0.61
脉滑数	痰白黏	0.44
	无痰白黏	0.31
脉数	无脉滑数	0.44
	脉滑数	0.86
舌苔黄腻	无脉滑数	0.47
	脉滑数	0.84
舌质红	无脉滑数	0.45
	脉滑数	0.86
舌苔黄	舌苔黄腻	0.88
	无舌苔黄腻	0.36
大便秘结	脉数	0.60
	无脉数	0.65
口渴喜冷饮	大便秘结	0.86
	无大便秘结	0.47

2.3 马尔科夫毯模型节点与父节点的条件概率

表 2 示, 痰白黏、口渴喜冷饮、舌苔黄腻、痰色黄的父节点只有一个, 为痰热壅肺证; 大便秘结父节点是痰热壅肺证、口渴喜冷饮。

表 2 慢性阻塞性肺疾病痰热壅肺证
马尔科夫毯模型的条件概率

节点	父节点	条件概率 = 1
痰白黏	痰热壅肺证	0.61
口渴喜冷饮	痰热壅肺证	0.71
大便秘结	痰热壅肺证	口渴喜冷饮 0.75
	痰热壅肺证	无口渴喜冷饮 0.29
舌苔黄腻	痰热壅肺证	0.73
痰色黄	痰热壅肺证	0.72

2.4 两种模型的比较

表 3 示, TAN 模型对于痰热壅肺证的分类正确率为 98.42%, 马尔科夫毯模型的正确率为 97.34%, 两种模型一致性比较为 97.68%。如果把两者结合起来判断, 那么整体的一致性对于痰热壅肺证的分类结果的正确率为 99.02%。

表 3 慢性阻塞性肺疾病痰热壅肺证 TAN 模型与
马尔科夫毯模型一致性比较

项目	正确/一致	错误/不一致	总计	正确率/一致率
TAN 模型	1740	28	1768	98.42%
马尔科夫毯模型	1721	47	1768	97.34%
TAN 与 马尔科夫毯模型的一致性	1727	41	1768	97.68%
一致性与痰热壅肺证	1710	17	1727	99.02%

3 讨论

贝叶斯网络由朴素贝叶斯分类法发展而来, 朴素贝叶斯分类是一种十分简单的分类算法, 认为输入变量的条件独立假设, 使得相应的计算非常方便, 但这是不符合实际情况的。为此, 人们开始探索各种有效的途径, 希望即能够直观表示变量的联合分布, 又便于分类预测的简化计算, 这就是贝叶斯网络, 包括 TAN 贝叶斯网络及马尔科夫毯网络^[1]。

TAN 贝叶斯网络是朴素贝叶斯网络的一种扩展, 其特点为所有输入变量节点与输出变量节点都有弧线连接, 这意味着对输出变量的分类预测是基于全体的输入变量的, 应考虑所有的因素, 尽管其中的某些输入变量可能对输出变量没有显著的影响; 输入变量之间存在弧线, 这意味着输入变量之间并非全部条件独立, 允许具有相互依赖关系; 对每个输入变量节点, 最多允许存在两个父节点, 其中一个为输出变量节点, 另外一个为输入变量节点^[1]。与 TAN 贝叶斯网络不同的是, 马尔科夫毯网络不再要求输出变量必须是所有输入变量的父节

点, 输出变量不再一定为根节点, 其上层允许有父节点。换句话说, 在马尔科夫毯网络结构中, 输入变量和输出变量具有完全相同的地位^[1]。

随着贝叶斯网络的发展, 在解决中医定量诊断中具有良好的应用前景, 预测效果较好, 可以体现出症状与证候之间的复杂关系, 这有助于确定证候诊断的标准和规范, 而且建立的证候诊断模型以概率形式给出诊断结果, 能有效辅助专家作出决策。一些学者运用贝叶斯网络在中医证候分类方面进行了探讨, 取得了初步成果^[6-11]。但目前尚存在诸多问题有待于解决, 主要表现在: 1) 缺少大样本的相关数据库做支撑进行数据挖掘; 2) 挖掘技术还不能完全融入并指导中医药临床实践, 尚不能有机地结合中医药传统理论对挖掘到的数据模式进行合理解释; 3) 对证候属性及诊断的挖掘尚处于探索阶段, 较少筛选出具有真正临床实际意义的症状与证候间的因果关系, 尤其是定量关系, 从而有效地指导临床实践。

本项目采用 TAN 模型及马尔科夫毯模型, 挖掘得到慢阻肺急性加重期痰热壅肺证的 DAG 及条件概率。结果显示, 痰热壅肺证因果关系如下: 1) 痰色黄、痰白黏、脉滑数、脉数、大便秘结、口渴喜冷饮; 2) 痰色黄、痰白黏、脉滑数、舌苔黄腻; 3) 痰色黄、痰白黏、脉滑数、舌苔黄、舌苔黄腻; 4) 痰色黄、痰白黏、脉滑数、舌质红。上述 4 条中, 共有的症状为痰色黄、痰白黏、脉滑数, 其中, 痰色黄对痰热壅肺证的贡献度最大, 为 0.72; 其次痰白黏, 为 0.61。TAN 模型中, 痰热壅肺且脉滑数的患者脉数的可能性为 0.86, 而马尔科夫毯模型中没有显示脉滑数对痰热壅肺证的条件概率, 提示脉滑数对痰热壅肺证的贡献度相对较小。马尔科夫毯模型显示, 口渴喜冷饮、大便秘结、舌苔黄腻对痰热壅肺证的贡献度均 >0.7。痰热壅肺证且舌苔黄腻的患者, 舌苔黄的可能性为 0.88, 从症状的角度看, 舌苔黄腻与舌苔黄关联最为密切。同理, 得到脉数、脉滑数, 舌苔黄腻、脉滑数, 舌质红、脉滑数, 口渴喜冷饮、大便秘结关联密切, 其条件概率均大于 0.7。

可见, 贝叶斯网络中 TAN 模型及马尔科夫毯模型均以因果关系和条件概率的形式表示症状对证候的关联及贡献度大小, 从而为证候诊断提供依据。但毕竟贝叶斯网络属于探索性知识挖掘方法之一, 两种或两种以上方法联合使用, 取长补短, 优势互补, 更能达到预期目的。

参考文献

- [1] 薛薇. 基于 SPSSModeler 的数据挖掘 [M]. 2 版. 北京: 中国人民大学出版社, 2016: 311-334.
- [2] 中华医学会呼吸病学分会慢性阻塞性肺疾病组. 慢性阻塞性肺疾病诊治指南 [J]. 中华结核和呼吸杂志, 2002, 25(8): 453-460.
- [3] 中华中医药学会内科分会肺系病专业委员会. 慢性阻塞性肺疾病中医证候诊断标准(2011 版) [J]. 中医杂志, 2012, 53(1): 177-178.
- [4] CAMPOS LDM, FEMÁNDEZ-LUNA JM, HUETE JF. Clustering terms in the Bayesian network retrieval model: a new approach with two term-layers [J]. Appl Sof Com, 2004, 4(1): 149-158.
- [5] RIBEIRO-NETO B, SILVA I, MUNTZ R. Soft Computing in Information Retrieval Techniques & Application [Z/OL]. (2000-01-01) [2017-10-10]. [https://link.springer.com/](https://link.springer.com/book/10.1007/978-3-7908-1849-9?no-access=true)
- [6] 孙文军, 曲淼, 唐后盛. 基于贝叶斯网络的广泛性焦虑症中医证候学规律研究 [J]. 北京中医药, 2014, 33(6): 403-405.
- [7] 吴荣, 聂晓燕, 王阶, 等. 基于贝叶斯网络的名老中医治疗冠心病辨证规律研究 [J]. 中国中医药信息杂志, 2010, 17(5): 98-99.
- [8] 张鑫, 陈波, 徐涛, 等. 基于贝叶斯网络的肺癌证候研究 [J]. 中国中医药科技, 2014, 21(6): 599-603.
- [9] 陈宝君, 包祖晓, 张丽, 等. 抑郁证肝阳虚证的宏观辨证指标研究 [J]. 中医药学报, 2011, 39(5): 5-7.
- [10] 徐璿, 许朝霞, 许文杰, 等. 基于贝叶斯网络原理的 835 例冠心病病例中医证候分类研究 [J]. 上海中医药杂志, 2014, 48(1): 10-13.
- [11] 吴宏进, 许家伦, 张志枫, 等. 基于数据挖掘的围绝经期综合征中医证候分类算法分析 [J]. 中国中医药信息杂志, 2016, 23(1): 39-42.

Constructing Symptom and Syndrome Association Mode of Acute Exacerbation of Chronic Obstructive Pulmonary Disease with Phlegm-heat Congesting Lung Syndrome: Based on Bayesian Network

WANG Zhiwan^{1,2,3}, LI Jiansheng^{2,3,4}, CHEN Chuxiang⁵, ZHANG Liang⁵, WANG Xiang⁶, LI Haixia^{2,4}

(1. First Affiliated Hospital of Henan University of Chinese Medicine, Zhengzhou, 450000; 2. Diagnosis and Treatment of Respiratory Diseases and New Drug Development Henan Collaborative Innovation Centre; 3. Key Laboratory of Respiratory Disease Prevention and Control by Chinese Medicine in Henan; 4. Henan University of Chinese Medicine; 5. The Chinese People's Liberation Army Information Engineering University; 6. Zhengzhou Management College Of Aircraft Industry)

ABSTRACT Objective To explore the contribution of symptoms to syndromes of chronic obstructive pulmonary disease acute exacerbation based on Bayesian network and construct the association mode of symptoms and phlegm-heat congesting lung syndrome. **Methods** Phlegm-heat congesting lung syndrome database was built adopting 1768 patients' data researched in 9 national provincial Chinese medicine hospitals for 2 years. Data were imported in Bayesian network after preprocessing. Taking phlegm-heat congesting lung syndrome as the target variable, taking the variables selected from the results of the decision tree as the input variables, Bayesian network algorithm was carried out by IBM SPSS Modeler 14.2 software. Comparing analysis was made on the results gained from TAN and markov blanket models using assessment and analysis. **Results** Phlegm-heat congesting lung syndrome causality showed below. First, Yellow phlegm, white sticky phlegm, slippery-rapid pulse, rapid pulse, constipation, thirst and liking cold drink. Second, Yellow phlegm, white sticky phlegm, slippery-rapid pulse and yellow greasy fur. Third, Yellow phlegm, white sticky phlegm, slippery-rapid pulse, yellow fur and yellow greasy fur. Forth, Yellow phlegm, white sticky phlegm, slippery-rapid pulse and red tongue. In the above 4 items, the common symptoms were yellow phlegm, white sticky phlegm and slippery-rapid pulse. Among them, yellow phlegm made the most contribution, 0.72. White sticky phlegm was the second, 0.61. Markov blanket model showed that the contribution of thirst and liking cold drink, constipation and yellow greasy fur to phlegm-heat congesting lung syndrome was more than 0.7. **Conclusion** There were 4 items of causality between phlegm-heat congesting lung syndrome and symptoms. Yellow phlegm, thirst and liking cold drink, constipation and yellow greasy fur made great contributions to phlegm-heat congesting lung syndrome.

Keywords chronic obstructive pulmonary disease; phlegm-heat congesting lung syndrome; Bayesian network; syndrome; symptom

(收稿日期: 2017-04-04; 修回日期: 2017-10-09)

编辑: 焦爽