

DOI:10.13288/j.11-2166/r.2022.08.005

机器学习方法在中医学传承研究中的应用及思考

刘福栋, 姜晓晨, 王桂彬, 庞博[✉], 花宝金

中国中医科学院广安门医院, 北京市西城区北线阁5号, 100053

[摘要] 推动中医药传承是保障中医学经久不衰的动力, 随着中医学理论体系的不断丰富, 传统的传承模式已经无法满足当今中医学传承的需求。在计算机技术飞速发展的当下, 机器学习在中医学传承中凸显出了重要地位。梳理并总结比较基于机器学习的两种基本形式, 即监督学习和无监督学习在相关中医药研究中的应用, 认为监督学习更适合“病-症-证”的研究, 在临床常见疾病的辨证模型、风险预测模型与证素危险因素研究中更具应用价值, 而无监督学习更适合“方-药”研究和隐性知识的发现。基于对当前机器学习平台的比较, 认为自研平台应当拓展算法类型, 以实现与专业数据挖掘平台的更好结合。基于中医药传承研究与大数据科学结合的发展现状, 认为机器学习将在中医理论的传承创新、临床决策和新药研发等方面产出更有意义的成果, 但还需进一步拓展学科交叉的广度和深度, 开发算法更加优化的机器学习平台。

[关键词] 中医学学术传承; 机器学习; 监督学习; 无监督学习; 大数据

传承是中医药事业发展的源泉, 是提升中医药服务能力的核心工程^[1], 但因为中医学独特的认知体系和经验方法论, 传统的统计方法难以全面真实地继承中医学的思辨体系和经验逻辑。近年来, 随着人工智能、信息技术的不断成熟, 机器学习(machine learning)方法作为数据挖掘技术的基本形式, 在中医传承研究中得到了广泛的应用。通过机器学习算法, 可以从证候、用药频次、用药配伍、核心处方等方面分析临证经验和用药规律, 可以从古代医家、现代医家的角度, 将机器学习的结果与医家的学术思想相结合, 构建中医认知模型, 促进中医学的传承和创新。但由于数据内部变量间的关系错综复杂, 建立的模型之间会存在差异^[2]。因此, 在众多机器学习方法中, 如何选择适合中医思维的技术, 客观真实地继承中医学的诊疗经验, 已成为中医传承体系构建的关键问题。现结合近年来机器学习在中医传承中的研究成果, 基于监督学习和无监督学习两种基本形式, 就其实践意义、成效、问题和未来发展方向分析如下。

1 监督学习方法在中医学传承研究中的应用

监督学习(supervised learning)通常需要先确定一个目标属性或因变量, 然后筛分数据找出与其他属性以及属性之间的关系。目前广泛应用的包括贝叶斯网络、决策树、Logistic回归分析、随机森林算法、支持向量机等。监督学习主要可以用来解决分类及回归问题, 适合“病-症-证”的研究, 在临床常见疾病的辨证模型、风险预测模型与证素危险因素研究中更具应用价值。

1.1 在传承名老中医临床经验中的应用

名老中医学术思想和诊疗经验是中医药知识的精华与载体, 是将中医药学理论、前人经验与当今实践相结合解决临床疑难问题的典范, 发现其中蕴含的概念、原理、规律、法则, 是肯定疗效、规范标准、发现机理、传承学术的基础^[3]。甘小金等^[4]在王子瑜教授及其学术传承人对150例子宫内膜异位症患者的症状信息进行辨证的基础上, 运用贝叶斯网络方法对子宫内膜异位症症状-证素信息做分类识别, 得出的结论与王子瑜的学术思想完全符合, 说明贝叶斯网络在传承名老中医学术经验方面具有较好的应用前景。秘红英等^[5]采用贝叶斯算法及回顾性临床研究方法对346例吴以岭教授治疗

基金项目: 中国中医科学院科技创新工程(CI2021A01805); 首都卫
生发展科研专项(首发2022-2-4155)

✉ 通讯作者: pangbotommy@gmail.com, (010)88001248

的冠心病心绞痛临证验案进行证型分类,并提取各个证型与症状间的关联性,分析吴以岭教授病案中药物的功效、用药频率、药物性味及归经,总结了吴以岭教授治疗冠心病的“病、证、症、法、方、药”特点,有利于其学术思想及临证经验传承发扬。

1.2 在中医证候学研究中的应用

近年来以名老中医学术思想、临床经验总结与思辨规律为代表的诸多研究,以其多样化的继承方法与传承方式,推动着中医学由“经验”向“知识”再向“证据”转化路径的日趋完善,促进了中医学术传承与发展。辨证论治是中医认识和诊治疾病的基本原则,探索名老中医专家辨证论治方法和规律对于传承其学术思想具有重要作用。利用机器学习方法,可以将“病-症-证-法-方-药”之间的关系数据化、可视化,从而合理揭示中医学内在的辨证规律。余学杰等^[6]对经专家辨证的803例患者的103种证候进行决策树方法分析,获取了9种病名和证候规则,如关于血瘀证的两条规则:“FB8=‘1’ and FB70E=‘1’为血瘀证(如果口唇紫黯且舌黯则为血瘀证);FB8=‘1’ and FB70E=‘0’ and FB9=‘1’为血瘀证(如果口唇紫黯且睑下青黑则为血瘀证)”,证实了利用决策树方法获取专家辨证规律的可行性。赵红波等^[7]对422例功能性便秘患者进行病例信息和临床资料采集,并采用Logistic回归分析法探讨功能性便秘的相关危险因素,结果证实功能性便秘与湿热质、气虚质、阴虚质、阳虚质、气郁质五种体质密切相关,其所运用的量化方法为我们后续的研究提供了思路。

1.3 在中药作用机制研究中的应用

中药具有多成分、多靶点和作用途径复杂等特点,单味中药及其复方在中医学整体观念和辨证论治理论体系的指导下具有十分显著的临床疗效。但中药的现代化研究,尤其是创新药物研发过程中存在诸多问题,对于中药尤其是复方的有效物质基础和分子作用机制尚不明确。而基于机器学习方法,结合网络药理学阐释中药和复方的作用靶点并针对复方进行优化,是中医药学发展守正创新的生动体现。苑婕等^[8]应用随机森林算法构建药物-靶点模型,预测出延胡索治疗心脑血管类疾病的关键靶点,并得到较好的文献验证,为应用现代科学方法研究中药作用机制提供了思路。郑婷婷等^[9]基于随机森林模型结合网格化搜索算法,从抑制肺癌细胞增殖方面对中药复方金复康的组成进行优化,优

化结果得到了实验验证,为复杂性中药复方的组方优化提供了方法学参考。陈永新等^[10]运用支持向量机法建立了区分20种中药寒热药性的数学判别模型,结果表明20种中药游离脂成分与其寒热药性存在明显的相关性,支持向量机可以作为区分寒热药性的统计学工具。

2 无监督学习方法在中医传承研究中的应用

无监督学习(unsupervised learning)无明确学习目标,通过聚类算法寻找数据集中的规律性并进行分析,主要用于缺少目标属性的分析情况,包括关联规则、聚类分析、信息熵、主成分分析等。该算法在不依赖任何预设目标的前提下发掘数据的关联和聚类,可以发现中医学理论中的潜在知识和经验。

2.1 在古今文献知识传承研究中的应用

中医药文化流传已久,古今医家流传的大量学术论著成为了中医学理论体系的重要组成部分,然而几千年来所积累的巨大信息量,依靠脑力运算十分困难,因此诸多学者运用无监督学习方法,以期准确提取古今医家的临证思维和用药规律。易亚乔等^[11]通过对历代医家治疗痴呆的156首方剂,运用频数统计与关联规则分析得出人参、茯苓、远志、石菖蒲、白术的加减为历代医家治疗痴呆的核心组合;王伟斌等^[12]利用古今医案云平台对《医案类聚》中治疗郁证的439条处方进行聚类分析、关联分析和复杂网络分析发现,古代治疗郁证以健脾补益气血、疏肝理气、养心安神为核心治法,以白术、茯苓、甘草、当归、人参、陈皮为常用组合,以丹栀逍遥散为治疗郁证的核心方药。这些研究皆有助于指导现代中医的临床实践。此外,魏丹妮等^[13]以Apriori算法对《中医方剂大辞典》中治疗4种月经病的1761首方剂进行了数据挖掘,并对关联规则进行深入分析,得出治疗月经不调、崩漏、闭经、痛经4种月经病的共同核心方剂为四物汤,治疗崩漏的核心药物为黄芪,治疗闭经的核心药物为桃仁和大黄,治疗痛经的核心药物为香附;Zhang等^[14]通过检索PubMed、中国知网等常用中、英文数据库搜集到124篇肺纤维化的中医药治疗文献,使用聚类结合链接分析,得出治疗肺纤维化的配伍组合为黄芪、党参、麦冬、五味子、地龙、丹参、当归、浙贝母、半夏、桃仁、苦杏仁、紫菀、甘草,其研究结果具有一定临床意义。

2.2 在临床文本研究中的应用

以临床病例为代表的临床文本是医生诊疗活动的准确记录,蕴含着医者的学术思想和临床用药的潜在规律。然而大样本的临床文本内容繁多且复杂,故采用信息技术对中医临床案例、数据进行分析,挖掘其内在组方配伍规律,是中医传承与发展的有效途径。王师菡等^[15]基于信息熵分析中国中医科学院广安门医院心内科名老中医专家的319首处方,提取出他们治疗冠心病的常用药对,为名老中医学术经验的传承和发展提供了新的思路与方法。郭仁清^[16]对国医大师周仲瑛辨治大肠癌388例1286诊次的原始病案进行关联规则分析,得出“扶正健脾,抗癌解毒”为基本治法、整体用药以四君子汤加减的结论,对于传承周仲瑛教授治疗大肠癌的用药经验具有实践意义。程建超等^[17]对安徽中医药大学第一附属医院792例患者的病例资料进行回顾性分析,通过聚类分析和关联规则算法,发现该院肺病学科治疗肺癌强调益气养阴、补脾益肺以治其本,清热解毒、祛痰散结以治其标的治疗原则,且在临床疗效中,白花蛇舌草、半枝莲配伍补虚药和其他攻邪药与肿瘤标志物的改善具有高关联度。姚魁武等^[18]基于血瘀证诊断的多中心、大样本临床资料,将血瘀证量化诊断流行病学调查的结果运用主成分分析进行处理,筛选出血瘀证与非血瘀证鉴别诊断较好代表性的条目,优化了血瘀证的量化诊断标准。

以聚类分析、关联规则为代表的无监督学习算法简单,易被接受与解释,在探索挖掘高频药对和用药规律方面得到广泛应用,适合对学习内容中的“方”“药”进行分析,有利于在继承学术思想的同时开发有价值的新处方,对于中医临床具有直接指导意义。

3 机器学习平台在中医传承研究中的应用

3.1 自主研发平台推动机器学习与中医传承结合

目前比较完善且应用广泛的中医经验传承信息和数据技术平台包括中国中医科学院研制的中医传承辅助平台^[19],该平台将一般统计法、文本挖掘、关联规则、复杂系统熵方法等挖掘分析方法加以集成,并将相关的分析算法进行软件化,成为中医组方用药经验传承的便利有效工具,近年来得到了广泛应用。中国中医科学院中医药信息研究所研制的古今医案云平台^[20]和中国中医科学院广安门医院研制的名老中医临床信息采集系统^[21]立足科研和

临床一体化的原则,集病案的采集、自动结构化、集中存储管理、后标准化处理、数据挖掘分析于一体,通过网络可视化展示增进其便利性,极大地推动了名医经验传承。然而这些自研平台在带来便利的同时,由于学科交叉深度不足,主要进行关联规则和信息熵算法,算法相对单一,存在创新性不足等问题。

3.2 专业平台推动机器学习与中医传承深度融合

除自主研发的分析平台外,还有许多专门用于或可以用于机器学习、数据挖掘的商业或非商业平台,如Hadoop, SPSS (Statistical Product and Service Solutions), MATLAB (Matrix Laboratory), Weka (Waikato Environment for Knowledge Analysis), Intelligent Miner, WinBUGS (Windows Bayesian inference using gibbs sampling)等。其中Weka是一款由新西兰怀卡托大学开发、基于JAVA环境下开源的机器学习和数据挖掘的非商业化软件,是现今较完备的数据挖掘工具之一,集合了大量能承担数据挖掘任务的机器学习算法。在Explorer模块中可以借助其实施过滤算法、学习算法、元学习算法、聚类算法、关联规则学习、属性选择、可视化等功能。

本研究团队通过对朴炳奎教授诊治1523例肿瘤医案数据进行机器学习和数据挖掘,依托Weka软件构建疾病、证候、症状、治法、处方、用药六类认知模型200余个,完成了名老中医诊治肺癌认知模型构建研究^[22]。并进一步通过引入原型范畴理论,探索了模型优化方法^[23],运用监督学习和无监督学习相结合的机器学习方法,继续扩大样本量至2000余例,引入朴炳奎诊治肺癌范畴化证素属性,有效提高了朴炳奎诊治肺癌“用药”认知模型分类的准确率。运用专门的机器学习、数据挖掘平台得出的数据更具科学化,但同时也存在输出的数据相对复杂且不够集成,需要有机地结合中医药传统理论进行深度分析,才能对挖掘到的数据模式进行合理解释。因此,临床中医师受到编程以及数据分析能力欠缺的限制,很难凭借自身的知识储备与能力完成研究任务。

4 基于机器学习的中医学传承研究存在的问题及发展对策

4.1 中医学传承需要新的技术和模式

中医药的传承方式自古以来以师承与学校教育为主,更有师承与学校规模教育并存的组合,有业

师授受、家学相传、私淑遥承等多种形式^[24]。此类教学-学习-领悟的培养模式培养了古今诸多名家,且历经千年,在医术、医理和医道方面对中医进行了全面的传承。但中医学术思想不是单纯的临床经验,其源于一病、一证、一法、一方、一药的诊治经验与心得体会,又在此基础上进行高度的抽象概括和理性提升^[25],是由感性的、繁杂的现象升华为理性的、系统的、可以再现的技术的过程^[26]。这无疑对学习提出了更高的要求,且随着中医学理论体系的逐渐积累,传统的领悟式学习方法已不能满足需求。因此,将新技术应用于名老中医经验传承,培养具备交叉学科知识的中医复合人才势在必行。

4.2 中医学的科学和人文双重属性是机器学习的难点

中医药学具有科学与人文的双重属性,是科技成果与人文准则相融合的医学^[27],且中医学理论体系是哲学观、整体观、恒动观的统一体,而机器学习作为一种数字化学习方式,必将在中医学传承中遇到诸多难点。

首先,中医学传承的过程不仅仅是医术、医理方面的传承,日常跟师学习过程中对医道的领悟同样是中医学传承的重点,但机器学习虽然可以在构建“病-症-证-法-方-药”认知模型的基础上基于大量的数据集进行自主学习,但无法学习与体现中医学所具备的人文思想。因此,在使用机器学习方法传承的同时,必须充分结合人的自主性和机器学习的客观性,实现中医药的全面传承。

其次,中医医案有重视个案、突出创新、实用性和文学性四大特点^[28],尤其是古代医案,记载格式个性化强,言辞古奥,在数据维度和完整性上与现代医案相差较大,但又极具传承价值。因此,需要经过研究者事先进行规范化处理,实现古今医案文献的共通共享,才能进一步引入机器学习方法。

最后,中医语义变迁和叙词表的变化为机器学习带来了困扰。比较典型的是药物计量单位的变化。据《中国科学技术史·度量衡卷》提供的数据^[29],宋元时期一两约合现在的 41.3 g,明清时期一两约合现在的 37.3 g,且医籍经过历代传抄,度量衡单位又十分混乱,汉以后唐宋医方中所用的计量单位也经常出现一些非日常所用的单位。李时珍《本草纲目·叙例》中的注释涉及陶弘景(南北朝)、李杲(金)和李时珍(明)三个不同时期的医学家对重量单位的记载,给药物剂量的挖掘和机

器学习造成了极大的阻碍,需要严谨的文献学考证以解决此类问题。

4.3 人工智能、数据挖掘和机器学习在中医药传承领域具有巨大发展空间

人工智能是机器形式的人类智能,计算机程序可以指定基于数据的决策并执行任务。机器学习作为人工智能的一个分支,它能够从给定的训练数据集中运用不同的算法进行数据解析及学习,并以此为基础构建模型从而对外部结果进行预判。机器学习使人工智能向前迈了一步,赋予了人工智能自主决策的能力。在数据挖掘领域,机器学习可以用于数据挖掘,也可以用于强化学习、自动控制等方面。就其应用范围而言,机器学习和数据挖掘拥有较大的重叠,但相比较而言,机器学习又具有更广阔的应用空间。有学者认为,数据挖掘更加侧重于“目的”,而机器学习则更偏重“方法”^[30]。由此可见,人工智能、数据挖掘和机器学习在技术领域存在诸多重叠。我们可以基于数据挖掘技术,通过对大样本的、不完全的、有噪声的、模糊的中医临床数据的研究,从现象中提取出隐藏的规律;并将复杂的中医学理论进行可视化数据表达,结合人工智能,构建中医认知模型,为传承中医理论、学术观点和临床经验提供依据;最终运用机器学习的方法,使用更大的训练集,赋予人工智能自主决策能力,不断对认知模型进行完善和优化。

5 小结

历经几代人的努力和大数据的飞速发展,机器学习与中医传承的融合不断深入,借助机器学习方法,挖掘新知识,对中医学理论进行定性、定量的描述已经成为当前研究的热点,使名老中医经验得到了有效的传承,中医学理论体系得到了不断丰富。未来基于机器学习技术,将中医经典古籍和名老中医经验构建为认知模型,不仅可以形成全新的中医药信息化传承模式,亦可以支撑临床决策支持系统的构建,提高诊疗的安全性,减少医疗错误的发生率。同时,中医学还蕴含高概念大数据时代的特征^[27],通过对大样本、多学派的数据挖掘和机器学习,实现数据可视化,可以对比学派之间的学术观点、诊疗经验和用药特点的差异,进一步指导各学派之间的融合发展,使中医药理论创新走上新高度。

综上所述,机器学习作为人工智能的分支和数据挖掘的基本方法,结合现代中医学蕴含的大数据

的时代特征, 进一步加深学科交叉的广度和深度, 开发适合中医传承的、算法与时俱进的、输出结果简洁精确的中医药传承平台, 必将在中医理论的传承创新、临床决策和新药研发等方面产出更有意义的成果。

参考文献

- [1] 孙光荣. 中医药学传承的内涵外延与路径[J]. 中医药通报, 2016, 15(4): 1-2.
- [2] 史忠植. 知识发现[M]. 北京: 清华大学出版社, 2002: 2-6.
- [3] 王永炎, 于智敏. 培养传承博士后与创新中医药学术[J]. 中医杂志, 2011, 52(5): 361-362.
- [4] 甘小金, 陈艳, 马秀丽. 基于贝叶斯网络的王子瑜教授治疗子宫内膜异位症的辨证规律研究[J]. 世界中西医结合杂志, 2019, 14(10): 1350-1352.
- [5] 秘红英, 魏聪, 李红蓉, 等. 基于贝叶斯算法的吴以岭教授治疗冠心病心绞痛的病案分析研究[J]. 中国中西医结合杂志, 2018, 38(5): 534-538.
- [6] 余学杰, 李书珍, 李晓燕, 等. 基于决策树提取中医专家辨证规律初探[J]. 辽宁中医杂志, 2015, 42(1): 19-24.
- [7] 赵红波, 吴晓晶, 杨云, 等. 慢性功能性便秘与中医体质类型及相关危险因素的 Logistic 回归分析[J]. 中医杂志, 2017, 58(16): 1393-1398.
- [8] 苑婕, 王珍珍, 宋丽娟, 等. 基于监督学习算法的延胡索成分-靶点-疾病网络的预测研究[J]. 海南医学, 2020, 31(13): 1638-1643.
- [9] 郑婷婷, 杨铭, 阙祖俊, 等. 基于随机森林回归模型的中药复方金复康对肺癌细胞增殖作用的组方优化[J]. 中国实验方剂学杂志, 2017, 23(4): 177-182.
- [10] 陈永新, 李峰, 孙稚颖, 等. 基于游离脂成分的 20 种中药寒热药性支持向量机分析[J]. 辽宁中医杂志, 2011, 38(1): 127-129.
- [11] 易亚乔, 方锐, 葛金文, 等. 基于数据挖掘的历代医家痴呆相关病症用药规律分析[J]. 中国中药杂志, 2018, 43(16): 3376-3381.
- [12] 王伟斌, 李敬华, 于琦, 等. 基于古代医案的郁证用药规律分析[J]. 中国实验方剂学杂志, 2020, 26(5): 162-167.
- [13] 魏丹妮, 王璐, 齐琦, 等. 基于数据挖掘的 4 种月经病中药方剂用药规律比较研究[J]. 中草药, 2018, 49(8): 1939-1945.
- [14] ZHANG S, WU H, LIU J, et al. Medication regularity of pulmonary fibrosis treatment by contemporary traditional Chinese medicine experts based on data mining[J]. J Thorac Dis, 2018, 10(3): 1775-1787.
- [15] 王师菡, 王阶, 何庆勇, 等. 基于信息熵法提取名老中医治疗冠心病心绞痛的药对[J]. 中华中医药学刊, 2008, 26(10): 2150-2152.
- [16] 郭仁清. 基于数据挖掘的国医大师周仲瑛辨治大肠癌处方用药规律研究[J]. 辽宁中医药大学学报, 2019, 21(4): 142-145.
- [17] 程建超, 童佳兵, 朱洁, 等. 基于 792 份住院病历探讨中医药治疗肺癌的处方规律[J]. 时珍国医国药, 2020, 31(9): 2278-2280.
- [18] 姚魁武, 王阶, 武继涛, 等. 血瘀证量化诊断流行病学调查数据的主成分分析[J]. 辽宁中医杂志, 2009, 36(1): 30-32.
- [19] 唐仕欢, 申丹, 卢朋, 等. 中医传承辅助平台应用评述[J]. 中华中医药杂志, 2015, 30(2): 329-331.
- [20] 李敬华, 王映辉, 李宗友, 等. 智能化技术在名医传承中的应用与实践[J]. 国际中医中药杂志, 2020, 42(11): 1049-1053.
- [21] 王映辉, 刘保延, 姚乃礼, 等. 动态结构化名老中医临床诊疗信息采集系统研究开发与应用[J]. 中国中医药信息杂志, 2008, 15(2): 106-107.
- [22] 庞博, 刘刚, 周雪忠, 等. 基于机器学习名老中医诊治肺癌认知模型构建方法[J]. 北京中医药, 2015, 34(12): 949-954.
- [23] 庞博, 徐心瑶, 周培培, 等. 基于原型范畴理论名老中医诊治肺癌认知模型优化思路与方法[J]. 北京中医药, 2018, 37(12): 1141-1145.
- [24] 王琦. 师承论[J]. 中医教育, 2006, 25(3): 65-68.
- [25] 于智敏, 王永炎. 对中医学术思想传承的思考[J]. 中国中医基础医学杂志, 2009, 15(5): 321, 324.
- [26] 王永炎. 临床经验刍议[J]. 现代中医临床, 2017, 24(2): 1-3.
- [27] 王永炎, 田金洲. 新形势下的中医药传承与创新[J]. 北京中医药大学学报, 2018, 41(7): 533-536.
- [28] 孟庆云. 论中医医案的价值、特点和研究方法[J]. 继续医学教育, 2007, 21(19): 4850.
- [29] 丘光明, 邱隆, 杨平. 中国科学技术史: 度量衡卷[M]. 北京: 科学出版社, 2001: 391.
- [30] 于兴文. 基于机器学习与数据挖掘的名老中医用药规律研究[D]. 南京: 南京中医药大学, 2019.

(下转第 738 页)