基于 BP 神经网络与决策树的大肠癌虚实证型分类对比*

李金城 从 瑶 陈秋芬 陈春益 刘秀峰

(广州中医药大学 广州 510006)

〔摘要〕 采用二进制编码对症状数据进行量化，将专家归纳的 8 个证型分为虚实两证并赋值量化，建立基于 BP 神经网络与决策树的大肠癌虚实证型分类模型，结果显示 BP 神经网络分类模型较决策树分类模型更适合于非线性映射关系的处理。

〔关键词〕 大肠癌；虚实证型；BP 神经网络；主成分分析；决策树

〔中图分类号〕R-056 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673-6036.2017.05.014

Comparison on Classification of Excess and Deficiency Syndromes of Colorectal Cancer Based on BP Neural Network and Decision Tree

LI Jin-cheng, CONG Yao, CHEN Qiu-fen, CHEN Chun-yi, LIU Xiu-feng, Guangzhou University of Chinese Medicine, Guangzhou 510006, China

〔**Abstract**〕 The paper quantizes symptom data through binary coding, divides 8 syndromes summed up by experts into excess and deficiency syndromes, values and quantizes them, and establishes the model for classification of excess and deficiency syndromes of colorectal cancer based on BP neural network and decision tree. The result shows that BP neural network classification model is more applicable for the handling of the nonlinear mapping relation compared with decision tree classification model.

〔**Keywords**〕 Colorectal cancer; Excess and deficiency syndrome; BP neural network; Principal component analysis; Decision tree

1 引言

大肠癌是严重危及人类健康的常见恶性肿瘤之一，在我国其发病率和死亡率呈上升趋势[1]。大肠癌属中医“积聚”、“肠覃”、“脏毒”、“锁肛痔”、“便血”等范畴，中医药治疗大肠癌有确定的疗效，是我国治疗肿瘤的一大特色[2]。中医治疗的核心是辨证论治，近年来有关中医辨证论治大肠癌的文献报道较多，但对大肠癌辨证分型划分不一，没有一个统一、客观的标准。这不仅不利于中医临床科研协作，而且也严重影响了中医药治疗结直肠癌有效性的评价。研究指出：大肠癌的各判定指标具有模糊性、不确定性，指标数量较多，彼此之间还存在非线性关联性[3]。针对这样一个复杂的系统，本文依据 BP 神经网络擅长处理非线性关系以及决策树擅长处理非数值型分类的理论，尝试应用两种方法分别模拟证型辨证分类，基于大肠癌证型高阶、多维等特点构建大肠癌证型分类器，以期促进对大肠癌虚实证型的研究。

〔修回日期〕 2017-04-03

〔作者简介〕 李金城，本科生；通讯作者：刘秀峰，教授，硕士生导师。

〔基金项目〕 广州中医药大学薪火计划资助项目（项目编号：XH20160105）。

2 资料与方法

2.1 数据来源

以中国知网全文数据库（CNKI）、维普全文数据库（VIP）及万方数据知识服务平台的文献为数据源，采用回顾性方法收集肠癌患者病例 188 例，筛选出符合标准的有 106 例，咨询专家后分成 8 个证型，虚实两类，其中男 55 例，女 51 例，年龄为 25～80 岁，平均年龄为 55 岁。

2.2 数据纳入标准

经细胞学或病理学检查诊断为结直肠癌的患者以及大肠癌术后患者（经手术切除原发病灶、术后病程 3 个月内）等。

2.3 数据排除标准

合并有严重的心、肺、脑、肝、肾等疾病的患者或临床资料不全的患者等。

2.4 量化标准

2.4.1 症状量化标准 对收集的数据进行统计，首先排除在案例中只出现一次的因素后初步筛选出大肠癌患者所表现出的 53 项体征，包含腹泻、面色萎黄、脉虚、脉细、舌红、舌淡白、纳差、乏力等。症状量化标准依据《中医症状鉴别诊断学》（第 2 版）[4]，对症状进行初步分类。接着通过研究文献，咨询专家，将已经确定好的大肠癌常见症状进行赋值量化，有该症状者赋值 1，无该症状者赋值 0。数据赋值量化结束后，将数据录入 Excel。数据录入采取双人独立录入，进行一致性校验。

2.4.2 证型量化标准 依据《中华肿瘤治疗大成》、《中华中医药学会标准·肿瘤中医诊疗指南》及临床经验进行大肠癌证型术语规范化，比较分析辨证结果，再与专家进行讨论，最终形成一致意见。初步将多种证型术语规范为以下 8 种：湿热内蕴、气滞血瘀、瘀毒内阻、脾失健运、肝脾不调、脾肾阳虚、肝肾阴虚、气血亏虚。咨询专家后，进而将上述证型分为虚实两证，使用二进制编码，将已经确定好的大肠癌证型进行赋值量化，有该证型者赋值 1，无该证型者赋值 0。数据赋值量化结束后，将数据录入 Excel。数据录入采取双人独立录入，进行一致性校验。

2.5 研究方法

对数据进行主成分分析，主成分的选择标准定为 90%。

3 结果

3.1 主成分分析法结果

数据经过主成分分析后，前 28 个主成分的贡献率涵盖了总共 53 个指标的 90.763% 的信息，说明前面的指标冗余性较大。通过主成分分析输入向量从 53 个减至 28 个。

3.2 建立基于主成分分析法的 BP 神经网络

3.2.1 BP 神经网络结构及隐层设置 由于任意函数都可以被一个有 3 层单元的前馈网络逼近[5]，所以本研究选用的 BP 神经网络由输入层、隐藏层及输出层 3 层单元组成。输入层由主成分分析得出的 28 个主成分决定；输出层由虚、实两个证型指标决定。隐藏层结点在设置时并无统一规定，孙文恒等[6]在研究胰腺癌诊断中采用一层隐藏层，为的是减少计算量和防止过度拟合，通过误差对比，综合考虑后选择隐含节点（神经元数）为 7。故隐藏层结点的个数先由 $m = \sqrt{n \times 1} + c$ [$c \in$ (1, 10)] 确定，然后调整参数 c，最后通过测试正确率确定隐藏层结点的个数为 15。设定的网络结构，见图 1。

3.2.2 BP 神经网络输出方式 将实证与虚证样本的期望输出值设为（1，0）与（0，1）。由于隐藏层与输出层之间的激活函数采用的是正切 S 型传递函数 tanh，故该函数的预测区间为（0，1）。白云静等[7]采用 BP 神经网络进行糖尿病肾病中医证型研究时，将预测结果≥0.6 视为诊断成立。对此本研究在观察测试集数据的输出值后，通过观察正确率进而不断调试区间，最终确定将预测结果≥0.65 设为诊断成立。

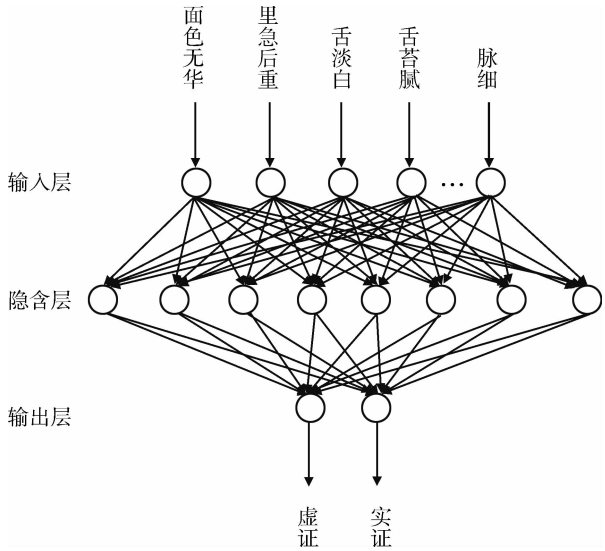

图 1 3 层 BP 神经网络结构

3.2.3 *BP 神经网络训练* 将主成分分析后的数据导入 python 数据框中，随机抽取前 80% 作为训练集，通过不断调试步长与迭代次数，使得网络的系统误差达到最小并且该神经网络趋于收敛。结果发现，在步长为 0.5、迭代次数为 50 000 时网络性能达标，训练自动停止。

3.2.4 *BP 神经网络测试* 待网络模型的权值趋于稳定、训练结束后，将数据框中的后 20% 数据作为测试集，规定输出值与原有值相等时，计数参数 *c* 自增 1，最后可计算出该神经网络模型的测试正确率。

3.3 建立基于 SPSS 软件的决策树分类模型

3.3.1 *决策树结构* 决策树的建模过程综合考虑患者的性别、年龄、腹部情况、大便情况、面部、脉象、舌质、舌苔和饮食等因素，构建大肠癌证型决策树分类模型，分析影响大肠癌证型决策树的主要因素，得出针对不同患者的症状与证型的相关关系。

3.3.2 *决策树训练及测试* 建立决策树模型进行虚、实证型分类，将预处理后的数据作为数据源，在验证选项框中，选择 80% 样本作为训练样本，20% 作为测试样本。经过测试，找到结点选择为父节点 20、子节点 10、效果更好，测试结果，见表 1。

表 1 决策树修剪结果

修剪严重性	10	20	30	40	50
树状图深度	3	4	4	3	2
训练样本证型为虚准确率(%)	91.5	65.1	66.0	65.1	68.1
训练样本证型为实准确率(%)	53.1	70.7	84.6	83.8	69.5
测试样本证型为虚准确率(%)	100	60.0	54.5	33.3	45.5
测试样本证型为实准确率(%)	12.5	71.4	88.9	81.8	69.2

因此，当“修剪严重性”在 10～50 间变化时，选择为父节点 20、子节点 10，能使得训练效果与测试效果较好。最终形成决策树，见图 2、图 3。

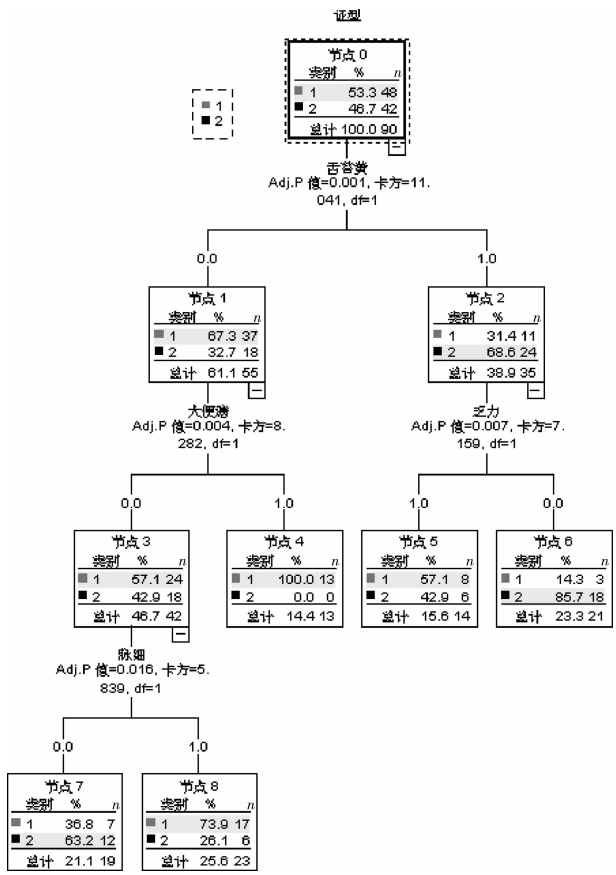

图 2 训练集决策树

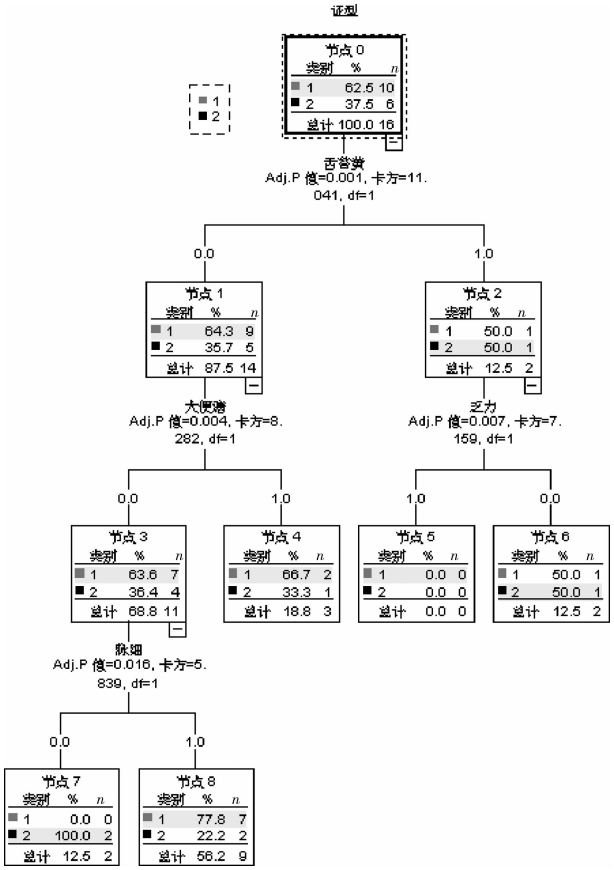

图 3 测试集决策树

训练和检验的标准误差分别为 0.050 和 0.084，符合误差要求。在该数据中，对于虚、实两种证型，舌苔是否黄是决定证型的主要因素，当舌苔黄为 1 时，样本为虚证的概率是 68.6%；当舌苔黄为 0 时，样本为实证的概率是 67.3%。

3.4 基于两种模型的结果对比

BP 神经网络模型与决策树模型的结果，见表 2。两个模型的样本总量均为 106 个，训练样本占样本总量的 80%，测试样本占样本总量的 20%。由表 2 可知，由于大肠癌症状与证型之间存在非线性关系，BP 神经网络分类模型对于非线性映射关系的处理比决策树分类模型好。决策树的准确率低可能与症状之间的相互交叉性有关，如果能解决症状之间的相互交叉性问题，决策树将是一种很好的分类算法。

表 2 两种模型结果对比

项目	BP 神经网络分类模型	决策树分类模型
训练样本准确率（%）	90.1	67.9
测试样本准确率（%）	85.3	65.7
平均值（%）	87.7	66.8

4 讨论

基于黑箱结构的人工神经网络能利用其自主学习能力[8-9]。采用 BP 神经网络建模，可将过程或对象看作一个“黑箱”，只要确定输入输出，就可以建立相应的模型，不必像传统的系统辨识那样将过程辨为线性还是非线性，这有利于对未知过程的系统进行建模。神经网络只能处理数值型数据，而决策树能处理非数值型数据这一优点恰好填补了本试验的空缺并与之形成对比。由于中医辨证学所研究的症状与证型之间是十分复杂的非线性关系，症状之间存在大量的多重共线性关系和协同关系，即某一个症状可以在多种证型中共同出现，而且可能对多种证型都具有重要的辨证价值[10]。而 BP 神经网络充分辨识在充分表现于外的“候”的表征信息的基础上，从样本中进行证候特征的规则提取，将其分布在网络的联接权举中，从而建立“候”与“证”的非线性映射函数[11]。因此，可以将其用于中医证型的非线性建模研究。本研究在 python2.7 环境下，对一组符合筛选标准的文献数据使用主成分分析法提取主成分后建立 BP 神经网络模型，之后再建立决策树模型。随机将样本总量的 80% 作为训练集，20% 作为测试集。结果显示，利用 BP 神经网络的自学习能力相较于决策树模型能更好地从样本中获取比较全面的证型内在规律，逼近证型的真实情况。因此 BP 神经网络模型在本研究中具有很好的诊断、预测能力。

5 结语

辨证论治是中医药治疗大肠癌的核心，是提高

（下转第 84 页）

引尽可能有一个合理的规范，确保同一概念的表述相同。除此之外，检索词库的使用尽可能地保证检索词的全面，进而使得科技查新的结果全面、客观。在应用检索词库进行信息检索时，除选择和使用检索词库中的主题词与同义词、缩写词、简称外，还要进一步考虑不同搜索引擎中的常用符号及特点，应用检索技巧，以提高全文检索系统或因特网搜索中文信息的查全率和查准率。

参考文献

1 Hilbert M, López P. The World's Technological Capacity to Store, Communicate, and Compute Information [J]. Science, 2011, 332 (6025): 60 - 65.

2 李慧美. 开阔国际视野，点亮学术人生——借助 CCSI - A&SHCI 进行人文、社会科学研究 [EB/OL]. [2016 - 10 - 31]. http://www.docin.com/p - 976524759 - f3.html.

3 林静，伊雷，陈珊珊，等. 大数据时代高校图书馆开展学科服务研究——学科馆员工作案例解析 [J]. 现代情报，2015，35 (12)：65 - 69.

4 孙淑萍，孙玉坤. 主题词、同义词、缩写词、简称对科技查新影响的实例分析 [J]. 情报探索，2013，(8)：85 - 87.

5 于曦. 面向嵌入式服务的科技查新服务体系优化 [J]. 图书馆工作与研究，2016，(5)：96 - 100.

6 张柏秋，吴晓鐀. 科技查新检索中的关键词选择 [J]. 情报科学，2008，26 (9)：1344 - 1348.

7 张岚，张柏秋，于非. 科技查新检索质量优化策略研究 [J]. 情报科学，2011，29 (6)：852 - 855.

8 朱康玲. 同义词的获取对医学科技查新查全率和查准率的影响 [J]. 中华医学图书情报杂志，2012，21 (3)：78 - 80.

9 宋玉，王筠，曲磊. 临床医学期刊论文的发表时滞与老化分析 [J]. 情报探索，2014，(4)：15 - 17.

10 杭文文，谢靖. 我国肿瘤学期刊引用半衰期研究 [J]. 江苏科技信息，2015，(26)：19 - 21.

11 刘伙玉. 基于 CNKI 的图书、情报学与档案学学科文献半衰期分析 [J]. 图书与情报，2015，(1)：106 - 111.

12 罗式胜. 文献半衰期的类型及其应用 [J]. 情报学报，1997，16 (1)：62 - 67.

13 马磊，宋建玮. IPC 分类法在科技查新工作中的应用 [J]. 图书馆学刊，2012，(3)：32 - 34.

14 王振风. 基于 Lucene 的分布式全文检索技术的研究与应用 [D]. 上海：东华大学，2015.

(上接第 64 页)

中医药治疗结直肠癌临床疗效的前提和基础。但目前临床研究对中医证型的分析缺乏多中心、大样本的流行病学研究，为辨证的客观化、标准化带来了一定的困难，从而制约了中国医临床研究水平的进一步提高。因此在今后的研究中应遵循循证医学的原则，本着科学严谨的态度，对大肠癌的症、证进行调研，开展多中心、随机对照试验，以揭示大肠癌的中医辨证论治规律，规范大肠癌中医证型。

参考文献

1 万德森. 结直肠癌的流行病学及其危险因素研究近况 [J]. 实用癌症杂志，2000，15 (2)：220 - 222.

2 陈叶，刘金涛，朱源，等. 大肠癌中医辨证及治疗概况 [J]. 中国肿瘤，2015，24 (4)：319 - 324.

3 许云，杨宇飞. 结直肠癌中医药研究进展与思考 [J]. 世界中医药，2014，9 (7)：828 - 532.

4 姚乃礼. 中医症状鉴别诊断学. 2 版 [M]. 北京：人民卫生出版社，2000：27 - 35.

5 王颖纯，白丽娜. 基于 BP 神经网络的中医脉诊体质类型判定 [J]. 中医杂志，2014，55 (15)：1288 - 1291.

6 连俊彦. 多因素回归生存分析探讨影响大肠癌病人术后生存期的因素 [J]. 实用医学杂志，2002，8 (6)：597 - 598.

7 白云静，孟庆刚，申洪波，等. 基于改进的 BP 神经网络的糖尿病肾病中医证候非线性建模研究 [J]. 北京中医药大学学报，2008，31 (5)：309 - 311.

8 Dayhof JE, Deleo JM. Artificial Neural Networks [J]. Cancer, 2001, 91 (8): 1615 - 1634.

9 Cross SS, Harrison RF, Kennedy RL. Introduction to Neural Networks [J]. Lancet, 1995, 346 (8982): 1075, 1079.

10 李建生，王至婉，余学庆，等. 基于慢性阻塞性肺病 (COPD) 急性加重期文献的多元统计方法在证候研究的应用探讨 [J]. 中医学报，2007，22 (6)：8 - 10.

11 孙贵香，袁肇凯. 人工神经网络在中医证候研究中的应用 [J]. 中华中医药学刊，2007，25 (7)：1450 - 1452.