·综述讲座·

基于大数据技术的中药饮片外观性状与内在成分数据的研究与应用Δ

谭超群[1,2*]，解达帅[1]，程小恩[2]，赵姝婷[2]，吴纯洁[1]，温川飙[2#]（1.成都中医药大学药学院，成都 611137；2.成都中医药大学医学信息工程学院，成都 610075）

中图分类号 R282.5 **文献标志码** A **文章编号** 1001-0408（2018）16-2287-04
DOI 10.6039/j.issn.1001-0408.2018.16.27

摘 要 目的：探讨大数据技术在中药饮片外观性状与内在成分数据结合分析（“内外结合”）中的应用。方法：总结现有中药饮片鉴别中外观性状与内在成分检测技术的不足，就中药饮片“内外结合”大数据库的研究现状与应用前景进行综述。结果与结论：现有的智能感观技术存在数据不完整、不准确和对多维数据缺乏整合的不足，内在成分检测方法也存在诸多不足。大数据分析流程包括前期数据采集、数据预处理、数据分析与处理、数据可视化处理以及数据应用几个阶段。收集中药饮片形、色、气、味外在性状与内在成分数据以及文献知识库信息，构建中药饮片“内外结合”大数据库，再根据大数据处理流程与算法，可实现人工智能辅助中药饮片真伪优劣鉴别，实现对中药产地的辨别区分，挖掘影响中药饮片质量的因素，为其质量控制提供规范化标准。大数据技术的应用能准确、高效地处理中药饮片“内外”多维数据，可为传统中药行业研究提供新思路，为中药饮片客观化发展提供新动力。

关键词 大数据技术；中药饮片；外观性状；内在成分

传统中药鉴别包括基源鉴定、显微鉴定、性状鉴定与理化鉴定4种方法。针对中药的形（形状）、色（颜色）、气（气味）、味（味道）的外观性状鉴定是对中药质量进行评价的重要指标，从其外观性状可以判断其原生植物（动物等）品种、生长年限、品质等。然而，目前对于中药外观性状的评价仍通过肉眼观察、口尝、鼻闻等方法来进行，不可避免地会因一些主观因素对鉴定结果的客观性与可靠性产生影响。基于这种情况，很多学者提出通过机器视觉、电子鼻、电子舌等现代技术对中药饮片的形、色、气、味等性状信息进行量化，客观化表达人工鉴别的主观经验，用以鉴别中药饮片[1]。当下大数据技术发展迅速，且已广泛应用于农业、医疗、教育、能源、国防、金融等诸多领域。引入大数据技术对中药饮片数据采集过程中积累形成的大量外观性状数据与内在成分数据结合（以下简称为“内外结合”）起来进行分析，建立其品质与外在信息间的耦合关系，可对中药饮片智能识别分类、产地智能分析推断的实现及中药质量标准的建立提供理论依据。

1 现有技术方法存在的不足

1.1 智能感官技术的不足

1.1.1 数据缺乏完整性与准确性 已有研究证明，中药材的形、色、气、味与其内在成分含量具有一定关联度，

[11] AILANI RK, AGASTYA G, AILANI RK, et al. Doxycycline is a costeffective therapy for hospitalized patients with community-acquired pneumonia[J]. *Arch Intern Med*, 1999, 159(3): 266-270.
[12] MOKABBERI R, HAFTBARADARAN A, RAVAKHAH K. Doxycycline vs. levofloxacin in the treatment of community-acquired pneumonia[J]. *J Clin Pharm Ther*, 2010, 35(2): 195-200.
[13] 范红，刘思彤，童翔. 呼吸喹诺酮类与β内酰胺类联合大环内酯类治疗非ICU住院社区获得性肺炎患者有效性和安全性的系统评价[J]. 中国循证医学杂志，2015，15(7)：802-803.
[14] 国家卫生计生委. 抗菌药物临床应用指导原则：2015年版[EB/OL]. (2015-08-27)[2017-04-10]. http://www.gov.cn/xinwen/2015-08/27/cntent_2920799.htm.
[15] MURRAY MP, GOVAN JR, DOHERTY CJ, et al. A randomized controlled trial of nebulized gentamicin in non-cystic fibrosis bronchiectasis[J]. *Am J Respir Crit Care Med*, 2011, 183(4): 491-499.
[16] ANTONIU SA, TROFOR AC. Inhaled gentamicin in non-cystic fibrosis bronchiectasis: effects of long-term therapy [J]. *Expert Opin Pharmacother*, 2011, 12(7): 1191-1194.
[17] 沈友良，鲍祥言. 咽部喷含庆大霉素致耳聋1例[J]. 新医学，1997(2)：108.

（收稿日期：2017-09-01 修回日期：2018-06-21）
（编辑：陈 宏）

Δ 基金项目：国家自然科学基金资助项目（No.81403105）；国家中药标准化项目（No.ZYBZH-Y-ZY-45）

* 硕士研究生。研究方向：中医药信息化。E-mail：645038306@qq.com
# 通信作者：教授，硕士生导师。研究方向：中医药信息化。E-mail：228237222@qq.com

但是对其形、色、气、味对应的物质基础研究较少[1]，因此采用智能感官技术进行鉴定缺乏与中药材或饮片内在成分的关联，影响数据采集的完整性。智能感官技术，例如电子鼻、电子舌等，由于仪器自身限制或传感器限制，对中药材的敏感度有限，尤其对辨识度不高的中药材检测正确率较低[2]，导致人们对其品质评判结果不能完全相信，从而极大地影响了人们的决策。

1.1.2　对多维数据缺乏整合　目前，智能感官技术在中药性状鉴别中的应用也越来越广泛，但大多数研究人员往往只依靠一两种技术对中药饮片的性状进行判定，多种分析技术的综合应用较少[3-6]，因此分析结果缺乏普适性，且各项技术得到的数据比较孤立，导致鉴别数据“各自为政”的现状，积累的大量数据分布在各自的“信息孤岛”中，未能得到整合与全面分析。而使用薄层色谱、液相色谱、质谱等多种方法用于内在成分的测定，获取的数据具有一定的复杂度，与外观性状数据的相关性研究较少，因此对多维数据的集群整合及数据分析存在一定的难度。

1.2　内在成分检测方法的不足

理化鉴别中一般采用光谱、色谱、差热分析等技术，即利用中药分子内部一些含氢基团振动的倍频和合频吸收来实现对中药的快速鉴别，但其检测结果多为定性判定，准确度有所欠缺；且在实际操作过程中，针对样本的测量需要大量有代表性且化学值已知的样品建立模型，在这样的情况下采用上述技术对小批量样品进行分析就显得不太实际。此外，由于仪器状态改变或标准样品发生变化，所建模型也需要不断更新，其稳定性与适用性均难以估值，加之在试验过程中所用模型并不是通用的，每台仪器的模型都不相同，又增加了使用的局限性。

2　中药饮片“内外结合”大数据库的构建

2.1　大数据处理流程简介

大数据处理流程包括前期数据采集、数据预处理、数据处理与分析、数据可视化处理以及数据应用几个阶段，即通过对多来源数据进行整合，结合计算机学习算法对数据进行分析预测，可得出用于展示交流的可视化图像或图形[7-8]，进而进行应用，详见图1。

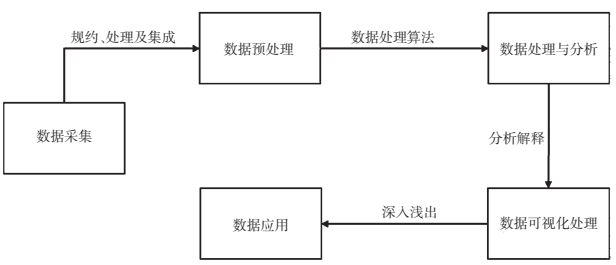

图1　大数据处理流程

2.2　中药饮片“内外”数据采集

中药饮片“内外”数据是指外中药饮片外观性状数据与内在成分数据，可利用多个数据来源渠道进行采集，例如，使用电子鼻、电子舌等设备可对中药材及饮片的形、色、气、味等外观性状数据进行实时采集；使用色谱（气相色谱、薄层色谱、柱色谱、高效液相色谱等）、光谱（紫外、红外等）、电泳、差热分析等技术可确定中药材及饮片的内在成分数据；还可检索现有文献知识库中涉及到的相应中药饮片的数据记录等。分析各类数据集合的关联度，可构建中药饮片“内外结合”大数据库（如图2所示），对所有结构化与非结构化的数据进行存储与管理，以便后续数据的查询与处理。

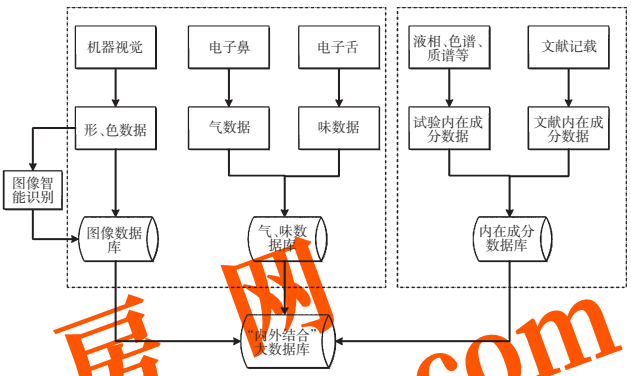

图2　中药饮片“内外结合”大数据库

2.3　数据预处理、处理与分析

由于不同来源的数据相互间易受到数据干扰产生噪声，存在数据值缺失、数据错误等问题[9]，因此需要对采集到的大量数据集合进行预处理工作，以便为后续数据处理与分析阶段提供准确、无误、真实、有效的数据，提高数据的整体质量，保证结果预测的准确性与可行性。具体操作包括数据清理、数据集成、数据规约等。数据清理是指对于中药饮片的外观形、色、气、味数据与内在成分含量数据及文献库提及的所有数据进行数据清理操作，包括数据过滤与修正[10-12]（如：对文献中记载的重复性数据加以过滤）和数据的不一致性检测（如：当试验与文献所得数据不完全一致时，应多方求证，确保数据的准确性、真实性）等。数据集成则是将不同来源的数据进行集成（如：对内在成分数据进行统一编码，统一存储格式，进行归一化操作等），形成统一的数据库，以提高数据的完整性和可用性[13-15]。数据规约用在中药饮片外观形、色、气、味数据与内在成分含量数据处理中[如：使用主成分分析法（PCA）进行数据降维，提取数据主要特征向量，旨在使用少数向量反映原始数据的特征，提高数据的价值密度]，以降低数据集规模。

数据分析是大数据处理与应用的关键环节，决定了大数据集合的价值性和可用性，以及分析预测结果的准确性。在数据分析环节，根据大数据应用情境与决策需

求选择合适的数据分析技术，可提高大数据分析结果的可用性。

2.4 数据应用

2.4.1 人工智能辅助中药饮片真伪优劣鉴别 中药饮片的质量鉴定是对中药饮片真伪优劣的检验，而通过数据挖掘技术对中药饮片“内外结合”大数据库进行处理分析，结合神经网络深度学习算法，即可实现对其“真伪优劣”的鉴别。如熏硫、炒制等传统中药炮制工艺[16-17]，对一些含糖量高但不易贮存的中药饮片的使用和贮存均有一定的积极作用。而以炒制为例，不同炒制程度的中药饮片具有不同的临床疗效，炒制温度过低或过高都会影响有效成分的活性[18]，因此判断适宜的炒制程度就显得至关重要。有研究获取不同炒制程度下的山楂“L*a*b*”颜色空间三维数值，其中L表示照度，a表示颜色从深绿色（低亮度值）到灰色（中亮度值）再到亮粉红色（高亮度值），b表示颜色从亮蓝色（低亮度值）到灰色（中亮度值）再到黄色（高亮度值）。将颜色数值与内在含量变化数据进行归一化整合，应用PCA法降维，根据人工神经网络（ANN）要求预设模型参数与反馈函数，确定最优权值与输出，建立外在性状-内在成分-炒制温度的算法模型。以“性状”数据作为ANN输入自变量（X），不同炒制程度（生、炒、焦山楂）下内在成分（有机酸、总黄酮、柠檬酸、金丝桃苷和5-羟甲基糠醛）的变化设为因变量（Y），结果该模型对3种不同炒制程度山楂的预测准确度分别为98.9%、92.5%、98.3%，得出山楂最合适的炒制温度为（150±5）℃；采用PCA法对上述预测结果进行验证，结果得PC1=94.237%，PC2=4.533%，PC3=0.693%[19-20]。可见，该算法模型对不同炒制程度下的山楂具有很好的预测性能，可实现对炮制火候的控制检测[30]，具体预测流程见图3。

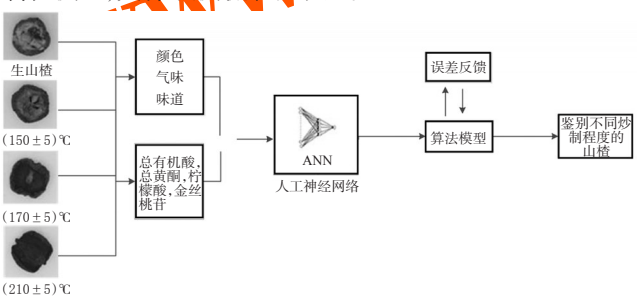

图3 山楂不同火候炒制程度的预测流程

2.4.2 对中药材产地进行分析推断 道地药材是指在特定的自然条件和生态环境的区域内所产的中药材，其生产较为集中，具有特定的栽培技术和采收加工方法，质优效佳。鉴于不同产地的中药材其功效可能有所差别，因此采用合理的检测技术，并结合有效的分析方法来分辨中药材产地显得尤为重要。

基于大数据的中药材产地判别，指在现有中药材产地数据之上进行各种算法的计算[21]，如机器学习算法K均值（Kmeans）、支持向量机（SVM）算法、深度学习算法等[22-25]，形成不同算法模型，挖掘大数据集合中的数据关联性，从而得到中药材产地预测结果。陶梦琳等[26]收集多产地黄连样本，采集“内外”数据，建立了基于SVM算法的黄连饮片产地区别模型，实现了产地的区分，可用于分析同种药材不同产地的差异。Yang SL等[27]采用实验室自助研发的机器视觉系统和AlphaMos公司研发的电子鼻、电子舌，分别获取贝母样品的“L*a*b*”颜色数值、18维气味特征值与7维味道特征值；采用PCA法对多维数据进行降维处理，采用SVM算法将不同产地贝母作为输出变量，建立多层数学分析模型，用于判断贝母饮片的不同产地。该算法通过强化学习不断改变权值实现了对贝母“道地性”的鉴别，随着训练数据的积累，预测的结果值不断优化，最终该模型的识别率达到了92.6%。试验中并分别用电子鼻与电子舌数据构建PCA鉴别模型对上述结果进行验证，表明该模型的鉴别效果较好[27]。整个数据处理环节以Matlab R2012a软件进行操作，具体识别流程见图4。

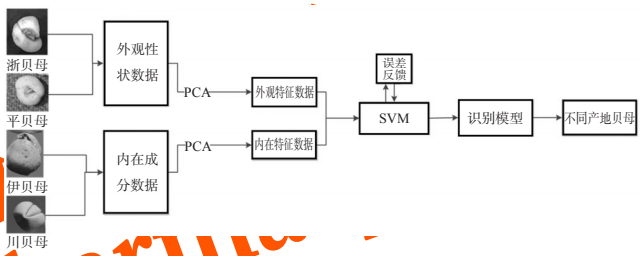

图4 不同产地贝母识别流程

此外，针对混合药材如不同品种、不同年份或辨识度低的药材饮片，也可通过训练海量数据与构建机器学习模型来提升数据分析与预测的准确性。

2.4.3 提高中药饮片质量控制标准 运用数据挖掘工具对中药饮片“内外结合”大数据库进行分析，建立算法预测模型来分析其各部分数据间的关系，有助于挖掘影响中药饮片质量的关键因素，进而有助于中药饮片质量标准的规范化，从而提高对中药饮片质量的监控与管理水平。

3 结语

随着大数据技术在多个领域的广泛应用，数据挖掘在中药中的应用也越来越多。建立中药饮片“内外结合”大数据库，可在整合海量数据的同时，结合机器学习算法挖掘数据潜力，并进行广泛应用。例如，其可用于中药真伪优劣品质的鉴别，实现人工智能识药；对中药产地进行辨别区分，分析同种药材不同产地的差异性，为道地药材的鉴定提供借鉴；挖掘影响中药饮片质量的因素，为其质量控制提供规范化标准，也为中药饮片质量监控与管理提供依据。大数据技术能有效整合多维、多变的数据，减少因信息单一而引致的错误判断，从而可为中药行业的现代化发展提供有力支撑。

中药饮片外观性状与内在成分数据是对其客观化评价的重要衡量标准，然而在其大数据技术研究过程中仍然存在一定问题，例如，数据来源的真实性与完整性难以得到保证，或不同中药饮片的外观性状数据未能全部获得；在饮片多方数据缺失严重的情况下，容易造成数据转换率折损、利用率较低等情况。随着当前人工智能技术的普及，引入深度学习方法对海量数据进行处理与分析，构建数学模型，有助于提高分析精准度及可靠性。应用大数据技术准确、高效地对中药饮片“内外”数据进行挖掘，可为传统中药行业研究提供新思路，为中药饮片客观化发展提供新动力。

参考文献

[1] 赵雷蕾，周洋，黎茂. 基于数据化表达的中药“形色气味”研究进展及思考[J]. 广东药学院学报，2015，31(5)：692-694.

[2] 黎江华，吴纯洁，孙灵根，等. 基于机器视觉技术实现中药性状“形色”客观化表达的展望[J]. 中成药，2011，33(10)：1781-1784.

[3] 吴继华，刘燕德，欧阳爱国. 基于机器视觉的种子品种实时检测系统研究[J]. 传感技术学报，2005，18(4)：742-744.

[4] 张俊雄，陈涛，于振东，等. 基于计算机视觉的新疆棉种颜色分选系统设计[J]. 农业机械学报，2009，40(10)：161-164.

[5] 常若葵，张伟玉，崔晶，等. 基于机器视觉的大米外特性评价[J].农机化研究，2009，31(12)：149-151.

[6] 夏于芬，梁光平. 大数据背景下的中药现代化[J].亚太传统医药，2012，11(21)：1-3.

[7] CUI M，LI HY，HU XQ. Similarities between “big data” and traditional Chinese medicine information[J]. *J Tradit Chin Med*，2014，34(4)：518-522.

[8] YEA SJ，SEONG B，JANG YJ，et al. A data mining approach to selecting herbs with similar efficacy：targeted selection methods based on medical subject headings[J]. *J Ethnopharmacol*，2016，8：27-34.

[9] 龙伟，邵馨，向剑，等. 中药方剂网络与中药化学空间的构建与分析[J]. 北京中医药大学学报，2011，34(11)：729-731.

[10] 陆爱军，刘冰，刘海波，等. 中药化学数据库关联规则的挖掘[J]. 计算机与应用化学，2005，22(2)：108-112.

[11] GUO J，SHANG E，ZHAO J，et al. Data mining and frequency analysis for licorice as a “Two-Face” herbin Chinese formulae based on Chinese formulae database[J]. *Phytomedicine*，2014，21(11)：1281-1286.

[12] 向杨峰. 基于数据挖掘的新药研发系统[D]. 北京：北京交通大学，2010.

[13] 付先军. 基于数据挖掘技术探讨治疗肺系疾病方剂中药物化学成分类别构成及其配伍关系[J]. 中医药信息学，2013，20(1)：28-30.

[14] 李振皓，钱忠直，程翼宇. 基于大数据科技的中药质量控制技术创新战略[J]. 中国中药杂志，2015，40(17)：3374-3378.

[15] 龚蓓，苏励，董亮，等. 基于大数据的风湿科常用中药饮片肾毒性初探[J]. 上海中医药杂志，2015，49(3)：7-9.

[16] 曹婷婷，孙志蓉，杨春宁，等. 硫黄熏蒸中药材的研究现状分析[J]. 中国现代中药，2016，18(5)：678-681.

[17] 李铎. 硫熏中药材快速检测装置设计研究[D]. 保定：河北大学，2016.

[18] 伍敏生. 硫熏对中药饮片质量的影响研究[J].中国中医药现代远程教育，2014，12(19)：158-159.

[19] 王洪建. 基于HSV颜色空间的一种车牌定位和分割方法[J]. 仪器仪表学报，2005，26(2)：371-373.

[20] XIE DS，PENG W，CHEN JC，et al. A novel method for the discrimination of hawthorn and its processed products using an intelligent sensory system and artificial neural networks[J]. *Food Sci Biotechnol*，2016，25(6)：1-6.

[21] 曾星翔. 通江银耳志[M]. 北京：方志出版社，2010：8-15.

[22] 施学丽，邓家刚，蒋筱，等. 195首治疗乳腺增生中药专利复方的用药规律分析[J]. 世界科学技术(中医药现代化)，2013，15(7)：1544-1550.

[23] YANG M，JIAO LJ，CHEN PQ，et al. Complex systems entropy network and its application in data mining for Chinese medicine tumor clinics[J]. *World Science Technology*，2012，14(2)：1376-1384.

[24] CHU H，SUN P，YIN J，et al. Integrated network analysis reveals potentially novel molecular mechanisms and therapeutic targets of refractory epilepsies[J]. *PloS One*，2017，12(4)：e0174964.

[25] TAN C，XIE D，LIU Y，et al. Identification of different bile species and fermentation times of bile arisaema based on an intelligent electronic nose and least squares support vector machine[J]. *Anal Chem*，2018，90(5)：3460-3466.

[26] 陶梦琳，顾文涛，侯珂惠，等. 基于支持向量机的“内外结合”中药质量控制新模式探索[J]. 中国药房，2016，27(1)：118-121.

[27] YANG SL，XIE SP，XU M，et al. A novel method for rapid discrimination of bulbus of Fritillaria by using electronic nose and electronic tongue technology[J]. *Anal Methods*，2015，7(3)：943-952.

（收稿日期：2017-11-05　修回日期：2018-07-05）

（编辑：孙　冰）