

复合结构智能化辨证选方模型的构建

周璐¹ 李光庚² 孙燕¹ 郑岩² 李宇航¹

(1 北京中医药大学中医学院, 北京, 100029; 2 北京邮电大学计算机学院, 北京, 100876)

摘要 通过回顾机器学习中的 C4.5 决策树、随机森林、支持向量机、BP 神经网络算法的原理与在中医辨证研究中所取得的成果, 创新设计出一种复合结构的智能化辨证选方模型, 并对模型进行了实现与测试。结果表明该模型输出结果的准确性高于仅使用单一算法建立的辨证选方模型, 这为进一步与“方-证要素对应”的组方原则相结合, 建立适用于复杂病机的临床诊疗辅助系统奠定了基础。

关键词 人工智能; 机器学习; 方-证要素对应; 辨证论治

Construction of Intelligent Syndrome Differentiation and Formula Selection of Compound Structure Model

Zhou Lu¹, Li Guangeng², Sun Yan¹, Zheng Yan², Li Yuhang¹

(1 School of Chinese Medicine, Beijing University of Chinese Medicine, Beijing 100029, China; 2 School of Computer Science, Beijing University of Posts and Telecommunications, Beijing 100876, China)

Abstract Through a review of the C4.5 Decision Tree, Random Forest, Support Vector Machine and BP Neural Network algorithm and the obtained in the study of TCM syndrome differentiation, a compound structure of intelligent syndrome differentiation and formula selection model was designed. The model was implemented and tested. Results showed that the accuracy of the corresponding results was higher than that of the single algorithm, laying the foundation for the model which was suitable for the complex Pathogenesis of clinical diagnosis and treatment of auxiliary system.

Key Words Artificial Intelligence; Machine Learning; Correspondence between syndrome and formula factors; Syndrome differentiation and treatment

中图分类号: R241; R242 文献标识码: A doi: 10.3969/j.issn.1673-7202.2018.02.057

辨证论治是中医的特色, 机器学习是当今人工智能领域的重要分支, 将机器学习与中医辨证论治进行结合, 建立智能化的辨证论治模型, 可辅助临床医师进行处方决策, 为中医复杂病机的辨证论治提供诊疗思路。因此, 分析机器学习在中医辨证研究中的应用, 探索智能化的辨证论治模型也是当今智能时代具有挑战意义的一项课题。

机器学习在 Machine Learning^[1] 中的定义为“如果一个计算机程序针对某类任务 T 的用 P 衡量的性能根据经验 E 来自我完善, 那么我们称这个计算机程序可从经验 E 中学习, 针对某类任务 T, 他的性能用 P 来衡量”。我们通过回顾学者所用的 C4.5 决策树算法、随机森林算法、支持向量机算法、BP 神经网络算法的原理与用于中医辨证所取得的成果。创新设计出一种复合结构的智能化辨证选方模型, 相对于单一算法建立的模型, 该模型可更准确辨识出

症状中所包含的证候, 实现精准的辨证选方。

1 机器学习算法应用于中医辨证研究的回顾

1.1 BP 神经网络 (Back Propagation Neural Network) BP 神经网络算法是基于反向传播算法^[2] 而实现, 具有建立多分类模型的能力。该算法主要通过给定包含样本类别与样本特征的训练集合, 自主建立多分类模型, 具有较好的自学习与自组织能力。

这种算法被中医较早应用到辨证研究之中, 如田禾和戴汝为^[3] 早在 1990 年用 BP 神经网络算法进行中医智能辨证的研究, 开发了用于中医儿科咳嗽的辨证系统。目前使用 BP 神经网络进行中医辨证的研究仍是热点, 如本课题组陈擎文^[4] 曾通过 BP 神经网络对《伤寒论》中的“方-证要素”进行探讨, 初步构建出基于《伤寒论》的方证对应及“方-证要素对应”的辨证论治模型。徐亮等^[5] 为探索挖掘名老中医辨证经验的新方法, 运用 BP 神经网络对 684 名

基金项目: 国家重点研发计划“中医药现代化研究”课题 (2017YFC1700303)

作者简介: 周璐 (1990.06—), 男, 硕士研究生, 研究方向: 中医辨证论治规律研究

通信作者: 李宇航 (1960.08—), 男, 博士, 教授, 博士研究生导师, 研究方向: 中医辨证论治规律, E-mail: liyuhang@bucm.edu.cn

老中医的医案构建辨证模型,其中用 669 例医案作为训练集,15 例气虚证医案作为测试集,预测一致性为 80%,证明了该方法的可行性。

1.2 C4.5 决策树(C4.5 Decision Tree) C4.5 决策树算法^[6]是一种可用于构建树形分类模型的决策树算法,通过对给定的训练数据计算信息熵、信息增益率建立分类模型,并可提取出树形的分类规则,实现对未知样本的类别进行可视化的分析与判断。

这种算法出现于神经网络之后,被中医较多的用于提取证候的诊断规则,如杨开明^[7]将 85 例糖尿病患者结合 C4.5 决策树算法建立了中医辨证决策树,提取出糖尿病的 8 种中医证候分类规则。李治和李国琳^[8]对 447 例 AECOPD 患者结合 C4.5 决策树算法建立起 AECOPD 的中医辨证分型决策树模型,提取出痰热郁肺、痰瘀阻肺、痰浊阻肺、外寒内饮 4 种中医证候的诊断规则。刘广等运用 C4.5 决策树算法结合 800 例中医胃炎病例建立了中医胃炎辨证分类决策树,并提取出胃炎的中医证候分类规则。

1.3 支持向量机(Support Vector Machine) 支持向量机是由 Cortes 和 Vapnik^[9]提出的一种可用于二分类问题的机器学习算法。该算法通过运用支持向量机的核函数等公式进行一系列的运算,得出支持向量机的类别判定公式 $y = W^T X + b$,并结合支持向量机中划分样本类别的超平面,实现对样本类别的划分,建立分类模型。

本算法的出现略晚于决策树,但其在中医辨证研究中的应用却较为广泛,如晏峻峰^[10]用支持向量机算法结合“脾虚”与“阳虚”的证素,构建辨别“脾虚”与“阳虚”的智能辨证模型。王阶等^[11]通过支持向量机算法从 115 例名中医冠心病治疗的医案中提取出血瘀、痰浊、气虚、阳虚、阴虚、内热、血虚、气滞 8 个证候要素的主要表现。此外,因为支持向量机算法构造的辨证模型具有很好的准确性,所以也常用于辨证模型之间的对照分析。如王华珍和胡雪琴^[12]运用随机森林算法对 2021 条“内生五邪”的病案建立辨证模型,并用支持向量机算法建立对照模型,分析辨证模型的准确性。

1.4 随机森林(Random Forest) 随机森林算法是由 Breiman^[13]所提出的一种组合分类器算法,可构造出多个树形分类模型。“随机”一词在这里有两层含义,第一层可以理解为在总训练样本中随机有放回的为森林中的每个决策树选取与总训练样本数相同的样本,作为构建决策树的训练集合;第二层

是对森林中的每个决策树从所有样本属性中随机不放回的选择部分样本属性。随机森林算法就是先通过样本与属性随机选择,再像生成决策树模型一样,生成由决策树组成的“森林”。随机森林通过采用森林内决策树投票的方式判定待测样本类别,而不是某一决策树的单一判断,因此具有较高的准确性。

与前 3 种算法相比,本算法出现较晚,在中医辨证研究中可谓是一种较新技术手段。如 2015 年孙超^[14]运用随机森林算法对糖尿病肾病建立辨证模型,并同时对症状的重要程度进行了客观化度量,为糖尿病肾病辨证的客观化诊断提供了依据。2016 年蔡晓路^[15]运用随机森林对类风湿性关节炎的证候建立辨证模型,并提取出对证候诊断有意义的重要症状,为研究证候的主证探索了一种有效方法。此外,随机森林算法还被用于中医望诊的研究,如阚红星等^[16]用随机森林算法对 2 型糖尿病中 3 种证候的舌象建立辨识模型,通过舌象图,辨别 2 型糖尿病的中证候。

以上所述的机器学习算法在中医辨证研究中有广泛的应用,其共同点是均可建立智能化的中医辨证模型,用于中医智能辨证。但是,由于证候是一个多维多阶的复杂巨系统,面对的是高度非线性的研究对象^[17],因此单一使用某一算法建立辨证模型,往往不能满足需求。例如临床医师在病历中记录的症状往往对应着多个证候的诊断,使症状与诊断结果间形成“多对多”的对应关系。若要求辨证模型也做到在记录下输入的多个症状后,输出多个证候诊断,形成输入症状与输出结果间“多对多”对应的关系,辨证模型则需要放宽输出的置信概率输出多个辨证结果,以覆盖可能对应的多个证候。但这种方式难以把握置信概率的大小,容易输出包含错误的辨证结果,无法保证正确性,难以进一步根据输出的辨证结果选择正确的方剂。因此探索可用于输入症状与输出方剂之间是“多对多”对应关系的辨证选方模型,对建立用于临床的中医智能化辅助诊疗系统具有较好的发展潜力。

2 复合结构的智能化辨证选方模型

以往因计算机运算速度所限,多种算法的结合难以实现,如今计算机运算性能显著提升,配置机器学习的运行环境日趋便利,数据库管理系统处理数据的功能不断丰富,都为本复合设计提供了可行性。使本观察组可以创新结合以上提及的多种算法设计一种复合结构的智能化辨证选方模型。使模型可对

辨证结果进行验证,剔除错误结果,提高模型在多方证对应时,输出结果的准确性,实现精准的辨证选方。

2.1 模型总体结构与实现精准辨证选方的方式

复合结构的智能化辨证选方模型是由辨证层级、验证层级、选方层级 3 个层级构成,分别执行辨证、验证、选方的任务,在层级之间,前一个层级的输出作为下一层级的输入,使各层级构成链式的数据传递,形成如图 1 所示结构与运行过程。使模型在输入症状与输出方剂之间存在“多对多”对应的联系时,准确输出辨证选方的结果。

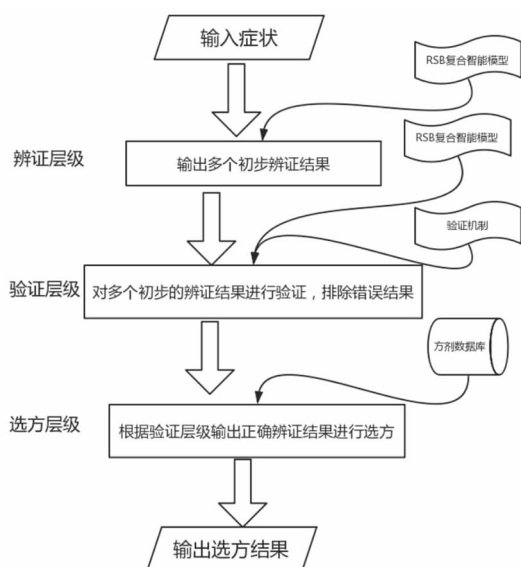


图 1 复合结构的智能化辨证选方模型结构

模型实现准确辨证选方的方式主要有以下 2 点:一是在辨证层级中构建复合型智能辨证模型,例如本文使用随机森林、支持向量机、BP 神经网络算法共同构建辨证模型,以 R 代表随机森林(Random Forest)算法构建的辨证模型,S 代表支持向量机(Support Vector Machine)算法构建的辨证模型,B 代表 BP 神经网络(Back Propagation Neural Network)算法构建的辨证模型。这种多算法共同构建的智能辨证模型,可简称为 RSB 复合型智能辨证模型。该辨证模型可在给定输出的辨证结果数量后,由 R、S、B 分别根据输入症状从各自不同的角度给出辨证结果,为结果验证提供更多的可能。二是在验证层级中植入一种可对辨证结果进行检验的验证机制,通过验证机制对辨证层级输出的辨证结果进行验证,排除错误结果,保留正确结果,并将正确的辨证结果送往选方层级实现精准选方。

2.2 模型各层级的结构

2.2.1 辨证层级 智能模型的第一层级是辨证层

级,模拟中医以症状辨证候的过程,与验证层相连形成链式关系。该层级是由用于辨别证候的 RSB 复合智能辨证模型构成,使用症状-证候数据(以症状作为特征,证候作为标签)训练。在模型训练完成后,通过设定输出的结果数量 N,输出前 N 个概率最高的辨证结果。但这里得出的 N 个结果是以放宽置信概率为代价,所以往往包含错误结果。因此在该层级初步得出辨证结果后,还需再将辨证结果送入验证层级进行验证,对错误进行排除。

2.2.2 验证层级 第二层级是验证层级,运用的是以证候推测症状的思想,因其上连辨证层级,下连选方层级,所以在 3 个层级的链式关系中扮演着承上启下的核心角色。该层级由验证机制、RSB 复合智能辨证模型组成。其中,验证机制的工作是根据辨证层级输出的证候,从症状-证候训练集中逆向提取出相应的症状,并在提取出相应的症状之后,再将提取的症状分别与辨证层级输入的症状取交集,依次提取出共有症状,模拟中医以证候推测症状的过程。RSB 复合智能辨证模型的工作则是根据依次提取出的共有症状输出概率最高或满足输出条件的辨证结果,完成模型的再次验证,以此来避免因放宽置信概率导致的误差。

2.2.3 选方层级 第三层级是选方层级,与验证层级相连,主要是存储证候与对应方剂的数据库。在该层级中,对相连的验证层级输出的多个证候查询出数据库中对应的证候与方剂,并且为保证方剂的使用安全在查询出方剂后,检索组合后的方剂是否包含“十八反十九畏”,若包含则用相近功效的中药进行代替,最终实现辨证选方。

2.3 模型验证 复合结构的智能化辨证选方模型是通过辨证层级、验证层级、选方层级完成辨证选方,按照这种层级结构的设计,在辨证选方时,输出的结果应有较高的准确性。因此,我们对该模型进行了实现,并同基于单一算法建立的辨证选方模型,比较输出结果的准确性。

2.3.1 模型的构建与测试数据 选择第 3 版“十三五”教材《伤寒论讲义》^[18]中,辨治要点所记录的主症、病机、方药等相关数据,建立用于构建辨证选方模型的训练集与选方数据库。其中,主症与病机数据作为训练集(主症作为特征,病机作为类别,形成症状-病机训练集),用于实现模型的辨证功能。方药数据则通过数据库管理系统(使用的是 MySQL 数据库管理系统),建立选方数据库,使模型可根据辨证结果查询对应的方剂,实现选方功能。对于模型

的测试数据,则是通过随机抽取训练集中的 6 个样本产生,除此之外,如临床中也存在着四逆散与半夏泻心汤合用,旋覆代赭汤与四逆散合用,四逆汤与柴胡加龙骨牡蛎汤合用,大柴胡汤与芍药甘草汤合用等多方证对应的情况。因此本研究也使用以上合用方剂的主症,与随机抽取的 6 个样本共同用于测试辨证选方模型输出结果的准确性。测试的示例与选方数据库中的内容如表 1 与表 2 所示。

表 1 测试数据与对应结果示例

模型应输出的方剂	模型应输出的病机	输入模型的测试数据(主症)
.....		
桂枝加附子汤	风寒袭表,营卫失调,阳虚不固	发热,恶风,漏汗不止,瘦短,肢挛
小陷胸汤	痰热互结心下	心下硬满,按之疼痛,胸闷喘满,咳吐黄痰,舌红,苔黄腻,脉浮滑
旋覆代赭汤+四逆散	胃虚气逆,痰气壅塞+少阴阳气郁遏不达	心下痞硬,暖气不除,四肢逆冷,脉弦
.....		

表 2 选方数据库内容示例

病机	方剂名称	方剂组成
.....		
风寒袭表,营卫失调,阳虚不固	桂枝加附子汤	桂枝、芍药、生姜、炙甘草、大枣、炮附子
痰热互结心下	小陷胸汤	黄连、半夏、栝楼实
少阴阳气郁遏不达	四逆散	柴胡、芍药、枳实、甘草
.....		

2.3.2 建立复合结构的智能化辨证选方模型 本模型使用 Python 编程语言结合 scikit-learn,按照上述模型的各层级结构进行建立。其中,构建 RSB 复合辨证模型所使用的训练集与选方层级所使用选方数据库,均如 2.3.1 中所述。对该模型验证层级中 RSB 复合模型所输出的条件,限定为置信概率大于 0.5,若未大于 0.5 但主症多于 3 个则取概率最高的结果。

2.3.3 建立基于单一算法的辨证选方模型 该辨证选方模型同样使用 Python 编程语言与 scikit-learn 构建。其构建方法是先调取 scikit-learn 中的随机森林、支持向量机、BP 神经网络算法,并分别使用上述 2.3.1 中的训练集进行训练,建立基于单一算法的辨证模型。再将构建出的辨证模型结合 2.3.1 中的选方数据库,实现根据辨证结果查询出对应方剂的功能,完成辨证选方模型的建立。基于单一算法的辨证选方模型的构建流程如图 2 所示。

2.3.4 模型的测试结果 基于单一算法构建的辨证选方模型与复合结构的智能化辨证选方模型的准确性测试如表 3 所示。其中表中模型的测试内容是方剂的主症,用于测试不同方法构建的模型能否根

据输入的方剂主症正确输出所对应的病机与方剂,若输出正确则以“√”表示,若输出错误则以“×”表示。结果显示,所有的辨证选方模型均能根据葛根汤、桂枝加附子汤、大青龙汤、小陷胸汤、猪苓汤、柴胡加龙骨牡蛎汤的主症,正确输出所对应的病机与方剂。但是对四逆散+半夏泻心汤、旋覆代赭汤+四逆散、四逆汤+柴胡加龙骨牡蛎汤、大柴胡汤+芍药甘草汤对应的主症,只有复合结构的智能化辨证选方模型输出的结果准确。

调用算法 (BP神经网络算法, 随机森林算法, 支持向量机算法)

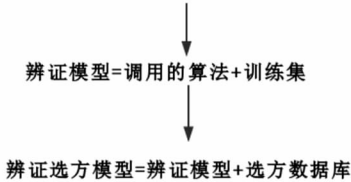


图 2 基于单一算法的辨证选方模型的构建流程

2.3.5 结果分析 通过对辨证选方模型的准确性比较,可看出单一算法构建的辨证选方模型与具有复合结构的辨证选方模型,在面对单一方剂的主症时,都有正确的结果。而一旦将 2 个方剂的主症合并后一同输入模型,则最终输出的选用方剂也应是 2 个,即临床中的合方。面对这种情况时,复合结构的智能化辨证选方模型的准确性明显高于单一算法构建的辨证模型。这是因为单一算法构建的辨证选方模型结构单一,仅输出概率最高的一个结果,这在输入模型的症状仅对应一个治疗方剂时,可保证准确。但面对 2 个方剂的主症合并后一同输入模型时,单一算法构建的辨证选方模型在辨证阶段仍只输出一个结果。即使要求其输出多个结果,由于结构单一,模型也难以具体确定输出多少结果或是确定输出的概率,从而容易导致错误的输出结果,并且不能排除。因此单一算法构建的辨证选方模型无法根据辨证结果从选方数据库选择正确的方剂。而复合结构的智能化辨证模型通过 RSB 复合智能辨证模型与验证机制的结合则实现了对辨证结果的二次验证,保证了辨证结果的准确。

3 展望

以上验证表明本创新模型可从多方证症状群中,精确提取对应的方证,提高辨证选方的准确性。但这仍然只是方证对应,若该模型能进一步结合一种可以紧扣病机的组方原则,使模型从方证对应的角度,难以找到适合的方剂时,能以病机为导向,从现有方剂中选择对病机起治疗作用的最佳药物组合,组建新方,则可扩展模型的应用范围。

表 3 单一算法构建的辨证模型与复合结构的智能化辨证模型的辨证准确性比较

模型的测试内容	各模型辨证的准确性			
	基于支持向量算法机构建的辨证选方模型(S)	基于随机森林算法构建的辨证选方模型(R)	基于 BP 神经网络算法构建的辨证选方模型(B)	复合结构的智能化辨证选方模型(RSB)
葛根汤的主症	√	√	√	√
桂枝加附子汤的主症	√	√	√	√
大青龙汤的主症	√	√	√	√
小陷胸汤的主症	√	√	√	√
猪苓汤的主症	√	√	√	√
柴胡加龙骨牡蛎汤的主症	√	√	√	√
四逆散 + 半夏泻心汤的主症	×	×	×	√
旋覆代赭汤 + 四逆散的主症	×	×	×	√
四逆汤 + 柴胡加龙骨牡蛎汤的主症	×	×	×	√
大柴胡汤 + 芍药甘草汤的主症	×	×	×	√

注:√表示输出的病机、方剂与输入模型的主症完全对应,辨证选方准确;×表示输出的病机、方剂同模型输入的主症不对应或不完全对应,辨证选方不准确。根据表中各模型的辨证选方准确性,所有的辨证模型均能根据葛根汤、桂枝加附子汤、大青龙汤、小陷胸汤、猪苓汤、柴胡加龙骨牡蛎汤对应的主症,正确输出所对应的病机与方剂。但对四逆散 + 半夏泻心汤、旋覆代赭汤 + 四逆散、四逆汤 + 柴胡加龙骨牡蛎汤、大柴胡汤 + 芍药甘草汤对应的主症,单一模型 R、S、B 由于难以确定第二个结果的输出概率,因此出现了结果不准确的情况,而包含 RSB 与验证机制的复合模型表现出精准的输出结果

我们在 2009 年提出了“方-证要素对应”的组方原则^[19],其中“方-证要素对应”是指“方剂要素”与“证候要素”的对应,即方剂中的主要药物组成与这些药物所主治的病机单元对应,使药物与病机之间形成靶向性的精确对应关系。运用“方-证要素对应”的组方原则,可在面对复合病机(即多个病机的组合),难以直接找到恰好对应的主治方剂时,将复合病机拆分为“证候要素”,使用与其对应的“方剂要素”进行组方。

综上所述,若本创新模型与“方-证要素对应”的组方原则结合,使用“方剂要素”与“证候要素”数据构建,则可建立出基于“方-正要素对应”的智能化辨证组方模型。该模型将能在“方-证要素对应”的基础上,进行智能组方,可用于辅助临床医生在病机比较复杂,方证对应难以适用时,从“方-证要素对应”的角度出发,紧扣病机,组建新方。希望能够利用人工智能技术,为中医复杂病机的“辨证论治”,提供新方法、新思路。

参考文献

[1] Mitchell TM. 机器学习[M]. 北京:机械工业出版社,2003:3.
[2] Williams D,Phillips G,Sekuler R. Hysteresis in the perception of motion direction as evidence for neural cooperativity[J]. Nature,1986,324(6094):253-255.
[3] 田禾,戴汝为. 基于人工神经网络的中医专家系统外壳 NNS [J]. 计算机学报,1990,12(5):397-401.
[4] 陈肇文.《伤寒论》的“方一证要素”对应体系及其神经网络数学模型的构建[D]. 北京:北京中医药大学,2011.
[5] 徐亮,陈守强,侯建辉,等. 基于 BP 神经网络的中医辨证模型构

建方法探讨[J]. 世界中医药,2016,11(2):335-338.
[6] Quinlan JR. C4. 5:Programs for machine learning[M]. San Francisco:Morgan Kaufmann Publishers,1993.
[7] 杨开明. 糖尿病中医临床数据挖掘技术研究[D]. 昆明:昆明理工大学,2013.
[8] 李治,李国琳. AECOPD 中医辨证分型决策树模型[J]. 长春工业大学学报:自然科学版,2013,34(3):296-299.
[9] Cortes C,Vapnik V. Support-vector networks[J]. Machine learning,1995,20(3):273-297.
[10] 晏峻峰. 基于数据挖掘技术的证素辨证方法研究[D]. 长沙:湖南中医药大学,2007.
[11] 王阶,吴荣,周雪忠. 基于支持向量机的名老中医治疗冠心病证候要素研究[J]. 北京中医药大学学报,2008,31(8):540-543,560.
[12] 王华珍,胡雪琴. 中医“内生五邪”的智能证型分类[J]. 计算机工程与应用,2011,47(6):156-160,163.
[13] Breiman L. Random forests[J]. Machine learning,2001,45(1):5-32.
[14] 孙超. 糖尿病肾病中医证候分类预测模型的构建及其识别性能的评估[D]. 北京:北京中医药大学,2015.
[15] 蔡晓路. 基于随机森林的类风湿关节炎证型判别模型研究 [D]. 北京:北京中医药大学,2016.
[16] 阚红星,张璐瑶,董昌武. 一种 2 型糖尿病中医证型的舌图像识别方法[J]. 中国生物医学工程学报,2016,35(6):658-664.
[17] 王永炎,张志斌. 再议完善辨证方法体系的几个问题[J]. 天津中医药,2007,24(1):1-4.
[18] 李赛美,李宇航. 伤寒论讲义[M]. 3 版. 北京:人民卫生出版社,2016.
[19] 李宇航. 谈“证候要素”与“方剂要素”[J]. 中华中医药杂志,2009,24(2):117-121.

(2017-11-15 收稿 责任编辑:杨觉雄)