[11] 梁文娜,李西海,李灿东.绝经后骨质疏松的核心病机——骨痿.中国老年学杂志,2015,35(18):5333-5335
[12] 黎晓敏,卢昌均,周艳芳,等.黄芪三仙汤干预成骨细胞护骨素及护骨素配体的表达.中国组织工程研究,2013,584(11):1939-1945
[13] 许兵,方剑利,刘慧,等.补肾活血方对去势大鼠骨质疏松的影响.中华骨质疏松和骨矿盐疾病杂志,2011,4(3):177-1782
[14] Hsu H,Lacey D L,Dunstan C R,et al.Tumor necrosis factor receptor family member RANK mediates osteoclast differentiation and activation induced by osteoprotegerin ligand.Proceedings of the National Academy of Sciences of the United States of America,1999,96(7):3540-3545
[15] Boyce B F,Xing L.Functions of RANKL/RANK/OPG in bone modeling and remodeling.Archives of Biochemistry and Biophysics,2008,473(2):139-146
[16] Tanaka S.Signaling axis in osteoclast biology and therapeutic targeting in the RANKL/RANK/OPG system.American Journal of Nephrology,2007,27(5):466-478
[17] 马慧萍,贾正平,张汝学,等.淫羊藿总黄酮含药血清促进骨髓间充质干细胞增殖与成骨性分化.中国骨质疏松杂志,2004,10(4):38-40,46
[18] Nian H,Ma M H,Nian S S,et al.Antiosteoporotic activity of icariin in ovariectomized rats.Phytomedicine,2009,16(4):320-326
[19] 马小妮,葛宝丰,陈克明,等.淫羊藿苷调节成骨细胞骨形成和破骨细胞骨吸收的机制.中国医学科学院学报,2013,35(4):432-438
[20] 郭海玲,赵咏芳,王翔,等.淫羊藿苷对人成骨细胞增殖及OPG蛋白表达的实验研究.中国骨伤,2011,24(7):585-588
[21] 于燕,颜虹,胡森科.淫羊藿及淫羊藿苷雌激素样作用研究.郑州:第十届中国科协年会论文集,2008
[22] Hofbauer L C,Kuehne C A,Viereck V.The OPG/RANKL/RANK system in metabolic bone diseases.Journal of Musculoskeletal & Neuronal Interactions,2004,4(3):268-275

（收稿日期：2017年11月28日）

·研究报告·

基于决策树和神经网络的高血压病痰湿壅盛证诊断模型研究

田艳鹏[1]，丁学义[1]，朱羽硕[2]，李运伦[2]，郭伟星[3]

（[1]山东中医药大学，济南 250355；[2]山东中医药大学附属医院，济南 250014；[3]山东省医学科学院，济南 250062）

摘要：目的：建立高血压病中医临床常见证候信息对痰湿壅盛证的诊断模型。方法：收集高血压病古今医案资料及临床病例资料，对所收集资料中患者的中医四诊信息进行归一化处理，采用C5.0、CRT、CHAID、QUEST决策树方法和神经网络方法，从35个中医临床常见证候中提取痰湿壅盛证的诊断规律，并形成诊断模型。结果：CHAID、CRT、QUEST及C5.0四种决策树算法对高血压病痰湿壅盛证的诊断准确率分别为82.9%、91.1%、92.4%、93.7%，其中C5.0决策树模型的准确率高于其它3种算法。采用径向基函数（RBF）神经网络和多层感知器（MLP）神经网络对原始证候数据分析，前者训练准确率为92.2%，测试准确率为91.4%，后者训练准确率为94.5%，测试准确率为90.4%，且RBF神经网络的诊断模型更优于MLP神经网络。结论：基于中医临床四诊信息数据，运用决策树和神经网络等数据挖掘方法，构建高血压病痰湿壅盛证诊断模型，能够直观地、清晰地对高血压病痰湿壅盛证进行诊断，归纳总结诊断规律，从而为高血压病痰湿壅盛证的中医证候规范提供依据。

关键词：数据挖掘；高血压病；痰湿壅盛证；证候规律；决策树

基金资助：国家自然科学基金项目（No.81473653），国家中医临床研究基地业务建设第二批科研专项（No.JDZX2015144），泰山学者工程专项经费资助项目

通讯作者：李运伦，山东省济南市历下区经十路16369号山东中医药大学附属医院，邮编：250014，电话：0531-68616038
E-mail：li.yunlun@163.com
郭伟星，山东省济南市历下区经十路18877号山东省医学科学院，邮编：250062，电话：0531-82929888，E-mail：szy206044@126.com

Study on the diagnostic model with syndrome of exuberance of phlegm-damp in hypertension based on decision tree and neural network

TIAN Yan-peng[1], DING Xue-yi[1], ZHU Yu-shuo[2], LI Yun-lun[2], GUO Wei-xing[3]

([1]Shandong University of Chinese Medicine, Jinan 250355, China; [2]Affiliated Hospital of Shandong University of Chinese Medicine, Jinan 250014, China; [3]Shandong Academy of Medical Sciences, Jinan 250062, China)

Abstract: Objective: To establish a diagnostic model of phlegm-damp syndrome in hypertension based on the common clinical syndrome of traditional Chinese medicine. Methods: The information of ancient and modern medical records and clinical cases of hypertension were collected. The collected four diagnostic information of TCM was normalized, and the C5.0, CRT, CHAID, QUEST decision tree methods and neural networks in SPSS 20.0 software were used for data analysis. The diagnostic rules of phlegm-damp syndrome were extracted from 35 common TCM clinical syndromes and finally the diagnostic models were established. Results: The diagnostic accuracy of CHAID, CRT, QUEST and C5.0 were 82.9%, 91.1%, 92.4% and 93.7%. The accuracy rate of C5.0 decision tree model was higher than that of the other three algorithms. Radial basis function neural network and multi-layer perceptron neural network were used to analyze the original syndrome data. The training accuracy of the former was 92.2% and the test accuracy was 91.4%, while the training accuracy of the latter was 94.5% and the test accuracy was 90.4 %. The diagnostic model of radial basis function neural network was better than multilayer perceptron neural network. Conclusion: We can directly and clearly diagnose phlegm-damp syndrome in hypertension and summarize the diagnostic rules through constructing diagnosis models of phlegm-damp syndrome in hypertension based on the four diagnostic information of TCM with the data mining methods such as decision tree and neural network, so as to provide evidence for TCM syndromes criterion of hypertension with syndrome of exuberance of phlegm-damp.

Key words: Data mining; Hypertension; Syndrome of exuberance of phlegm-damp; Syndrome differentiation regularity; Decision tree

Funding: National Natural Science Foundation of China (No.81473653), The Second Batch of Scientific Specialized Program of Construction of Research Bases of National Chinese Medicine Clinical Research (No.JDZX2015144), Fund of Taishan Scholar Project

高血压病是世界范围内的现代流行性疾病[1]，伴有进行性的靶器官损害，发病率呈逐年上升之势，是现代心脑血管疾病的主要危险因素之一。中医药以其独特的辨证方法和治疗观念在高血压病的防治过程中日益体现出其独特优势。数据挖掘综合运用可视化和信息技术、数据库理论及统计学专业知识、从大量数据中提取有用的信息[2]，为临床医学和中医理论辨证提供了充分的科学证据，并在中医经典理论研究、中医临床辨证等方面得到了广泛应用。新的数据挖掘方法不仅能够分析中医古籍和现代医案的用药规律，还能够综合症状和证候为中医药的临床发展提供数据支撑[3]。黄嘉韵等[4]基于数据挖掘中的决策树算法探究了鼻衄的中医辨证规律，在建立的决策树模型中筛选出5条判断规则，准确率达93.1%。本研究以高血压病为研究对象，通过收集大量的高血压病古今医案资料和临床病例资料，利用C5.0、CRT、CHAID、QUEST决策树方法以及径向基函数（radial basis function，RBF）神经网络和多层感知器（multilayer perceptron，MLP）神经网络等数据挖掘方法，分析高血压病痰湿壅盛证的诊断规律，并形成诊断模型，从而为高血压病痰湿壅盛证的中医证候规范提供依据，为进一步深入的研究奠定基础。

资料

1. 一般资料　利用网络和手工检索有关高血压病的各类名医医案书籍及期刊文献。并收集2014年6月至12月就诊于山东中医药大学附属医院的高血压病患者病历资料。共纳入高血压病研究病例463例，包括中医古籍医案385例及临床高血压病病例78例。

2. 诊断标准　高血压病诊断标准参照《中国高血压防治指南》（2010版）。证候及辨证分型标准参照《中药新药临床研究指导原则》（2002年）和前期研究中制定的高血压病痰湿壅盛证证候诊断量表[5]。辨证标准：主症：眩晕、头痛、头重昏蒙，胸闷，呕恶，多痰涎；次症：心悸，肢体困倦，口淡食少，大便黏腻不爽，舌胖大苔腻，脉滑，医案中的病例也依据上述方法。

3. 纳入标准　医案入选标准：①符合头痛、眩晕的诊断标准；②症状和药物描述比较完善；③单纯中药治疗。临床病例入选标准：①年龄在18~75岁之间；②符合高血压病诊断标准；③平素服用降压药物，仍不能良好控制血压者；④单纯中药治疗；⑤1级高血压病程必须>3个月；⑥对调查知情同意。

4. 排除标准　医案排除标准：①失治误治；②有西药介入；③对于一稿多投的文献，只取1篇；④综述性文章。临床病例排除标准：①继发性高血压病；②过敏体质及对多种药物过敏者；③妊娠或准备妊娠或哺乳期妇女；④患有严重的精神性疾病者；⑤具有其他可能导致心排出量增加和收缩压升高的疾病，如重度贫血、甲状腺功能亢进症、主动脉瘘等；符合任何1项或以上者均应排除，不能入选。

5. 痰湿壅盛证分布情况　463例高血压病患者经临床辨证

为痰湿壅盛证的患者253例，非痰湿壅盛证患者210例，且痰湿壅盛证和非痰湿壅盛证患者基线情况差异无统计学意义。

方法

1. 资料采集及辨证分型　利用网络检索中国期刊全文数据库（CNKI）、万方数据库、维普数据库（VIP）、中华医学会（CMA）等国家大型电子图书馆，并采用手工检索山东中医药大学图书馆及山东省图书馆有关高血压病的各类名医医案书籍及期刊文献。对门诊和入院患者进行基本资料及中医四诊信息的采集，由2名副主任医师及以上、具有5年及以上相关工作的临床经验的临床医生进行辨证分型的判断，并将古今医案资料与临床病例资料进行联合统计。

2. 数据预处理　对所收集的古今医案及临床病历资料进行整理，建立高血压病痰湿壅盛证证候信息数据库（包括病案资料来源、患者情况、四诊信息和病因病机信息等），进而对各证候要素进行语言规范化处理，剔除出现次数较少的信息后，以证候要素作为自变量，“是否为痰湿壅盛证”作为因变量，采取“0,1”赋值法，是为“1”，否为“0”。根据统计结果，剔除出现频率小于10%的证候因子后，将数据进行主成分分析，最终筛选出35个证候因子作为自变量

3. 高血压病痰湿壅盛证诊断模型的构建及验证　经数据预处理，最终选择35个证候因子作为自变量，“是否为痰湿壅盛证”作为因变量进入决策树模型和神经网络模型的筛选过程。采用SPSS 20.0软件中的C5.0、CRT、CHAID、QUEST决策树方法和神经网络方法进行识别规律的挖掘。并采用10倍交叉验证方法对形成的识别模式进行验证。

结果

1. 主成分分析结果　数据经过主成分分析，发现前35个主成分的贡献率超过了全部指标的90%，说明部分指标的冗余性较大，通过主成分分析，输入向量减至35个。

2. 高血压病痰湿壅盛证CHAID决策树模型结果　见图1。应用CHAID算法，对35个证候因子进行决策树分析，在构建模型的过程中考虑到样本数量的限制，为保证并促进树模型的良好生长，父节点数设定为100，子节点50。筛选出头重昏蒙、呕恶、多痰涎、苔白腻和肢体困重5个属性形成决策树模型。该模型深度为3，共有11个节点，6个终结点，形成6个对是否为痰湿壅盛证的识别路线，头重昏蒙为最佳识别属性。10倍交叉验证结果显示其证候判别准确率为82.9%。

3. 高血压病痰湿壅盛证CRT决策树模型结果　见图2。应用CRT算法，对35个证候因子进行分析，父节点50，子节点25。筛选出包括头重昏蒙、呕恶、多痰涎、苔白腻、胸满闷和痞满6个属性形成决策树模型，该模型深度为4，共有15个节点，8个终结点，形成8个对是否为痰湿壅盛证的识别路线，准确率为91.1%。

4. 高血压病痰湿壅盛证QUEST决策树模型结果　见图3。应用QUEST算法，对35个证候因子进行分析，选取父节点为

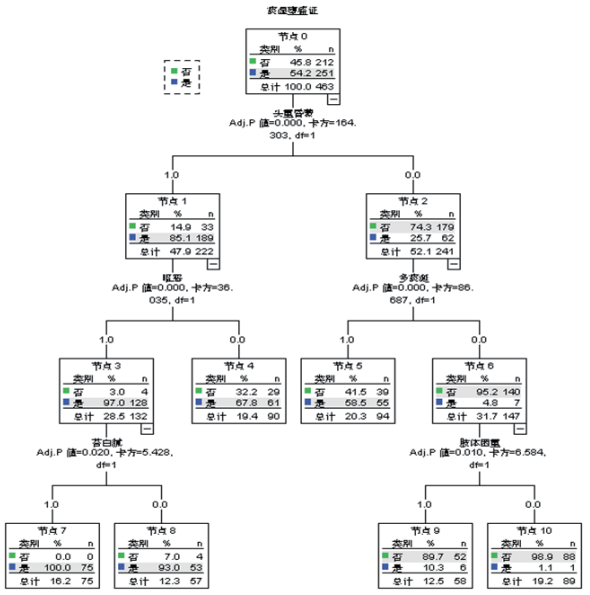

图1　463例高血压病患者痰湿壅盛证CHAID决策树识别模型

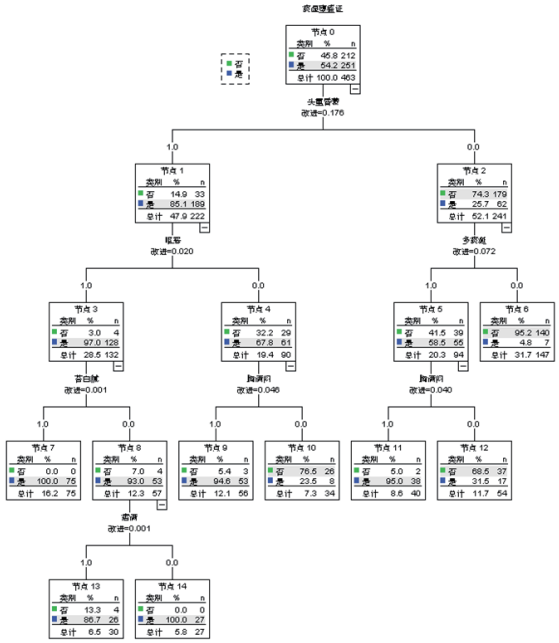

图2　463例高血压病患者痰湿壅盛证CRT决策树识别模型

30，子节点为10，筛选出头重昏蒙，多痰涎，呕恶，胸满闷和苔白腻5个属性形成决策树模型。模型深度为4，共有15个节点，8个终结点，形成8个对是否为痰湿壅盛证的识别路线，准确率为92.4%。

5. 高血压病痰湿壅盛证C5.0决策树模型结果　见图4。应用C5.0算法，对35个证候因子进行分析，筛选出头重昏蒙，呕恶，多痰涎，胸满闷，苔白腻，精神倦怠和痞满7个属性形成决策树模型，模型深度为5，共有21个节点，11个终结点，准确率为93.74%。

6. BP神经网络结构及隐层设置　由于任意函数都可以被

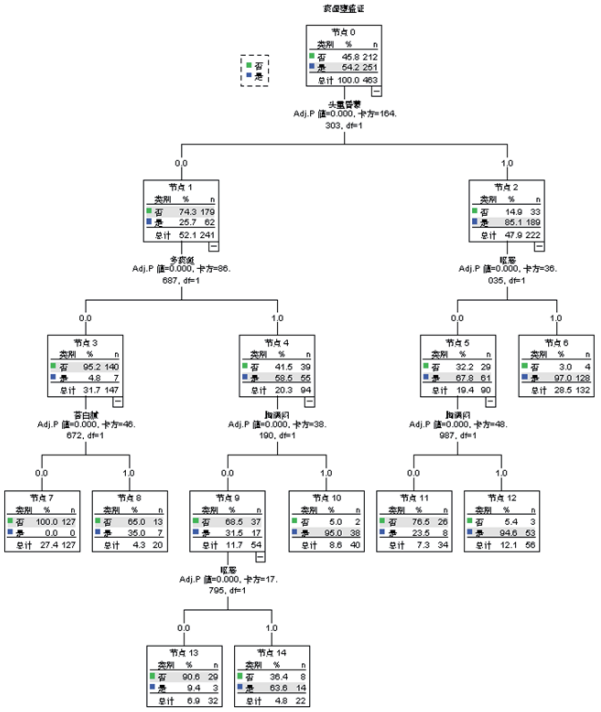

图3 463例高血压病患者痰湿壅盛证QUEST决策树识别模型

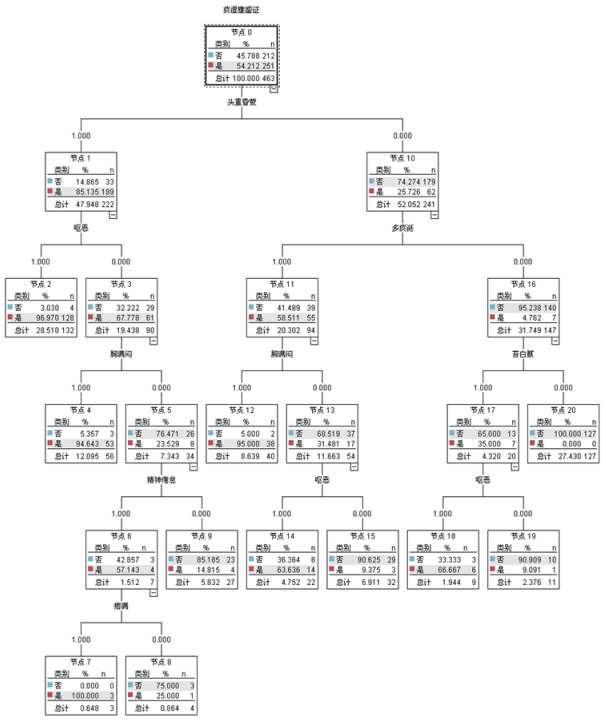

图4 C5.0算法的决策树模型示意图

1个有3层单元的前馈网络逼近，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35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，通过误差对比综合考虑后确定神经元数。

表1 基于RBF神经网络的高血压病痰湿壅盛证诊断模型自变量的重要性

症状	重要性	标准化的重要性(%)
头重昏蒙	0.171	100.0
多痰涎	0.142	82.9
胸满闷	0.108	63.3
呕恶	0.107	62.5
苔白腻	0.106	62.0
脉弦	0.061	35.6
痞满	0.054	31.3
舌胖大	0.052	30.1
脉滑	0.048	28.0
大便黏腻不爽	0.042	24.8
精神倦怠	0.042	24.5
泄泻	0.036	21.1
肢体困重	0.032	18.4

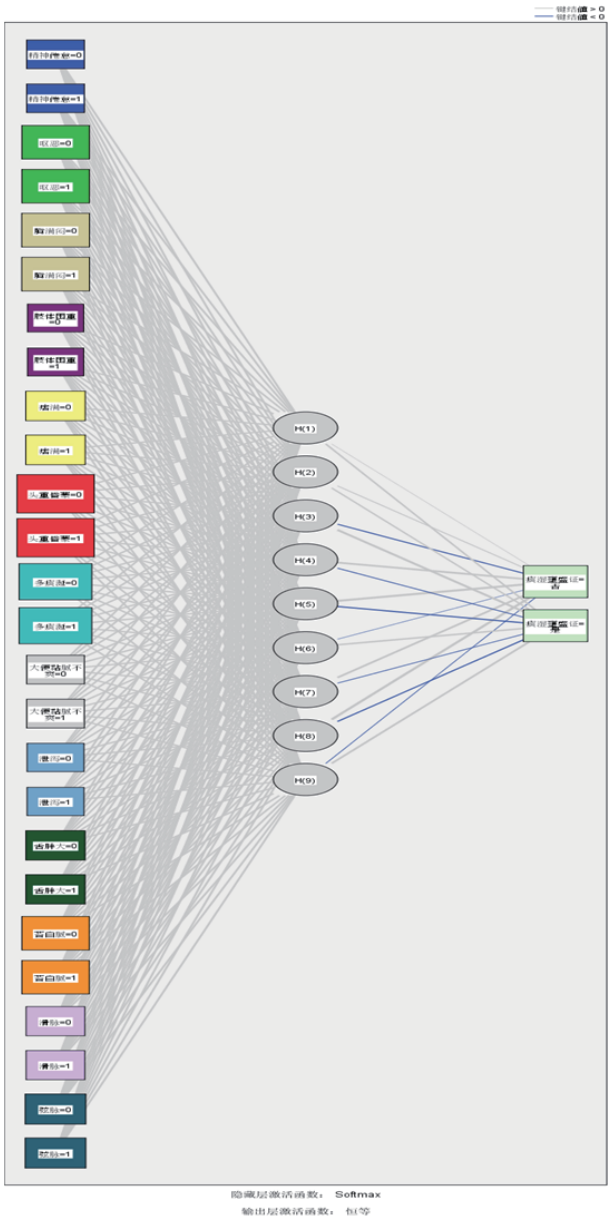

图5 RBF神经网络图

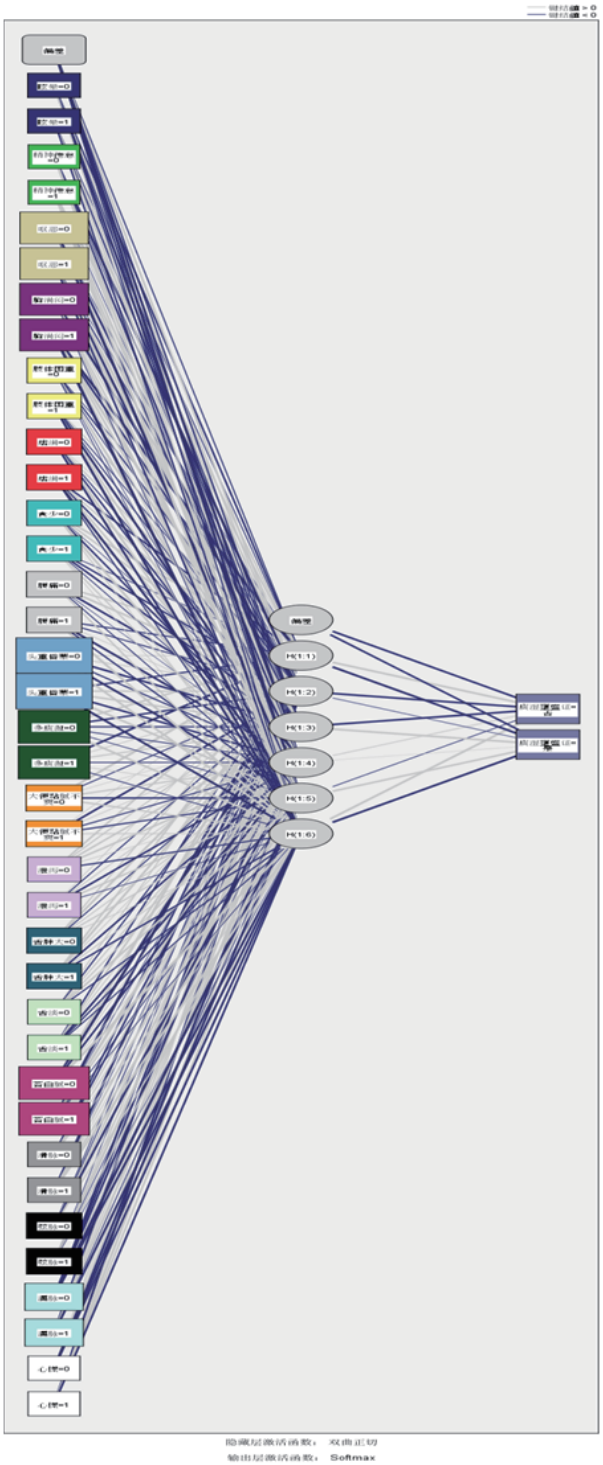

图6 MLP神经网络图

7. 基于RBF神经网络的高血压病痰湿壅盛证诊断模型 见表1、图5。运用RBF神经网络对赋值后的证候数据进行分析，得出RBF神经网络图，证候诊断的训练准确率为92.2%，测试准确率为91.4%。高血压病痰湿壅盛证证候标准化的重要性>60%的自变量属性依次为头重昏蒙、多痰涎、胸满闷、呕恶、苔白腻。其标准化的重要性所占比例依次为100.0%、82.9%、63.3%、62.5%、62.0%。

8. 基于MLP神经网络的高血压病痰湿壅盛证诊断模型 见表2、图6。运用MLP神经网络对赋值后的证候数据进行分析，得出MLP神经网络图，证候诊断的训练准确率为94.5%，测试准确率为90.4%。高血压病痰湿壅盛证证候标准化重要性>60.0%的自变量属性依次为头重昏蒙、多痰涎、苔白腻、胸满闷和呕恶，其标准化的重要性所占比例依次为100.0%、80.6%、77.3%、71.0%、69.2%。

表2 基于MLP神经网络的高血压病痰湿壅盛证诊断模型自变量的重要性

症状	重要性	标准化的重要性（%）
头重昏蒙	0.171	100.0
多痰涎	0.138	80.6
苔白腻	0.132	77.3
胸满闷	0.121	71.0
呕恶	0.118	69.2
脉濡	0.051	29.8
大便黏腻不爽	0.035	20.8
脉弦	0.031	18.3
腹痛	0.029	17.1
痞满	0.029	16.9
舌胖大	0.025	14.8
食少	0.024	14.4
肢体困重	0.022	12.8
泄泻	0.016	9.5
舌淡	0.015	8.7
脉滑	0.014	8.3
眩晕	0.013	7.7
心慌	0.009	5.1
精神倦怠	0.007	4.0

讨论

脾居中焦，为人体气机升降开阖之枢机，痰浊中阻，阻遏清阳，发为眩晕。早在金元时期，《丹溪心法》中已有“无痰则不作眩”的论述。中医学的“痰”有广义和狭义之分。广义之痰是由于津液和水谷精微停聚在机体任何部位而成，无形可见且变幻多端。狭义之痰是指唾出体外，有形有物，可见之痰。随着现代生活水平的提高，因嗜食膏粱厚味而滋生疾病的疾病不断增高。有学者研究探讨现代临床高血压病与痰湿中阻之间的内在关系，从中得出饮食劳倦可损伤脾胃，聚湿生痰，阻滞清阳，蒙蔽清窍而发病的结论[6]。赵立诚认为痰浊在高血压病发生发展过程中占有重要的地位，主张从“痰”论治高血压病[7]。朱妍等[8]认为血瘀痰凝，瘀阻脉道，郁而蕴蒸，凝聚化毒为高血压病的主要病机，治疗当以祛痰化瘀法为主。本研究在立足于高血

压病痰湿壅盛证古今医案资料和现代临床病例资料的基础上，基于中医临床四诊信息，建立了中医证候信息数据库，为下一步运用CHAID、CRT、QUEST、C5.0决策树方法和RBF、MLP神经网络等数据挖掘技术解读其中蕴含的信息，探寻高血压病的证候规律奠定了基础，最终构建了高血压病痰湿壅盛证的证候诊断判别模型。

CHAID算法即卡方自动交互检测法(chi-squared automatic interaction detector)[9]，属决策树中算法的一种，它具有目标选择及变量筛选等功能[10]，主要针对预先给定的结果变量，对众多分类变量进行比较和筛选，找到最优分类变量和结果，并根据卡方检验的结果自动判断分组[11-12]。CRT与CHAID算法均属于分类树方法，主要根据自变量对因变量进行分类检测，其核心是以已知的类别作为对象建立判别函数，然后将分类指标带入此函数，根据所得函数值判断该指标所应归入的类别[13]。QUEST算法属二次元分类方法[14]，它是在CHAID算法的基础上进行改进，使用方差分析、卡方检验、聚类分析和判别分析等方法，生成精确的二叉树模型。C5.0是基于ID3和C4.5算法形成的决策树方法，主要使用信息增益率选择属性。RBF神经网络是1988年Moody和Darken提出的一种神经网络结构，属于前向神经网络类型，它能够以任意精度逼近任意连续函数，特别适合于解决分类问题。多层感知器是一种前馈人工神经网络模型，其将输入的多个数据收集映射到单一的输出的数据集上。在本研究中，采用C5.0决策树算法对高血压病痰湿壅盛证的证候属性进行分析，判断准确率达到了93.74%，高于其他3种算法。决策树算法筛选出头重昏蒙、呕恶、多痰涎，胸满闷，苔白腻，精神倦怠，痞满7种中医属性，这些全部出现于决策树的根节点中，准确率较为理想，符合中医辨证思路。在归纳、整合4种模型各自筛选出的证候属性后，发现头重昏蒙、呕恶、多痰涎、苔白腻为4个类证方法的共性四诊信息。若进一步结合如胸满闷、痞满、肢体困重等症状，可在四诊信息层面形成良好的痰湿壅盛证组合判别模式，与中医理论较为吻合，可为临床辨证提供相对客观化的依据，但仍需将这些结论在临床中加以检验。此外，本研究还采用RBF和MLP两种神经网络方法对入选的35种分类属性做了重要性分析，两种方法均得出头重昏蒙为高血压病痰湿壅盛证的最佳识别变量。从模型的测试样本准确率来看，RBF神经网络模型更优于MLP神经网络模型。

综上所述，运用决策树和神经网络等数据挖掘技术对高血压病痰湿壅盛证的证候分布规律的进行分析，建立高血压病痰湿壅盛证诊断模型，能够直观地、清晰地对高血压病痰湿壅盛证进行诊断，归纳总结诊断规律，从而为高血压病痰湿壅盛证的中医证候规范提供依据。但本研究存在样本量较少，无法开展不同年龄组合、不同证候组之间的分层次研究，且自变量均为中医临床四诊信息，无临床检测指标的纳入，这也是下一步研究的重点。此外，当自变量相对庞大时，各自变量作为影响因素对证型的诊断意义不等，需进一步比较研究。

参 考 文 献

[1] 《中国高血压防治指南》修订委员会.中国高血压防治指南(2010年修订版).北京:人民卫生出版社,2011:2-30

[2] 任智军,朱东华,谢菲.科技文本的可视化分析研究.北京理工大学学报:社会科学版,2007,9(1):12-17

[3] 任建业,许鸣,陆嘉惠.基于数据挖掘的中医临床用药规律和证型研究进展.中华中医药杂志,2017,32(10):4579-4582

[4] 黄嘉韵,郭宏,邝艳萍.基于决策树算法的鼻鼽辨证规律初步研究.中华中医药杂志,2016,31(11):4770-4773

[5] 孙洁.高血压病痰湿壅盛证文献计量学研究.济南:山东中医药大学,2012

[6] 龚一萍.试论痰浊内阻在高血压病形成中的作用.湖南中医药大学学报,2001,21(2):32-33

[7] 张国华.赵立诚从痰论治原发性高血压病经验.浙江中医杂志,2006,41(4):206-207

[8] 朱妍,韩学杰.高血压病从痰瘀论治的理论研究.中西医结合心脑血管病杂志,2006,4(10):890-891

[9] 张芬,余金明,王家宏,等.Exhaustive CHAID分类树与logistic回归在脑卒中危险因素中的应用.中国预防医学杂志,2011,12(7):573-576

[10] 吴超,张晓祥.数据挖掘在疾病诊断相关组项目中的应用.中国数字医学,2010,5(5):70-72

[11] 帅健,李丽萍,陈业群.决策树模型与Logistic回归模型在伤害发生影响因素分析中的作用.中华疾病控制杂志,2015,29(2):185-189

[12] 张玥娇,徐昕,代礼,等.CHAID模型在巩义市新生儿低出生体重影响因素筛选中的应用.现代预防医学,2012,39(1):18-22

[13] 张超.分类树中CRT算法与判别分析的比较及其医学应用.数理医药学杂志,2008,21(2):139-141

[14] Loh W Y,Shih Y S.Split selection methods for classification trees. Statistica Sinica,1997,7:815-840

(收稿日期：2017年11月28日)