·数据挖掘研究·

淋巴瘤医案不同聚类分析方法比较研究

朱垚[1]，陆明[2,3]，杨涛[4]，倪海雯[5]

（1.南京中医药大学第一临床医学院，江苏 南京 210023；2.南京中医药大学针灸推拿学院·养生康复学院，江苏 南京 210023；3.南京医中数据挖掘中心，江苏 南京 210029；4.南京中医药大学人工智能与信息技术学院，江苏 南京 210023；5.南京中医药大学附属医院，江苏 南京 210029）

摘要：目的 以淋巴瘤临床医案为范例数据，对不同聚类分析方法挖掘结果进行比较，从而分析中医医案药物聚类挖掘方法的优化方案与结果差异。**方法** 对淋巴瘤医案进行统一预处理与规范，运用分散性聚类中的快速聚类、结构性聚类中的层次聚类进行挖掘分析，并从算法特点、终值偏倚与临床拟合 3 个维度综合比较。**结果** 研究共涉及患者 138 人次，病例 354 诊次，药物 451 味。分散性聚类中药物分散性聚类所得群集类 26 类，群集数最大 29，最小 1；方剂分散性聚类所得群集类 22 类，群集数最大 19，最小 3，位点值最大 14，最小 3。结构性聚类中 F10 药物凝聚层次聚类，群集数最大 25，最小 12；结构性聚类中 F20 药物凝聚层次聚类，群集数最大 21，最小 8；结构性聚类中 F30 药物凝聚层次聚类，群集数最大 15，最小 5。**结论** 对于中医临床医案单病种数据挖掘研究，方法的选取主要取决于样本的总体数量与药物的总体频数。数据量较小时宜选取结构性聚类，药物结构性聚类挖掘设计宜采用较高药物频幅，挖掘终值偏倚较低，研究结果临床拟合度较好；数据量较大时宜选取分散性聚类，分散性聚类挖掘设计宜采用方剂分散性聚类，挖掘终值偏倚较低，研究结果临床拟合度较好。

关键词：中医；医案；数据挖掘；分散性聚类；结构性聚类；方法学；差异比较

中图分类号：R273 **文献标志码：**A **文章编号：**1672-0482(2021)01-0105-08

DOI：10.14148/j.issn.1672－0482.2021.0105

引文格式：朱垚，陆明，杨涛，等.淋巴瘤医案不同聚类分析方法比较研究[J].南京中医药大学学报，2021，37(1)：105-112.

Comparative Study on Different Cluster Analysis Methods of Lymphoma Medical Records

ZHU Yao[1]，*LU Ming*[2,3]，*YANG Tao*[4]，*NI Hai-wen*[5]

(1. *The First School of Clinical Medicine, Nanjing University of Chinese Medicine, Nanjing*, 210023, *China*; 2. *School of Acupuncture and Tuina, School of Regimen and Rehabilitation, Nanjing University of Chinese Medicine, Nanjing*, 210023, *China*; 3. *Data Mining Center, Medchitec Co. Ltd., Nanjing*, 210029, *China*; 4. *School of Artificial Intelligence and Information Technology, Nanjing University of Chinese Medicine, Nanjing*, 210023, *China*; 5. *The Affiliated Hospital of Nanjing University of Chinese Medicine, Nanjing*, 210029, *China*)

ABSTRACT: OBJECTIVE To compare the mining results of different cluster analysis methods with lymphoma clinical medical records as sample data, and to analyze the optimization plan and result differences of the drug cluster mining methods in traditional Chinese medical records. **METHODS** Through performing unified preprocessing and standardization of lymphoma medical records, as well as using fast clustering in decentralized clustering, and hierarchical clustering in structural clustering for analysis and mining, such three dimensionalities as characteristics of algorithms, terminal value bias, and clinical fitting were analyzed and compared comprehensively. **RESULTS** This study involved 138 patients, 354 visits and 451 kinds of medicines. In the decentralized clustering, the drug decentralized clustering results had 26 clusters, while the largest number of clusters was 29 and the smallest was 1. The prescription decentralized clustering results included 22 clusters, the largest number of clusters was 19, the smallest was 3, the largest dot value was 14, and the smallest was 3. As for the F10 drug hierarchical clustering of structural clustering, the largest number of clusters was 25, and the smallest was 12. As for the F20 drug hierarchical clustering of structural clustering, the largest number of clusters was 21 and the smallest was 8. As for the F30 drug hierarchical clustering of structural clustering, the largest number of clusters was 15, and the smallest was 5. **CONCLUSION** For the research on data mining of traditional Chinese medicine (TCM) clinical medical records for a single disease, different clustering methods have been applied to study the drug combination or core prescriptions used in the clinical application of Chi-

收稿日期： 2020-07-29

基金项目： 江苏省中医药管理局课题(YB2017014)；国家中医药管理局第四批全国中医优秀人才研修项目(国中医药办人教函〔2017〕24 号)；南京中医药大学横向课题(2018045)；江苏省六大人才高峰项目(RJFW-40)；江苏省“333 高层次人才培养工程”；江苏省科技型企业技术创新基金(BC2015022)

第一作者： 朱垚，男，副教授，E-mail：zhongyiyaochuanren@126.com

通信作者： 倪海雯，女，主任医师，主要从事血液病的中医药治疗研究，E-mail：1499169641@qq.com

nese medicine. The selection of the methods mainly depends on the total number of samples and the overall frequency of drugs. When the amount of data is small, structural clustering should be selected while a higher drug frequency range should be used in the design of drug structural clustering mining, so as to get a lower final value of the mining bias and a better research result of the clinical fitting degree. When the amount of data is large, the decentralized clustering should be selected, and the prescription decentralized clustering in the design of decentralized clustering mining should be adopted to get a lower final value of the mining bias and better research result of the clinical fitting degree.

KEYWORDS: traditional Chinese medicine; medical records; data mining; decentralized clustering; structural clustering; methodology; difference comparison

基于中医临床医案的数据挖掘与知识发现是近年来中医药研究的热点之一。中医医案数据挖掘虽然研究方法众多，但符合中医临床辨治特点且能有效总结专家经验的挖掘方法，主要还是以频数解构、关联规则、聚类分析为主。聚类分析是将数据分到不同类的过程，同一个类中的数据有较大相似性，而不同类间的数据差异性较大。聚类分析是一种探索性的数据挖掘方法，在分类过程中，不必预先制定分类标准，聚类分析能够从样本数据出发自动分类。通过对中医医案实际数据样本集的系列研究发现，同一医案数据样本集所采用的聚类分析方法不同，结论亦有较大差异。不同研究者对于同一组数据进行相同的聚类分析方法，由于研究设计不同，所得到的聚类值也不尽相同。因此，基于中医药学知识发现的客观规律与前期大量中医医案数据挖掘实践[1-5]，本团队提出药物分散性聚类与方剂分散性聚类的概念，并根据聚类分析在中医医案研究中的主要目的，确立药物分散性聚类与方剂分散性聚类的内涵及外延，以期扩展和丰富聚类分析在中医医案数据挖掘领域的应用。本文采用淋巴瘤医案为范例数据，对不同聚类分析方法的挖掘结果进行比较，从而分析中医医案采用不同聚类分析方法的优化方案与结果差异，为中医药数据挖掘提供更加优化的研究范式。

1　资料与方法

1.1　数据采集

本研究采用数据挖掘中的单源数据库类研究，旨在研究单一来源数据库的医案数据结构特点，其优势在于能够有效降低建库噪点，减少研究复核能耗，留存单源数据研究样本，为多源数据库类研究形成比对数据，确保后续多源数据库类医案研究的差异值显现最大化。

选择目标单源数据库为中国知网（CNKI），进行“淋巴瘤”单病种、单源数据库类医案研究论文检索。论文数据库保留 Download Index 索引清单，进行双人复核，确保论文数据库与索引清单的量值一致。检索式：（SU＝"淋巴瘤"）AND（SU＝"国医大师" OR SU＝"名老中医"）AND（SU＝"验案" OR SU＝"经验" OR SU＝"治验" OR SU＝"医案"），检索时间设置为 1988 年 1 月 1 日－2019 年 8 月 15 日。

1.2　数据库构建

在论文数据库的基础上，采用 Medcase Ver3.8 诊籍中医师工作室-名老中医经验传承辅助平台，进行临床医案数据文本提取，建立淋巴瘤专病医案数据库。录入完成后进行二次数据审核；不同研究人员进行录入及审核，控制相异率＜3%。

1.3　纳入与排除标准

纳入标准：①论文医案描述中明确“淋巴瘤”诊断；②论文医案描述中存在中药内服干预方案；③论文医案描述中有复诊信息存在可供进行疗效判别的症状。

排除标准：经校验复核确认在不同论文中，记录了同一主诊医师相同的医案，排除时序首位记录外的所有医案。

1.4　诊断与中医证候分型标准

中医辨证分型及证候要素参照《现代中医肿瘤学》[6]《中药新药临床研究指导原则》[7]；西医诊断标准参照《WHO（2008）造血与淋巴组织肿瘤分类诊断标准》[8]根据受累淋巴结或结外肿瘤组织病理报告确定诊断。

1.5　数据预处理

针对淋巴瘤专病医案数据库中文本医案，在医案录入与数据提取过程中对明显的症状、病机、治法、药物、理化检查等数据源中的错误等非研究性数据噪点，进行溯源性预处理，达到降噪、优化的目的。溯源数据值与修正数据值由不同研究人员实时双备份标记。

1.6　数据规范化

预处理后的医案数据库按照研究分析类型的不同，进行数据规范。规范化中医药术语分项集进行，症状项集、诊断项集、病机项集规范参考《中医诊断

学》[9]；药物项集规范根据临床经验导向型知识获取原则，遵照中医医案真实世界研究范式，药物名称参考《中药学》[10]，其他类型数据规范化采用《中医临床医案数据挖掘研究数据规范化标准》[11]。对于明确为非标准简写或非标准全称药物，均按照中药标准名称进行规范；对于中医临床疗效有差异或专家使用强调道地药材功效的药物，则保留原医案药名规则，以促进较多临床型知识规则的获取；对于真实医案中未能明确炮制方法或生、熟特性的药物，保留原始医案药名，由研究者根据最终获取的知识规则，对药物的炮制方法与生、熟特性进行临床读判与学术研究。非标准全称完善为标准全称，如生薏仁、生苡仁统一规范为生薏苡仁，熟薏仁、熟苡仁统一规范为熟薏苡仁，但生薏苡仁与熟薏苡仁从临床使用角度看，疗效有差别，因此按照真实世界研究范式，遵从临床专家的使用习惯，以便多维度获取临床知识规则。在此次聚类研究中，不再进行合并性药物名称规范，如附子原始医案中标明生附子与熟附子的采用生、熟区分，未注明生熟均按照医案原文保留为附子，作为不同研究项集素材存在，以获得更多层次知识规则，供研究者进行临床拟合分析。对于多种不同炮制方法产生不同疗效特点的药物，不进行合并性药物名称规范，如清半夏、法半夏、矾半夏、姜半夏、竹沥半夏、制半夏（原文未标明，不做硬性划分）等不做统一性规范。对于同一大类药物，子类内涵临床选用存在差异的，均予以保留，不进行合并性药物名称规范，如金钱白花蛇主指小白花蛇，白花蛇作为大类名称包含大白花蛇。临床医家常用小方或成药，按照临床使用习惯，作为独立项集素材存在，不做进一步药物细化分拆及药名规范化处理，如黛蛤散、猴枣散等。

1.7　数据分析

Medcase V3.8 诊籍中医师工作室-名老中医经验传承辅助平台系江苏省科技创新专项研究成果，由南京中医药大学国医大师周仲瑛工作室、第二临床医学院、人工智能与信息技术学院联合研发[12]，已在江苏地区中医临床、教学、科研单位广泛使用。研究采用 Medcase 系统中的 XMiner V1.0 中医药数据挖掘平台进行格式化和编码，并根据文本特征计算数据权重，参考《中医临床医案数据挖掘研究数据分析操作标准》[13]操作执行系统常规极值处理，实时记录调参标值，基于 Pycharm 的 Kmeans 工具与 Hierarchy 工具进行运算数据的可视化表达。

1.8　方法学差异性研究设计

对于淋巴瘤临床医案的范例数据，在先期统一预处理与规范后，运用分散性聚类中的快速聚类、结构性聚类中的层次聚类进行分析挖掘，并在研究中设计不同的挖掘路径。在分散性聚类中根据方剂药物重频规则，采用药物分散性聚类与方剂分散性聚类并行挖掘；在结构性聚类中根据药物频幅的节段，设置 F10、F20、F30 3 个药物凝聚层次进行并行挖掘。将最终的挖掘结果根据算法特点、终值偏倚与临床拟合 3 个方面进行综合分析比较，从而得出中医医案药物聚类挖掘方法的优化方案与结果差异。

此次比较研究中采用的聚类分析方法，大类选取的是以 K-Means 聚类为代表的分散性聚类和以系统聚类为代表的结构性聚类。分散性聚类亚类选取笔者所在团队提出的药物聚类分散性与方剂分散性聚类。结构性聚类亚类选取则是定向性药物凝聚层次聚类，根据不同的药物频幅，设定高、中、低 3 段频幅的药物凝聚层次聚类。

药物分散性聚类是分散性聚类中的一种结合中医临床用药特点优化的聚类亚型。其核心内涵是数据降维，在中医临床医案的药物数据挖掘中，仅计算药物在所有诊次中全部药物的绝对值及药物间的绝对距离。药物分散性聚类研究的具体实施操作是将药物项集纵向矩阵化，根据每个药物在矩阵中的共现特征，赋予相应权值，根据可视化需要进行降维处理，采用药物唯一性定位，根据欧氏距离进行药物聚类分析。

方剂分散性聚类是分散性聚类中的一种结合中医临床组方特点优化的聚类亚型。其核心内涵是数据升维，在中医临床医案的药物数据挖掘中，仅计算共现药物在全部诊次中的相对值及共现药物的位点数值。方剂分散性聚类研究是将药物项集横向矩阵化，根据不同诊次共现药物在矩阵中的特征，赋予相应权值，按照可视化需要进行升维处理，采用共现药物的位点频次，根据欧氏距离进行共现药物诊次特征聚类分析。

凝聚层次聚类是结构性聚类的常见聚类亚型，但在中医药领域中的运用，缺乏统一划分凝聚层的优化方案。笔者所在团队根据前期中医医案挖掘实践，在中医临床医案的药物数据挖掘中，采用药物频幅权重划分聚类的凝聚层，常规划分标准为所有诊次全部药物中出现频率每 10 次为一个频幅节点。药物凝聚层次聚类研究的具体实施操作是先根据药

物频幅的权重进行分层，将全部目标药物进行频幅分布挖掘，再根据分布结果设定频幅分层节点，确定高、中、低 3 段频幅节点后，采用结构性聚类中的定向性药物凝聚层次聚类分析。

本次淋巴瘤研究中的药物凝聚层次聚类中，F10 频幅即经过权重分层，设定频幅分层节点为全部药物频次中大于 10 次的药物凝聚层，F20 频幅即经过权重分层，设定频幅分层节点为全部药物频次中大于 20 次的药物凝聚层，F30 的频幅即经过权重分层，设定频幅分层节点为全部药物频次中大于 30 次的药物凝聚层，然后针对 3 个频段凝聚层进行结构性聚类，获得终值。

2　结果

2.1　研究总体描述

本次研究符合纳入标准的医案 138 则，共计 138 人次，354 诊次，其中男性 176 诊次，占总诊次数的 49.72%；女性总共 178 诊次，占总诊次数的 50.28%。年龄最大患者 86 岁，最小者 6 岁。研究涉及病机 59 条，症状 215 种，脉象 18 种，舌象 80 种，药物 451 种。

2.2　医案疗效评估

本次医案 138 则，其中仅记录初诊的单诊次医案 54 则，记录复诊的多诊次医案 84 则，复诊症状改善阳性医案 81 则，复诊症状改善率为 96.43%；诊疗超过 5 诊次的长诊次医案为 17 则，诊疗时间超过 6 个月的长疗程医案 18 则，全部长诊次医案及长疗程医案复诊临床症状改善率均为 100%。

2.3　药物分散性聚类

结果见表 1。

表 1　药物分散性聚类群集值

群集类	群集数	群集值
1	23	人参、凤凰衣、炒白术、刺猬皮、清半夏、厚朴、橘红、猫爪草、生姜、生白术、白苏子、知母、石见穿、紫草、红藤、茯苓、莱菔子、葶苈子、败酱草、金沸草、陈皮、鸡内金、龙胆草
2	16	佛手、平地木、当归尾、木鳖、板蓝根、熟庄黄、牵牛子、白芷、白鲜皮、芍药、天花粉、藤梨根、钩藤、雄黄、黄柏、龟板
3	28	乌梅、乌药、土茯苓、天竺黄、天麻、枣皮、桂枝、桑螵蛸、橘络、栀子、海螵蛸、炒白芍药、炮姜、玉竹、百合、益智仁、竹茹、紫菀、茜草炭、蒲黄、覆盆子、诃子、铅丹、银柴胡、阿胶珠、骨碎补、龙骨、龟板胶
4	21	代赭石、八月札、制首乌、鳖甲、南沙参、地榆、墨旱莲、枳实、枸橘李、泽漆、漏芦、白残花、白英、白薇、红景天、白花蛇舌草、蜂房、马勃、鸡血藤、龙葵、龙葵子
5	10	天花粉、天葵子、柴胡、炒枳壳、白蒺藜、蒲公英、连翘、金银花、香附、麻黄
6	29	仙茅、全蝎、冬凌草、墓头回、夜交藤、黑大豆、威灵仙、小蓟、山豆根、忍冬藤、杏仁、桑叶、炒冬瓜子、焦山楂、王不留行、白及、石打穿、秦艽、糯稻根、紫苏子、芙蓉叶、芡实、紫苏叶、茯苓皮、茵陈、莲子、谷芽、黄芪
7	13	制半夏、合欢皮、姜黄、川芎、桂枝、桔梗、海藻、生甘草、穿山甲、红花、胆南星、金钱草、青皮
8	24	牡丹皮、北沙参、大青叶、荞麦、山楂炭、川厚朴、川贝母、开金锁、桑叶、桑寄生、沙参、炒谷芽、赤芍药、黄柏、猴枣散、白扁豆、白蔹、石膏、紫花地丁、苍耳草、薄荷、蛇床子、铁皮枫斗、鸭跖草
9	16	僵蚕、前胡、壁虎、旱莲草、杜仲、枇杷叶、枳壳、油松节、水牛角、益母草、紫苏叶、红豆杉、苏木、葛根、金刚骨、青风藤
10	1	黄芪
11	21	地骨皮、天南星、射干、山慈菇、桑白皮、椿根皮、水牛角片、鹿茸草、泽兰、淫羊藿、牡丹皮、瓦楞子、紫荆皮、肉苁蓉、菊花、七叶一枝花、防风、马齿苋、鬼箭羽、鸡矢藤、黄芩
12	1	太子参
13	11	三叶青、土贝母、川楝子、昆布、枸橘核、炙甘草、生薏苡仁、红枣、青黛、生黄芪、黄药子
14	11	乳香、五灵脂、制乌头、地龙、山慈姑、当归、木鳖子、枫香脂、没药、香墨、麝香
15	23	八月扎、地肤子、夏枯草、大腹皮、山药、干姜、泽泻、海浮石、海蛤壳、牛黄、牡蛎、猪苓、玄参、玉米须、白扁豆衣、白鲜皮、糙米、肉桂、车前子、车前草、郁金、金樱子、麦芽
16	9	三七、何首乌、天门冬、牛蒡子、白茅根、紫苏梗、胡麻仁、芦根、西洋参
17	16	大血参、小血参、山萸肉、核桃、白附子、白首乌、石菖蒲、紫油桂、红参、葱白、藿香、蜂蜜、赤芍药、附子、青蒿、黑小豆
18	17	冬凌草、地鳖、土鳖虫、生山楂、徐长卿、桃仁、海藻、溪黄草、灵芝、生甘草、石斛、神曲、肿节风、白芥子、蛇蜕、阿胶、鹿角胶、麦冬
19	16	丹参、仙鹤草、半枝莲、半边莲、卷柏、女贞子、枸杞子、炒苍术、炙甘草、生地黄、皂角刺、石韦、羊蹄根、花生衣、苦参、菟丝子

（续表）

群集类	群集数	群集值
20	20	三棱、乌头、伸筋草、大枣、山海螺、木瓜、木通、水蛭、浮小麦、滑石、独活、甘草梢、细辛、羌活、苍术、莪术、肉豆蔻、通草、雷公藤、马钱子
21	17	五味子、佩兰、制山甲、大黄、白扁豆、栀子、桑枝、火麻仁、瓜蒌子、白芍药、石决明、紫河车、茯神、虎杖、金荞麦、首乌藤、鱼腥草
22	22	乌梢蛇、穿山甲、槟榔、橘核、沉香、炒杏仁、炒酸枣仁、瓜蒌、白花蛇、百部、绿豆、羚羊角、荔枝核、葶苈、薤白、路路通、金钱白花蛇、八月札、香茶菜、香菌、鳖甲、鹿衔草
23	40	七叶参、党参、合欢花、垂盆草、大蒜、天葵、小茴香、小麦、生川续断、延胡索、灯芯草、桑椹子、檀香、款冬花、沉香曲、淡豆豉、炙牛角腮、瓜蒌实、炮姜炭、炮甲珠、焦山栀、狗脊、玫瑰、珍珠母、瓜蒌子、白花蛇舌草、石莲子、竹沥、紫石英、炒川续断、艾叶、紫苏梗、茵陈蒿、莲须、菖蒲、蟾皮、贝母、远志、青葙子、饴糖
24	17	炙黄芪、升麻、木馒头、木香、炒扁豆、炒麦芽、砂仁、竹叶、莲子心、葎草、薏苡仁、蛇莓、蜈蚣、蝉蜕、补骨脂、酸枣仁、黄连
25	22	七叶一枝花、制大黄、天冬、巴戟天、柏子仁、天花粉、浙贝母、炒稻芽、炙龟板、熟地黄、牛膝、瓜蒌子、瓜蒌皮、紫菀、绞股蓝、荆芥、蛇六谷、金银花、鬼针草、鹿角、黄精、黛蛤散
26	5	槟榔炭、水红花子、芜荑、荷叶、贯众

注：此群集标列参数 Mark Parameter=[K=20.0000；inertia=0.0349]；Format Export by Medcase Chart © 2020。

本次药物分散性聚类共计挖掘获得群集类 26 项，其中群集数＜10 区间的群集类有 4 项，群集数在 10～20 区间的有 12 项，群集数＞20 区间的有 10 项。药物分散性聚类所得药物组合的药味数量普遍偏大，不完全符合中医临床组方原理；虽然此类方法目标药物没有重复性，但挖掘所得药物组合中部分存在个别药物的临床低解释性特征；个别群集仅为单味药物，虽有可能为专病单方，但从临床实际出发可能性较低，应配合其他药物组合使用。

2.4　方剂分散性聚类

结果见表 2。

表 2　方剂分散性聚类群集值

群集类	位点数	群集数	群集值
1	14	5	清半夏、浙贝母、猫爪草、陈皮、黄芪
2	11	13	人参、清半夏、大枣、女贞子、山药、枸杞子、甘草、生姜、白花蛇舌草、茯苓、菟丝子、陈皮、黄芪
3	11	3	半枝莲、白花蛇舌草、蒲公英
4	10	3	甘草、红豆杉、茯苓
5	8	10	仙鹤草、鳖甲、北沙参、半枝莲、太子参、女贞子、漏芦、肿节风、鸡血藤、麦冬
6	8	6	制半夏、太子参、柴胡、甘草、生白术、蒲公英
7	7	8	乳香、五灵脂、地龙、木鳖子、枫香脂、没药、香墨、麝香
8	6	19	僵蚕、夏枯草、姜黄、川芎、当归、柴胡、桔梗、浙贝母、海藻、猫爪草、生甘草、穿山甲、红花、连翘、金银花、青皮、香附、黄芪、黄药子
9	6	16	丹参、仙鹤草、党参、半枝莲、卷柏、女贞子、枸杞子、甘草、生地黄、白花蛇舌草、石韦、羊蹄、花生衣、苦参、菟丝子、黄芪
10	6	12	三七、三棱、人参、党参、北沙参、太子参、山萸肉、昆布、水蛭、海藻、生地黄、莪术
11	6	8	党参、炒白术、炙黄芪、当归、木香、茯苓、补骨脂、酸枣仁
12	6	8	山药、泽泻、猪苓、甘草、石见穿、茯苓、车前子、车前草
13	6	3	夏枯草、山慈菇、莪术
14	5	23	三七、党参、前胡、北沙参、川芎、当归、枇杷叶、柴胡、桔梗、泽泻、浙贝母、猪苓、全瓜蒌、甘草、石膏、紫苏叶、菊花、葛根、赤芍药、车前子、连翘、陈皮、黄芩
15	5	15	党参、凤凰衣、刺猬皮、厚朴、生白术、白芍药、穿山甲、红藤、茯苓、蜈蚣、败酱草、金荞麦根、陈皮、鱼腥草、鸡内金
16	5	5	大枣、干姜、炙甘草、生姜、黑小豆
17	5	3	半枝莲、白花蛇舌草、百合
18	4	18	北沙参、半夏、壁虎、大枣、女贞子、枸杞子、玄参、甘草、生地黄、生姜、生白术、茯苓、菟丝子、金钱白花蛇、陈皮、麦冬、麦芽、黄芪
19	4	16	僵蚕、冬凌草、土鳖虫、地龙、壁虎、徐长卿、浙贝母、海藻、玄参、生地黄、生甘草、石斛、肿节风、白芥子、金刚骨、麦冬
20	4	9	射干、水牛角、牡丹皮、甘草、七叶一枝花、金荞麦根、马齿苋、鸡矢藤、黄芩
21	3	15	僵蚕、凤凰衣、炮山甲、半枝莲、壁虎、猫爪草、生白术、白花蛇舌草、茯苓、莪术、金刚骨、青风藤、鸡内金、黄芪、龙葵
22	3	5	丹参、夏枯草、浙贝母、牡蛎、玄参

注：此群集标列参数 Mark Parameter=[K=22.0000；inertia=1737.6799]；Format Export by Medcase Chart © 2020。

本次方剂分散性聚类共计挖掘获得群集类 22 项，其中群集数＜10 区间的群集类有 12 项，群集数

在 10～20 区间的有 9 项，群集数＞20 区间的有 1 项。方剂分散性聚类研究结果所得药物组合的药味数量普遍偏小，相对符合中医临床组方原理；虽然此类方法目标药物有重复性，但挖掘所得药物组合具有临床高解释性特征；未出现单味药物，最低群集为 3，符合临床角药小方特征，更加符合中医临床实际。

2.5　F10 药物结构性聚类

结果见表 3。

表 3　F10 药物凝聚层次聚类群集值

群集类	位点值	群集数	群集值
1	0.280 9	25	附子、炙甘草、肉桂、干姜、黄芩、大枣、人参、生姜、炒麦芽、天冬、败酱草、厚朴、鱼腥草、金荞麦根、白鲜皮、地肤子、土茯苓、猪苓、茯苓、泽泻、桂枝、黄精、熟地黄、蛇六谷、苦参
2	0.287 2	12	淫羊藿、炮山甲、麦芽、枸杞子、女贞子、菟丝子、白花蛇舌草、半枝莲、神曲、山楂、阿胶、灵芝
3	0.270 5	20	紫草、知母、牡丹皮、石见穿、山萸肉、赤芍药、鸡内金、猫爪草、山药、浙贝母、夏枯草、莪术、三棱、陈皮、炒白术、半夏、黄芪、茯苓、生白术、党参
4	0.292 9	16	玄参、牡蛎、白芍药、甘草、海藻、昆布、红枣、金银花、连翘、炒枳壳、天葵子、天花粉、蒲公英、芍药、香附、柴胡

注：此群集标列参数 Mark、Parameter＝[Frequency amplitude＞10；Pick points＜30]；Format Export by Medcase Chart © 2020。

在药物频幅大于 10 的 F10 药物凝聚层次聚类分析结果中，满足标列参数 Mark Parameter 符合 Frequency amplitude＞10 且 Pick points＜30 的条件下，共计挖掘获得群集类 4 项，其中群集数在10～20 区间的有 2 项，群集数≥20 区间的有 2 项。F10 药物凝聚层次性聚类所得药物组合的药味数量普遍偏大，不完全符合中医临床组方原理；挖掘所得药物组合中存在部分药物临床解释性低。

2.6　F20 药物结构性聚类

结果见表 4。

表 4　F20 药物凝聚层次聚类群集值

群集类	位点值	群集数	群集值
1	0.217 3	8	桃仁、昆布、海藻、金银花、连翘、玄参、牡蛎、夏枯草
2	0.226 2	9	香附、柴胡、桔梗、黄芩、天花粉、蒲公英、当归、川芎、白芍药
3	0.287 2	21	枸杞子、女贞子、菟丝子、甘草、黄芪、生姜、大枣、人参、半夏、麦芽、茯苓、生白术、党参、酸枣仁、淫羊藿、补骨脂、砂仁、木香、陈皮、炒白术、山药
4	0.232 2	10	泽泻、桂枝、赤芍药、枳壳、附子、炙甘草、白芥子、肉桂、黄精、山慈菇

注：此群集标列参数 Mark Parameter＝[Frequency amplitude＞20；Pick points＜25]；Format Export by Medcase Chart © 2020。

在药物频幅大于 20 的 F20 药物凝聚层次聚类分析结果中，满足标列参数 Mark Parameter 符合 Frequency amplitude＞20 且 Pick points＜25 的条件下，共计挖掘获得群集类 4 项，其中群集数＜10 区间的有 2 项，群集数在 10～20 区间的有 1 项，群集数＞20 区间的有 1 项。F20 药物凝聚层次性聚类所得药物组合的药味数量较 F10 的域宽等级有所缩减，相对符合中医临床组方原理；挖掘所得药物组合中存在少量药物的临床解释性低。

2.7　F30 药物结构性聚类

结果见表 5。

表 5　F30 药物凝聚层次聚类群集值

群集类	位点值	群集数	群集值
1	0.358 2	14	白花蛇舌草、半枝莲、炙甘草、麦冬、太子参、砂仁、山萸肉、生甘草、生地黄、熟地黄、薏苡仁、仙鹤草、玄参、牡蛎
2	0.202 9	5	蒲公英、制半夏、金银花、连翘、海藻
3	0.342 7	15	党参、丹参、当归、川芎、桔梗、莪术、猫爪草、山慈菇、浙贝母、夏枯草、穿山甲、僵蚕、香附、柴胡、黄芩
4	0.285 2	11	麦芽、神曲、赤芍药、白芍药、枳壳、黄芪、生白术、鸡内金、茯苓、甘草、女贞子
5	0.244 0	9	菟丝子、枸杞子、生姜、大枣、人参、陈皮、炒白术、清半夏、山药

注：此群集标列参数 Mark Parameter＝[Frequency amplitude＞30；Pick points ＝Total]；Format Export by Medcase Chart © 2020。

在药物频幅大于 30 的 F30 药物凝聚层次聚类分析结果中，满足标列参数 Mark Parameter 符合 Frequency amplitude＞30 且 Pick points＝Total 的条件下，共计挖掘获得群集类 5 项，其中群集数＜10 区间的有 2 项，群集数在 10～20 区间的有 3 项。F30 药物凝聚层次性聚类所得药物组合的药味数量较 F20 的域宽等级进一步缩减，更加符合中医临床组方原理；挖掘所得药物组合大部分具有临床高解释性特征。

3　讨论

3.1　算法特点比较

分散性聚类的算法原理是首先选择聚类的类

数，其次产生任意类数个聚类，确定聚类中心，再对每个点确定其聚类中心点，计算其聚类新中心，重复多次，最终确定中心点不再改变。分散性聚类的优势在于解决聚类问题简单快捷；算法对大数据集处理可保持伸缩性和高效率；当群集值较密集时效果较好。劣势在于类数的平均值可被定义的情况下才能使用，可能不适用于某些应用；必须事先给出类数，在运算过程中对初值敏感，不同初值可能会导致终值差异；不适合于发现非凸形状的类或者大小差别很大的类；对噪声和孤立数据较为敏感。结合此次淋巴瘤医案数据挖掘结果，在医案数量为354诊次且目标药物数量为451种时，药物数量相对阈值不大，采用分散性聚类运算的效率优势没有完全显现，且运算结果群集数偏大，群集值较多，临床解释性相对较低。

本次淋巴瘤研究中的结构性聚类采用层次聚类中自下而上的凝聚层次聚类，其主要算法原理是首先将每个对象作为一个类，然后运算合并这些子类为越来越大的类，直到满足终结条件而停止运算。实际聚类分析中，结构性聚类是使用最多的一种聚类方法，其优势在于结构性聚类既可以对样本聚类，也可以对变量聚类，变量可以是连续性变量也可以是分类变量；结构性聚类的类间距离计算方法和结果表示方法非常丰富，可视化效果较好。结构性聚类劣势在于与其分析过程相关，由于每一步聚类都需要计算类间距离，当变量较多或样本数据量较大时，运算速度较慢，运算效率较低。结合此次淋巴瘤医案数据挖掘，采用结构性聚类中的凝聚层次聚类运算效率与分散性聚类基本相当，而运算结果群集数相对偏小，群集值较合理，临床解释性相对较高。

因此，样本数量的大小在一定程度上决定了数据挖掘中分散性聚类与结构性聚类的选取，样本数量较大时分散性聚类运算较为高效，样本数量较小时结构性聚类适应性较好，可视程度更佳，在中医药领域运用时更加符合临床实际，具有较好的临床解释性。

3.2　终值偏倚比较

分散性聚类中药物分散性聚类的设计特点是按照所有样本医案数据中药物出现单次计算药物距离，所得终值为药物间绝对距离。分散性聚类中方剂分散性聚类的设计特点是根据所有样本医案数据中相同药物组合共现数计算药物距离，所得终值为药物间相对距离。从挖掘终值结果来看，在目标药物绝对数量相对较低时，按照临床实际的方剂组方原则与方剂常规药味数为标准，药物分散性聚类产生的极值偏倚较大，方剂分散性聚类产生的极值偏倚较小，并能明确展示药物组合实际出现的位点数与位点值，更加利于临床分析与数据溯源。

在结构性聚类中，根据药物频幅的节段设置F10、F20、F30 3个药物凝聚层次进行并行挖掘。F10药物凝聚层次聚类为药物频幅大于10的药物系统聚类；F20药物凝聚层次聚类为药物频幅大于20的药物系统聚类；F30药物凝聚层次聚类为药物频幅大于30的药物系统聚类。从挖掘终值结果来看，在药物绝对数量相对较低时，药物频幅越高的凝聚层次聚类群集数区间相对更加集中，相对于临床实际方剂药味数均值，产生的极值偏倚也相对较小。

由此可见在样本药物绝对数量相对较低时，结构性聚类的群集数较分散性聚类的群集数相对较小，极值偏倚也较小，更加接近临床实际方剂药味数均值，而方剂分散性聚类与高频幅药物凝聚层次聚类在终值表达与可视化展示方面更具优势。

3.3　临床拟合比较

本次淋巴瘤医案样本分散性聚类中，药物分散性聚类群集数大于20的有10项，群集数等于1的有2项，符合临床处方组方规律的群集数仅为14项。从中医临床诊疗实际出发，群集数大于20的多为大方，研究偏倚风险较大，临床解释困难，群集数为1的多为单药、单方，不符合聚类分析的运用目的。而剩余的14项也可能存在数据噪声与临床意义不可解释性的问题。分散性聚类中方剂分散性聚类因为算法及设计更加符合中医临床方剂的使用特点，因此干扰噪点较低，虽然群集类也有22类，但整体群集值较药物分散性聚类群集值大幅下降。其中群集值大于10的共有10项，而剩余12项均为个位群集值，更加符合中医临床核心药物组合或经典方剂的解释，功效主治特征更加突出，有较强的临床解释性。由此可见，在分散性聚类方法下，方剂分散性聚类较药物分散性聚类具有更好的临床拟合度，而群集类的数量对临床拟合度影响较小，群集值对临床拟合度影响较大，群集值在(10±5)范围内为临床拟合的最佳阈值，具有较高的临床可解释性。

本次淋巴瘤医案结构性聚类中，从临床拟合度来看，挖掘结果序位应为F30＞F20＞F10，低频幅药物凝聚层次聚类的群集值较为离散，临床意义不可解释性较高，临床拟合度较低；而高频幅药物凝聚

层次聚类群集值较为集中，频幅越高群类别越集中，临床解释性越好，临床拟合度越高。由此可见，在结构性聚类方法下，根据药物频幅优先选取高频药物进行凝聚层次聚类具有更好的临床拟合度，而聚类分析结果位点值对临床拟合度影响较小，群集数随着药物频幅的上升存在优化可能，群集值随着药物频幅的上升反而成下降趋势，更加符合中医临床特征；从结构性聚类最终群集值 10±3 阈值范围来看，结构性聚类临床拟合的最佳阈值范围也较分散性聚类群集值更为聚合，临床可解释性进一步提升。

4　结论

综上所述，结合淋巴瘤医案数据挖掘结果进行比对研究，在淋巴瘤医案数量为 354 诊次且目标药物数量为 451 种时，分散性聚类分析与结构性聚类分析在知识规则的结果表达上各有特点，但从算法特点、终值偏倚、临床拟合三个维度整体评价，结构性聚类在此研究中更具有优势。而此次基于淋巴瘤医案的数据挖掘方法比较研究中，结构性聚类的亚类中药物频幅大于 30 的 F30 药物凝聚层次聚类分析是符合中医临床数据挖掘研究范式的最优方案。

对于中医临床医案单病种数据挖掘研究，采用不同的聚类方法研究临床中药运用的药物组合或核心处方，方法的选取主要取决于样本总体数量与药物总体频数。数据量较小时宜选取结构性聚类，药物结构性聚类挖掘设计宜采用较高药物频幅，挖掘终值偏倚较低，研究结果临床拟合度较好；数据量较大时宜选取分散性聚类，分散性聚类挖掘设计宜采用方剂分散性聚类，挖掘终值偏倚较低，研究结果临床拟合度较好。但无论中医医案类数据挖掘选取何种聚类方法，在对挖掘终值的分析上均需要研究者有较为深厚的临床经验，才能更全面地根据挖掘结果获取新的知识。

参考文献：

[1] 苏克雷，叶娟，张业清，等.基于数据挖掘的江浙沪名老中医膏方医案关联解析[J].中华中医药杂志，2019，34(6)：2721-2727.

[2] 朱青，朱垚，陆明.基于国医大师周仲瑛临证肝胆医案的经验解构研究[J].中华中医药杂志，2017，32(4)：1814-1817.

[3] 黄磊，朱垚，陆明，等.周仲瑛临证医案参附药对经验解构[J].中国中医基础医学杂志，2016，22(6)：863-865.

[4] 夏娟，朱垚，陆明.基于国医大师周仲瑛临证医案的交泰丸运用经验解构[J].江苏中医药，2016，48(5)：14-16，18.

[5] 厉励，朱垚，陆明.近现代内分泌代谢性疾病“瘀热”医案解构研究[J].中国临床研究，2016，29(2)：253-256，259.

[6] 周晓鸽.WHO(2008)造血与淋巴组织肿瘤分类[J].诊断病理学杂志，2008，15(6)：510-512.

[7] 王永炎，严世芸.实用中医内科学[M].上海：上海科学技术出版社，2009：702-706.

[8] 中药新药临床研究指导原则[M].北京：中国医药科技出版社，2002：383-388.

[9] 吴承玉，王天芳.中医诊断学[M].上海：科学技术出版社，2018.

[10] 唐德才，高学敏，吴庆光，等.中药学[M].北京：人民卫生出版社，2016.

[11] 中医临床医案数据挖掘研究数据规范化标准[S].南京：江苏地区备案企标，2019.

[12] 杨涛，陆明，朱垚.基于 FP-Growth 的中医药数据关联分析平台的设计和应用[J].时珍国医国药，2016，27(12)：3050-3052.

[13] 中医临床医案数据挖掘研究数据分析操作标准[S].南京：江苏地区备案企标，2019.

（编辑：叶亮）