

• 数据挖掘与循证医学 •

常用关联与聚类分析方法对中医处方数据的适用性探讨

张 迪¹, 陈艺幻², 张伟娜^{1*}

1. 中国中医科学院 中医药信息研究所, 北京 100700

2. 中国中医科学院 中医临床基础医学研究所, 北京 100700

摘 要: **目的** 探究常用关联与聚类分析方法对中医处方数据的适用性, 基于现有方法提出一种在中医处方数据挖掘研究中关联与聚类分析部分的良好实践方案。**方法** 结合中医处方数据特点, 从算法原理层面指出各分析方法与中医处方数据的不匹配之处; 以列举典型真实研究示例的形式展示不同方法在应用时可能出现的不合理现象; 对具备优化价值的方法提出其应用时的优化方法; 对于关联分析方法, 基于真实医案数据横向对比所有算法的计算结果以呈现算法间的差异。**结果** 除 Pearson、Spearman 和 Phi 相关系数外, 所有药物关联分析方法应用于中医处方数据时均存在不同程度的局限性, 在对比研究中, 支持度与置信度对高频药物的计算结果均偏高, 欧氏距离与汉明距离易高估低频药物间的关联性, 余弦相似度与 Jaccard 系数对频次差异较大的药物组合的计算结果偏低, 互信息与改良互信息对全部组合的计算结果几乎均存在异常, 提升度对高频药物的计算结果偏低; 在常用的药物聚类方法中, 系统聚类与极大团算法对药物的划分效果最佳, 但前者易在类中纳入低相关药物, 后者的类别划分结果则过于模糊, 经优化后, 两者的局限性均得到改善。**结论** 在涉及药物关联计算时, 应首选 Pearson、Spearman 和 Phi 相关系数; 在进行药物聚类分析时, 可借鉴优化后的应用方法, 利用系统聚类分析药物类别的凝聚过程, 利用极大团算法探索药物的多维应用方向。

关键词: 中医; 处方; 数据挖掘; 关联分析; Pearson 相关系数; Spearman 相关系数; Phi 相关系数; 系统聚类分析; 极大团算法
中图分类号: TP18; R28 **文献标志码:** A **文章编号:** 0253-2670(2025)11-3974-11

DOI: 10.7501/j.issn.0253-2670.2025.11.018

Discussion on applicability of common association and cluster analysis methods to traditional Chinese medicine prescription data

ZHANG Di¹, CHEN Yihuan², ZHANG Weina¹

1. Institute of Information on Traditional Chinese Medicine, China Academy of Chinese Medical Sciences, Beijing 100700, China

2. Institute of Basic Research in Clinical Medicine, China Academy of Chinese Medical Sciences, Beijing 100700, China

Abstract: Objective To investigate the applicability of common association and clustering analysis methods to traditional Chinese medicine (TCM) prescription data, and propose an optimal framework for association and clustering analysis in TCM prescription data mining research based on existing methods. **Methods** By considering the characteristics of TCM prescription data, this study identifies mismatches between analytical methods and TCM prescription data at the algorithmic principle level. Typical real-world research examples are presented to demonstrate potential unreasonable phenomena in the application of different methods. For methods with optimization potential, improved application approaches are proposed. For association analysis methods, a comparative evaluation of all algorithms is conducted using real medical case data to highlight differences in computational outcomes. **Results** Besides Pearson, Spearman, and Phi correlation coefficients, all drug association analysis methods exhibit varying degrees of limitations when applied to TCM prescription data. Comparative experiments revealed the following issues: Support and confidence metrics overestimate results for high-frequency drugs; Euclidean and Hamming distances tend to overestimate associations between low-frequency drugs; cosine similarity and Jaccard coefficient yield underestimated results for drug combinations with significant frequency disparities; mutual information and modified mutual information produce abnormal results for nearly all combinations; and lift underestimates results for high-frequency drugs. Among common drug clustering methods, hierarchical clustering and maximal clique

收稿日期: 2025-03-21

基金项目: 国家重点研发计划 (2023YFC3502900)

作者简介: 张 迪, 硕士研究生, 从事中医古籍数字化研究。E-mail: zhangdi@itcm.ac.cn

*通信作者: 张伟娜, 研究员, 从事中医古籍数字化研究。E-mail: 523404824@qq.com

algorithms demonstrate the best partitioning performance, though the former may include low-relevance drugs in clusters, while the latter produces overly ambiguous category divisions. After optimization, the limitations of hierarchical clustering and maximal clique-based algorithms were mitigated. **Conclusion** For drug association calculations, Pearson, Spearman, and Phi correlation coefficients could be prioritized. In drug clustering analysis, the optimized application methods can be adopted: hierarchical clustering is recommended to analyze the agglomeration process of drug categories, whereas the maximal clique algorithm can be used to identify multidimensional application directions of drugs.

Key words: traditional Chinese medicine; prescription; data mining; association analysis; Pearson correlation coefficient; Spearman correlation coefficient; Phi correlation coefficient; system cluster analysis; maximal cliquet algorithm

处方数据挖掘是中医药现代化研究的重要方向之一，基于数据挖掘技术，研究者可通过对大量中医处方数据进行计算分析，挖掘出潜在有价值的药物配伍规律，如治疗某一疾病的核心药对或药组、潜在新方等^[1-2]。虽然此类研究的挖掘结果仍需进一步加工、分析与验证才能保证其可靠性，但数据挖掘方法作为一种低成本的干实验方法，将其用于研究初期，无疑可有效缩短研究者的研究周期，降低研究成本。

但现有中医处方数据挖掘研究所广泛应用的互信息、K 均值聚类等计算分析方法，多是直接借鉴自其他领域的方法，由于中医处方数据在形式、价值度量与维度上存在着相对于一般数据的特殊性，这些方法在实际应用时可能会出现研究者预期之外的错误，进而导致研究的不科学性。而现有研究中频繁出现的难以用中医理论解释的数据挖掘结果，也从侧面证实了这一现象的存在^[3]，但目前尚未有研究对这一问题进行深入探究。由于药物关联与聚类分析是中医处方数据挖掘研究中最常使用的分析方法，本研究选取了目前最常用的或被相关研究推荐的 10 种药物关联分析方法与 7 种药物聚类分析方法，从原理层面探究了各种方法对中医处方数据的适用性，并以举例或对比实验的形式对方法存在的局限性进行了验证，提出了部分应用价值较高的分析方法的优化方案，以期今后相关研究提供重要参考。

1 中医处方数据及药物关联与聚类分析方法概述

1.1 中医处方数据的概念

中医处方数据目前尚无公认的标准概念，但根据现有研究，可将其概括为医家在长期临床实践中积累的处方信息记录，这种记录可以包括药物、药物剂量、针灸选穴、对应症状等各类信息。由于药物与针灸选穴等处方主体部分是相关研究所分析的主要对象，因此本研究所指处方数据仅包括处方的主体部分，暂不考虑药物剂量、对应症状等附加信息。

对于部分研究，其研究对象并非处方，而是《中医方剂大辞典》《普济方》等方剂著作中的选方，或从其他文献中收集的方剂，由于此类数据在数据结构上与处方数据一致，也可看作是一种特殊的处方数据并适用本研究中对诸多挖掘方法的分析，但此类数据反映的医家思想并不统一，因此在挖掘结果的解读与分析上，应结合数据来源类型加以解读。

1.2 中医处方数据与一般数据的差异性

中医处方数据与其他领域研究中所分析的一般数据主要存在 3 种重要差异：一是数据形式的差异，二是价值度量的差异，三是数据维度的差异。这 3 种差异可能导致在其他研究领域中所适用的数据挖掘方法在中医药领域中可能会出现研究者预期之外的算法局限。

1.2.1 中医处方数据与一般数据的形式差异 在数据形式上，中医处方数据实际上是一种共现数据，即其数据集中所记录的是药物在各处方中的出现情况；而对于一般数据，记录的则是个体的属性信息。在应用数据挖掘方法时，中医处方数据一般被作为二值数据处理，为了更直观地对比 2 种数据的差异，本研究以汽车配置信息为例，设计了 1 个同样为二值数据的一般数据示例进行对比，如表 1 所示。

表 1 中分别记录了 2 类数据集内 2 个不同个体（药物 1 与药物 2、汽车 1 与汽车 2）的信息。其中药物 1 仅在处方 1 中有所应用，药物 2 仅在处方 2

表 1 中医处方数据与一般数据示例

Table 1 Examples of TCM prescription data and general data

中医处方数据			一般数据		
不同处方	药物 1	药物 2	配置	汽车 1	汽车 2
处方 1	1	0	全景天窗	1	0
处方 2	0	1	座椅加热	0	1
...	0	0	...	0	0
处方 n	0	0	语音控制	0	0

中有所应用；汽车 1 与汽车 2 仅在天窗和座椅上存在差异，其余配置均相同。从而推断药物 1 与药物 2 从未在同一处方中出现过，因此两者在应用时很可能适用于完全不同的病机；而汽车 1 与汽车 2 几乎全部配置相同，因此两者很可能具有一致的市场定位。即对于完全相同的数据，中医处方数据中药物的关联性和一般数据中个体的关联性呈现出了完全相反的结果。导致这一现象出现的本质原因在于对数据集中 0 值记录的分析方式差异。对于一般数据，2 个个体在同一属性上标注为 0，表明两者的该属性相同；对于中医处方数据，2 个药物在同一处方中标注为 0，仅能表明两者存在微弱的同时不适用于某一病机的潜在可能性。因此那些关注 0 值一致性的数据挖掘方法在用于中医处方数据时，其结果一般是不可靠的。

1.2.2 中医处方数据与一般数据的价值度量差异在价值度量上，一方面，由于中医处方数据的生成过程取决于临床中患者的实际就诊情况，这便不可避免地会引入频次偏差，即不同证型、不同病情的患者在临床中出现的频次是不一致的，但不同证型、不同病情的患者在生命价值上是平等的；另一方面，中医师在遣方时，往往会习惯于高频应用甘草等药性平和、适用范围广的药物，但这些药物对疾病往往不具备特异性的治疗作用，而那些使用频次低、药物偏性强的药物，则可能是治疗疾病的核心药物。因此，对于中医处方数据，药物的使用频次虽然是一种价值的参考指标，但绝不是价值的决定性指标。

而对于一般数据，尤其是商业数据，频次反映着用户交易量、网站流量等营销者最关注的信息，往往是价值最重要的度量指标。由于现有的数据挖掘方法多主要用于购买行为分析、客户群体分析等商业用途，如在中医药数据挖掘领域应用极为广泛的关联规则分析，又被称为购物篮分析，就是一种分析客户购买行为的分析方法。那么在应用这些分析方法时，就必须衡量频次因素在算法中的使用占比，若频次因素对计算结果的影响过大，则分析结果的全面性就可能不足，需要考虑更换分析方法。

1.2.3 中医处方数据与一般数据的维度差异中医处方数据在数据的组织结构上一般由方与药构成，其中，药是分析的对象，方则是药物的特征。数据挖掘方法通过分析药物是否在各处方中使用，来实现对药物的关联与聚类分析。对于一般数

据，往往分析对象的数量要远大于分析对象特征的数量，如鸢尾花数据集包含 150 个样本，却仅有 4 个特征；而对于中医处方数据，由于常用药物的种数是相对有限的，处方数量则没有上限，这便导致了其特征数量可能会接近甚至超过分析对象的数量，若分析方法无法处理这种高维数据，就会导致结果的不可靠性。

另一方面，药物的属性也是多维的，一种药物往往具有多种功效，那么在对其进行类别划分时就必须考虑到其潜在的其他划分方法，即应实现类的重叠划分。但传统的数据挖掘方法并不都具备此特性，这也会影响挖掘结果的可靠性。

1.3 药物关联与聚类分析的概念

药物关联与聚类分析是目前中医处方数据挖掘研究中应用最广泛的分析方法，可对处方数据集中的药物分布、药物组合、基础方等进行挖掘。在本研究中，药物关联分析指运用算法度量处方数据集中药物或药物组合间关联程度的方法；药物聚类分析指运用算法自动地将处方数据集中的药物划分为多个由高关联药物组成的药物群的方法。从算法设计的角度，前者是后者的重要组成部分，后者则是对前者的一种拓展与延伸。

2 整体分析流程

本研究旨在基于现有的常用关联与聚类分析方法，提出一种在中医处方数据挖掘研究中关联与聚类分析部分的良好实践方案。整体上，本研究采用先原理分析、后举例验证的形式分析各数据挖掘方法对中医处方数据的适用性，详述应用较多或效果较好的分析方法，简述应用较少或效果较差的分析方法。

对于关联分析方法，由于所有方法均以计算药物间关联度为目的，因此可以进行横向对比，以突出不同计算方法存在的局限性，故本研究还基于真实医案数据设计了对比实验。对于聚类方法，由于其准确性主要受内部选择的关联分析算法影响，差异性主要体现在结果形式上，因此仅通过所举案例就能对比出方法间的差异，故无须设计对比实验。

在本研究中，由于存在计算结果相对准确的常用关联分析方法，但常用聚类方法应用于中医处方数据时均存在不同程度的局限性，因此在聚类方法部分，本研究主要聚焦于效果较好的、具备较高优化价值的分析方法，并提出其应用时的优化方案。本研究的整体分析流程如图 1 所示。

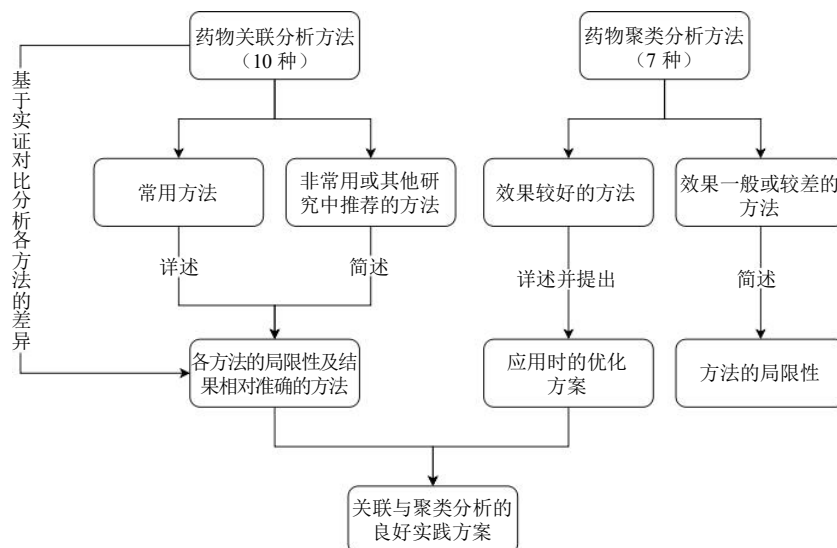


图 1 分析流程

Fig. 1 Analysis process

3 常用药物关联分析方法的适用性分析

药物关联分析方法常用于药物组合的挖掘，也被用于药物聚类分析时药物间联系密切程度的度量。在药物聚类分析中，药物间关联度高的药物往往被聚为一类，因此药物关联分析方法的选择也决定着药物聚类分析的准确性。现对常用药物关联分析方法进行梳理分析。

3.1 关联规则分析中的药物关联度算法

3.1.1 关联规则分析概述 关联规则分析最初被设计为一种用户购买行为分析方法，在中医处方数据挖掘研究中则常用于挖掘药物组合间的关联关系。该方法由频繁项集挖掘与规则挖掘 2 步组成，其中前者指出现频率较高的药物组合，常用 Apriori^[4]、FP-Growth^[5]等算法进行挖掘，但各算法除运算速度外并无明显区别，一般无须过度关注；后者则以支持度、置信度、提升度 3 种药物关联度算法最为常用。

3.1.2 常用关联规则分析指标应用于中医处方数据时存在的局限性 虽然关联规则分析是目前中医处方数据挖掘研究中应用最广泛的分析方法之一。但在实际应用中，3 种药物关联度算法均存在一定问题。

支持度指 2 个药物组合共现的概率，因此该指标大小受药物频次影响极大；置信度指 2 个药物组合出现在处方中的条件概率，由于高频药物更容易出现在处方中，因此涉及甘草等高频药物的药物组合置信度往往也较高；提升度指药物组合的联合概

率与边缘概率之比，但对于高频药物，其边缘概率较高，而联合概率最大值为 1，故提升度往往偏小。因此，上述常用的关联规则分析指标均存在不同程度的频次偏见问题，对高频或低频药物的计算结果往往不够准确。如赵乙川等^[6]在分析郑洪新教授治疗不寐兼郁证处方时，将关联规则挖掘结果按支持度降序排列，令其结果的支持度均在 80% 以上，排除了所有使用频次在 190 次以下的药物组合，而保留规则的置信度均在 85% 以上，普遍高于 95%，但绝大部分规则的提升度均极接近 100%。该示例所呈现的分析结果与前文中对各关联规则分析指标的描述一致，这说明传统的关联规则方法作为分析用户购买行为的方法，存在较严重的频次偏见，对中医处方数据的适用性较为有限。

3.2 欧氏距离

3.2.1 欧氏距离的计算方法 欧氏距离指欧几里得空间中两点的真实距离。对于欧几里得空间中的 2 个点 A 与 B ，其欧氏距离用公式可如下表示。

$$D(A, B) = \sqrt{\sum_{i=1}^n (A_i - B_i)^2}$$

对于中医处方数据， i 可理解为处方编号， n 可理解为总处方数，当处方中包含指定药物时，记录该药物为 1，否则为 0。欧氏距离是 K 均值聚类算法的默认关联度计算方法，也是目前中医处方数据挖掘研究中应用较多的关联度计算方法之一。

3.2.2 欧氏距离应用于中医处方数据时存在的局

限性 由计算方法可知,欧氏距离是典型的关注 0 值一致性的关联度计算方法,由于应用频次较低的药物在多数处方中的值均为 0,其距离的计算结果也必趋近于 0,因此会特别高估低频药物间的关联性。而若将其用于聚类分析中,则会导致所有低频药物被聚为一类,甚至由于低频药物种类过多,绝大多数药物都被聚为一类而形成超大类,从而得出无效结果或诱导产生错误的分析。

以代俊泽等^[7]的研究为例,该研究在应用聚类方法分析中药治疗糖尿病合并骨质疏松症的处方时,选择了欧氏距离作为关联度计算方法,共得到 4 类聚类结果。而其第 1 类结果所呈现的 12 味中药正是其聚类分析结果中频次最低的 12 味中药,且该类药物数量远超其他类别,验证了前文中所提到的欧氏距离局限。

3.3 互信息与改良互信息

改良互信息是中医处方数据挖掘研究中应用最为广泛的药物关联度计算方法,该方法由互信息算法改良而来,因此在分析该方法的适用性前,需要先对互信息算法进行介绍。

3.3.1 互信息的计算方法与通俗理解 互信息指 1 个随机变量因另 1 个随机变量而减少的不确定性,对于随机变量 A 和 B ,其互信息可如下表示。

$$I(A, B) = \sum_{i=1}^n \sum_{j=1}^m P(A_i, B_j) \log \frac{P(A_i, B_j)}{P(A_i)P(B_j)}$$

$P(A_i, B_j)$ 表示 A 和 B 所有可能取值组合的概率, $\frac{P(A_i, B_j)}{P(A_i)P(B_j)}$ 表示

所有组合的联合概率与边缘概率之比,即关联规则中的提升度

3.3.2 互信息应用于中医处方数据时存在的局限性 根据互信息公式可知, A 和 B 的互信息取决于 A 和 B 所有可能取值组合下的提升度,即互信息所衡量的不确定性的减少,实际上同时受 2 个事件正相关与负相关程度的影响。类比至中医处方数据,若 2 味药物具有强烈的互斥性,也会被认为是高关联药物,这显然与中医处方数据挖掘中所需要的药物间关联度不是同一概念。因此互信息是完全不适用于中医处方数据的药物关联度计算方法。

3.3.3 改良互信息的计算方法及应用于中医处方数据时存在的局限性 为解决互信息的局限性,有学者提出了改良互信息的概念,目前应用最广泛的是中医传承辅助平台团队提出的方法^[8],该方法通过为低共现药物设置惩罚系数来区分药物关联的

正负性,若药物共现频次低于预设值,则强制调低其关联度计算结果。

这种基于共现频率与惩罚系数的改良互信息指标与互信息相比具有更高的可行性,但同时也引入了频次偏见,导致低频药物的相关性难以被挖掘。且该方法并未消除互信息在原理上与中医处方数据挖掘中所指的药物关联度的差异性。因此其对中医处方数据的适用性仍十分有限。如张翱等^[9]的研究中基于改良互信息指标采用复杂系统熵聚类算法得到的多个分类中均含有薏苡仁、黄柏、苍术、牛膝、土茯苓,而这 5 个药物恰是其数据来源中使用频次前 5 的中药,这说明改良互信息指标确实存在对低频药物的歧视现象,可能会影响分析结果的客观性。

3.4 Pearson、Spearman 和 Phi 相关系数

相关性检验是统计学中检验变量间相关性的方法,其中最常使用的是 Pearson 和 Spearman 相关性检验,对于二值变量,常使用 Phi 相关系数计算其相关性。在中医处方数据挖掘研究中,以 Pearson 相关性检验的应用最为广泛,如石若玉等^[10]基于 Pearson 相关系数对中医药治疗 H 型高血压病处方中高频药物的相关性进行了分析。

3.4.1 3 种相关系数应用于中医处方数据挖掘时的关系及计算方法 从算法的角度看, Spearman 相关可看作是一种对于等级变量的 Pearson 相关,而 Phi 相关系数则是两者的特例。因此对于中医处方数据, Pearson、Spearman 和 Phi 相关系数是等值的, Phi 相关系数可看作是前两者的一种简化形态。将 2 个药物的关系汇总于列联表, Phi 相关系数可如下表示。

$$\varphi = \frac{n_{00}n_{11} - n_{01}n_{10}}{\sqrt{(n_{00} + n_{01})(n_{10} + n_{11})(n_{00} + n_{10})(n_{01} + n_{11})}}$$

3.4.2 3 种相关系数在计算上的合理性分析 不难看出, Phi 相关系数是通过衡量 2 个变量一致(即同为 0 或同为 1)与不一致的分布来计算变量间相关性的。这种思想与互信息类似,但互信息将变量一致与不一致时的点互信息累加,而 Phi 相关系数则将两者的分布相减,更切合实际应用。同时,由于 Phi 相关系数在计算时还对变量的各种状态频次做了除法运算,一定程度上消除了变量间的频次差异,可近似看作是一种基于概率的算法。因此,在中医处方数据挖掘研究中,更推荐使用 Pearson、Spearman 和 Phi 相关系数作为药物间关联程度的度量指标。

3.5 其他关联度计算方法

除前文所论述的药物关联分析方法外，有学者提出由于欧氏距离是一种应用于连续变量的距离，应当使用汉明距离代替欧氏距离，同时应关注基于集合重叠程度计算的 Jaccard 系数指标^[11]。余弦相

似度作为计算机领域最常用的关联度计算方法之一，其性质与 Jaccard 系数类似，且在中医处方数据挖掘研究中同样有一定应用。上述方法均为对前文所述方法的补充，但在实际应用中仍存在一定问题，其在中医处方数据挖掘研究中的局限性见表 2。

表 2 其他关联度计算方法存在的局限性

Table 2 Limitations of other correlation calculation methods

挖掘方法	局限性
汉明距离 ^[12]	汉明距离通过计算使用其中 1 味药时，不使用另一味药的处方数来度量药物间的关联性，其本质与欧氏距离并无较大差异；即对于低频药物，即使 2 种药物不存在交集，由于药物频次较低，依然会得出较小的汉明距离
Jaccard 系数 ^[13]	Jaccard 系数通过计算含 2 味药物处方的交集与并集之比度量药物间的关联性。其交集的最大值取决于频次较低的药物，并集的最大值取决于频次较高的药物，因此当 2 种药物频次差异较大时，其 Jaccard 系数会偏低
余弦相似度 ^[14]	余弦相似度通过计算含 2 味药物处方（即药物向量中值为 1 的部分）的分布差异来度量药物间的关联性；由于高频药物向量中的 1 值必然显著多于低频药物，即使高频药物完全包含低频药物，其余弦相似度必然也会偏低，因此余弦相似度与 Jaccard 系数存在相似的局限性

3.6 各关联度计算方法的实证对比分析

3.6.1 实验设计与结果 为了更直观地观察应用各关联度计算方法分析药物关联程度时存在的问题，本研究基于《丁甘仁医案》数据集^[15]中的真实医案数据

对各关联度计算方法进行对比分析。使用各关联度方法两两计算药物间的关联程度，列举不同频次分布的药物组合示例，对各关联度计算方法的结果进行分析。各药物组合示例信息与结果如表 3 所示。

表 3 各关联度指标的计算结果对比

Table 3 Comparison of calculation results of various correlation indexes

序号	各药物组合信息						支持度	最高置信度
	药物1	频次	药物2	频次	共现频次	组合性质		
1	陈皮	119	厚朴	21	13	经典药对	0.036	0.619
2	茯苓	116	党参	45	20	经典药对	0.056	0.444
3	栀子	40	茵陈	13	6	经典药对	0.017	0.462
4	川贝母	70	天花粉	31	11	经典药对	0.031	0.354
5	陈皮	119	茯苓	116	68	经典药对	0.190	0.586
6	白芍	84	当归	79	36	经典药对	0.101	0.456
7	六神曲	29	山楂	29	8	经典药对	0.022	0.276
8	甘草	137	茯神	76	27	普通组合	0.075	0.355
9	甘草	137	半夏	107	33	普通组合	0.092	0.308
10	荷叶	49	杏仁	48	6	普通组合	0.017	0.125
11	远志	51	生地黄	47	6	普通组合	0.017	0.128
序号	提升度	欧氏距离	汉明距离	余弦相似度	Jaccard系数	互信息	改良互信息	统计学相关系数
1	1.862	10.677	114	0.260	0.102	0.011	-0.838	0.152
2	1.372	11.000	121	0.277	0.142	1.008	1.008	0.098
3	4.131	6.403	41	0.263	0.128	0.015	-0.476	0.216
4	1.815	8.888	79	0.236	0.122	0.007	-0.775	0.124
5	1.764	9.950	99	0.579	0.407	1.266	1.266	0.373
6	1.942	9.539	91	0.442	0.283	1.073	1.073	0.278
7	3.405	6.481	42	0.276	0.160	0.015	-0.532	0.212
8	0.928	12.610	159	0.265	0.145	1.182	1.182	-0.029
9	0.806	13.341	178	0.273	0.156	1.275	1.275	-0.100
10	0.913	9.220	85	0.123	0.066	0.000	-0.793	-0.014
11	0.896	9.274	86	0.123	0.065	0.000	-0.798	-0.016

组合 1~7 均为在真实中医临床环境中常用的经典药对,组合 8~11 则为在临床遣方时无显著关联的普通药物组合。通过药物频次与总处方数,基于超几何分布的生存函数^[16]计算各组合药物间随机状态下共现频次 $\geq$ 真实频次的概率,组合 1~7 的平均概率为 0.47%,最高概率为 2.37%;组合 8~11 的平均概率为 65.97%,最低概率为 49.62%。即在中医理论与数学概率 2 个层面,组合 1~7 的药物间关联程度均显著高于组合 8~11。

在频次组合设计上,组合 1、2 由高频药物与低频药物组成;组合 3、4 由中频药物与低频药物组成;组合 5~7 分别由频次相似的高、中、低频药物组成;组合 8、9 由高频药物组成;组合 10、11 由低频药物组成。由于当药物的频次差较大且呈非正相关时,所有算法均能正确区分药物间的关联性质,故不作讨论。

3.6.2 实验结果分析 为确保对各方法计算结果评估的客观性,本研究主要对药物关联的性质进行分析,当某种算法对经典药对关联度的计算结果小于普通药物组合时,则认为算法对该组药物的度量不够准确。观察可知,支持度对组合 1~4 与组合 7 的计算结果均低于组合 8、9,而前者为经典药对,后者为普通药物组合。这说明对于存在低频药物的组合,其支持度可能会偏低;而对于高频药物组合,则可能会偏高。

最高置信度对组合 4 与组合 7 的计算结果低于组合 8、9。这说明置信度存在与支持度类似的问题,但若共现频次接近组合中的低频药物频次,置信度仍可正确度量其关联性。

组合 1、2 与组合 5、6 的欧氏距离与汉明距离均高于组合 11、12。说明对于存在高频药物的组合,欧氏距离与汉明距离可能会低估其关联性;而对于低频药物组合,则可能会高估其关联性。

余弦相似度与 Jaccard 系数对组合 1~4 的计算结果均低于或接近组合 8、9。说明余弦相似度与 Jaccard 系数会低估频次差异较大的药物组合的关联度,高估高频药物的关联度。

互信息与改良互信息对全部组合的计算结果几乎均存在异常,其中互信息的计算结果无明显规律,改良互信息的计算结果则与共现频次高度相关。这说明互信息可能不适用于任何频次药物组合关联度的计算,改良互信息的适用范围则与支持度一致。

提升度与统计学相关系数的计算结果始终准确,但观察可知,提升度对低频药物关联程度的计算结果普遍高于高频药物,这说明提升度对高频事件存在一定程度的不敏感性。

3.6.3 结论 由上述分析可知,现有的主流药物间关联度计算方法中仅 Pearson、Spearman 和 Phi 相关系数是相对准确的,提升度能保证对药物间关联程度性质判断的准确但无法保证数值大小的可靠,其余关联度指标则连药物间关联度性质判断的准确也无法保证。故研究者在进行药物关联分析或确定聚类方法的关联度指标时应尽量选择 Pearson、Spearman 和 Phi 相关系数,避免选择其余指标。

4 常用聚类分析方法的特性及适用性分析

聚类分析目前常用于基础方、新方、治法和病机等的挖掘,近年来在处方数据挖掘研究中应用较多的聚类分析方法包括系统聚类、K 均值聚类、基于熵方法的聚类等。从分析形式的角度看,因子分析、极大团算法、社区发现算法与隐结构模型等其他常用分析方法也可视为特殊的聚类分析方法。整体而言,在上述方法中,系统聚类与极大团算法是目前相对最为成熟且适用于中医处方数据的分析方法。

4.1 系统聚类分析

系统聚类是计算机领域的经典聚类算法,也是在中医处方数据挖掘研究中应用最为广泛的聚类方法。如于佳佳等^[17]运用系统聚类法对《中华医典》中治疗胃食管反流病的古代中医方剂、医案进行分析,得到了 6 组中医药治疗胃食管反流病的核心药物组成。

4.1.1 系统聚类应用于中医处方数据时存在的局限性 系统聚类在划分药物类别时,首先将所有药物视为单独的一类,然后迭代合并最相似的类别,最终实现药物从小类到大类的凝聚。该方法的可解释性与可视化效果较好,无须提前设置聚类数量,算法结果稳定,因此对中医处方数据的适用性较强。

但应用该方法分析中医处方数据时,仍存在一定问题。首先,该方法的聚类结果无法反映药物与最终归属类别的关联程度,仅能对类别做出划分,因此计算得出的药物很可能实际上并不适合归入其所属分类,对于处方挖掘研究,若分析时不加以明确,则可能对研究者产生误导。其次,该方法作为传统的聚类方法,不支持类的重叠划分,而中药在真实临床环境中往往适用于多种中医病机。因此

该方法分析得出的药物划分实际上是较为粗浅的，难以深入挖掘处方中药物间复杂的配伍关系。

4.1.2 系统聚类分析在应用时的优化方法 在应用系统聚类分析中医处方数据时，可在完成类别划分后，补充计算各药物与类中其他药物的平均相似度，基于该结果结合中医理论对类别属性再做分析。对于类的重叠划分问题，则推荐借助极大团算法等擅长划分重叠类别的方法，对数据进行综合分析。

以《丁甘仁医案》数据集为例，检索正文中包含“咳”字的医案，探究丁甘仁治咳的用药规律。经检索共得到 63 个处方。为提高分析结果的可靠性，取频次 ≥ 10 的药物作为分析对象，共计 21 味药物。将 Pearson 相关系数作为药物间关联程度的度量指标，将所有药物间的平均距离作为类间距离，对各药物执行系统聚类，最后补充计算各药物与类中其他药物的平均 Pearson 相关系数。结果如图 2 所示。

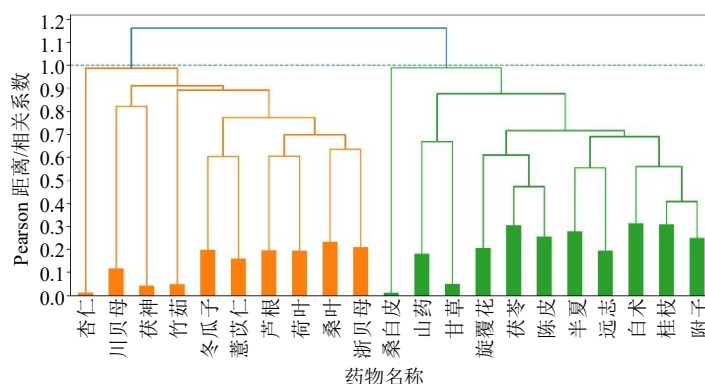


图 2 系统聚类分析结果

Fig. 2 Result of systematic clustering analysis

图 2 中树状部分的纵坐标为 Pearson 距离（1-皮尔逊相关系数），柱状部分为药物与所属类的平均 Pearson 相关系数。桔色药物中杏仁、茯苓、竹茹的柱状部分较低；绿色药物中桑白皮、甘草的柱状部分较低。提示上述药物与药物类别的整体属性关联性相对较弱，分析时应加以鉴别。

根据中医理论，图 2 中桔色药物以凉性药物为主，主治热证；绿色药物以温平性药物为主，主治寒证。因此，杏仁、茯苓作为温平性药物，桑白皮作为凉性药物，与类中其余药物的关联性较差。中医认为，痰饮为阴邪，多见于寒证，故竹茹作为化痰药物，在热证中应用相对较少。甘草具有调和药性的作用，其应用与患者的寒热属性无显著关联。因此，上述药物的低关联性均符合中医理论，通过补充计算各药物与类中其他药物的平均相似度，既可以辅助明确药物类别的整体属性，也可以对医家遣方时的配伍思想作出更深入的分析。

4.2 极大团算法

4.2.1 极大团算法的概念与特性 极大团算法是一种基于图论的复杂网络分析算法，将药物间的关联关系构建为网络，若网络中存在子网络，其内部

所有的节点间都有边，则可称之为 1 个团；若 1 个团不是任何其他团的子集，则可称其为极大团，求解极大团的算法即极大团算法。

目前最常用的极大团算法为 Bron-Kerbosch (BK) 算法^[18]。极大团算法与传统聚类最大的区别在于，极大团算法不强制要求覆盖所有样本，极大团内的每一味药与其他药都是高关联的，因此这种算法可以直接得出数据集中的骨干结构，而无须做额外的分析。此外，极大团允许样本重叠划分，且与其他算法相比，极大团算法得到的结果通常具有更高的重叠率，更符合中药多功效的特性。

4.2.2 极大团算法应用于中医处方数据时存在的局限性 极大团算法在中医处方数据挖掘研究中的局限性来自其结果高度重叠带来的划分模糊性。如邱磊等^[19]运用 BK 算法得到的处方挖掘结果，各类中有半数以上甚至接近全部的药物是与其他类重复的药物，这种类间差异不显著的情况，会令研究者很难概括各类的实际含义，进而干扰后续分析。计算《丁甘仁医案》治咳处方中存在的药物极大团，同样存在该现象，如图 3 所示。

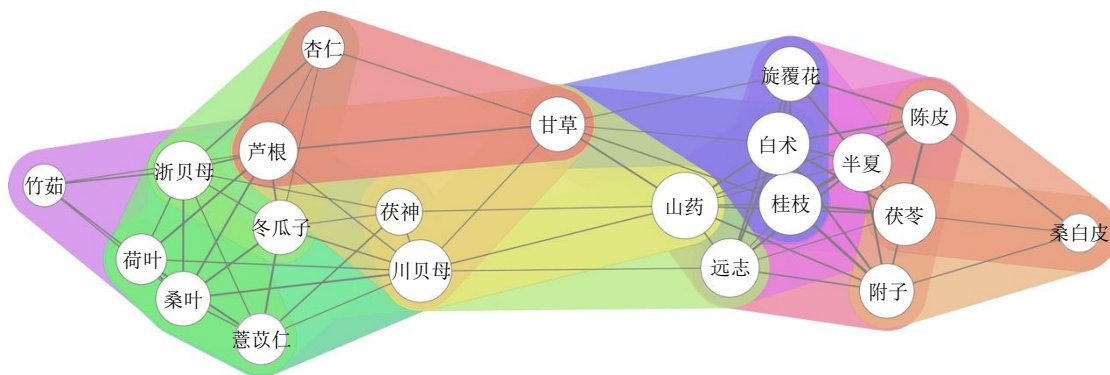


图 3 极大团算法的计算结果

Fig. 3 Calculation result of maximal clique algorithm

图 3 中共存在 24 个极大团, 且不同极大团内的药物高度重叠, 因此在研究中直接应用极大团算法对中医处方数据进行分析是几乎不具备可行性的。

4.2.3 极大团算法在应用时的优化方法 为降低极大团算法划分结果的模糊性, 同时保留其划分重叠类别的能力, 可基于相似度算法对极大团算法的

划分结果作进一步合并。

仍以《丁甘仁医案》治咳处方数据为例, 利用 rapidfuzz 的 partial_ratio 算法^[20]两两计算各极大团的相似度, 为相似度 > 70 的极大团对建立边, 边权重为相似度, 通过 Louvain 算法将各极大团划分为多个社区, 最后合并各社区内极大团中的所有药物, 结果如图 4 所示。

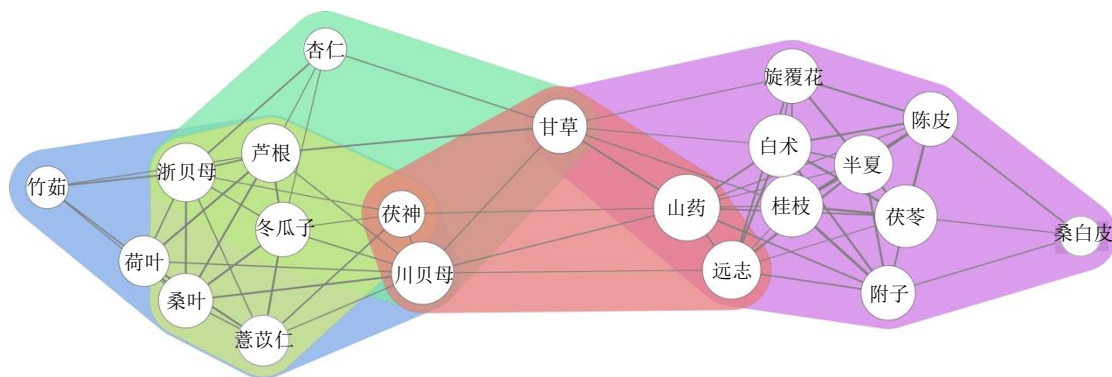


图 4 优化后的极大团算法计算结果

Fig. 4 Calculation result of optimized maximal clique algorithm

图 4 中共有 5 个药物类别, 结合中医理论, 可将其属性从左至右概括为清热类药物、泻火消痢类药物、清热润肺类药物、养阴安神类药物、温补化痰类药物。各类药物的主治差异清晰, 且存在合适的重叠, 划分效果较好。即通过对极大团的合并处理, 其划分结果的模糊性显著降低, 同时还可一定程度上体现出中药的多维度治疗作用, 令其更加适用于中医处方数据挖掘。

4.3 其他聚类分析方法

除系统聚类与极大团算法外, 其他分析方法对中医处方数据的适用性则相对较弱, 或发展尚未成熟, 其局限性见表 4。

5 结论与讨论

5.1 结论

整体而言, 常用的药物关联分析方法均存在不同程度的局限性, 仅 Pearson、Spearman 和 Phi 相关系数对药物关联程度的度量是相对准确的。因此仅推荐使用此方法进行药物关联分析。

常用的药物聚类分析方法中, 系统聚类和极大团算法对药物类别的划分效果较好, 但前者易在类中纳入低相关药物, 后者的类别划分结果则过于模糊, 因此本研究同时提出了 2 种算法在应用时的优化方法。与现有常用方法相比, 优化后的方法对中医处方数据适用性更佳, 可胜任更加多元与深入的

表 4 其他聚类分析方法的局限性
Table 4 Limitations of other cluster analysis methods

挖掘方法	局限性
K 均值聚类 ^[21-22]	作为传统的聚类算法，存在与系统聚类相似的局限性，但该算法仍存在其他局限性：在计算过程中存在随机行为，导致其结果存在一定的不确定性；需提前确定聚类数量；默认采用欧氏距离作为药物间关联程度的度量指标，令其结果易受药物频次影响；该算法相对于系统聚类，运算速度较快，但中医处方数据受药物种类限制，往往规模不大，故其优势并不突出
复杂系统熵聚类 ^[23-24]	本质上是一种通过剪枝改良的极大团算法，但其仅保留节点相似度最高的固定边数，令结果规模受限于固定边数的设置，类包含最大药物数无法超过该值，不符合客观规律；此外，该算法采用改良互信息作为药物间关联程度的度量指标，导致结果极易受药物频次影响
因子分析 ^[25-26]	无法挖掘药物间的非线性关联关系，且公共因子数量较难确定；虽然可实现类的重叠划分，并呈现药物与所在类的关联强度，但前者的划分效果不如极大团算法，后者的呈现效果不如直接补充计算各药物与其所属类的关联强度；由于该方法普遍包含于统计学工具中，对研究者的专业技能要求较低，若无条件对系统聚类与极大团算法的计算结果作进一步处理，该方法也可作为二者的辅助分析方法
隐结构模型 ^[27-28]	实际上是一种引入了隐变量的树状贝叶斯网络模型，最常使用的 LTM-EAST 算法通过启发式算法寻找拥有高贝叶斯信息标准评分（Bayesian information criterion，BIC）的结构作为最优结构，但传统的 BIC 对隐结构模型的适用性还有待进一步验证，隐变量的推测准确率仍有提升空间 ^[29] ，且该算法不支持类的重叠划分
社区发现 ^[30-31]	在中医处方数据挖掘研究中应用较少，最常用的 Louvain 算法仅具备划分能力，不具备任何其他优势；虽然其他社区发现算法 ^[32-34] 在灵活性与可解释性等方面可能存在优势，但开发程度较差

研究。在聚类分析部分，可通过系统聚类分析药物类别的树状凝聚过程，通过极大团算法探索药物的多维应用方向。

5.2 本研究的意义

本研究作为首个从算法角度论述常用数据挖掘方法对中医处方数据适用性的研究，与既往研究^[1-3,11,35-39]相比，本研究提出了中医处方数据在形式、价值度量及维度层面的特殊性，并重点关注方法应用时存在的局限性。本研究还结合算法的计算过程与中医处方数据特点，指出了常用分析方法产生局限性的原因，通过实证与举例分析呈现了方法局限性对分析结果的影响，并对系统聚类和极大团算法的分析过程进行了优化。本研究可成为相关研究者选择分析方法、提出或改良分析方法的重要参考。

5.3 展望

本研究对近年来在中医处方数据挖掘研究中常用的药物关联与聚类分析方法进行了较为细致与全面的梳理，重点分析了各方法应用于中医处方数据时存在的局限性。从研究结果上看，现有研究在药物关联和聚类分析部分的结果质量是不容乐观的。这说明目前的中医处方数据挖掘体系尚未真正成熟，还需针对中医药发展中面临的实际问题与

已有方法产生局限性的原因，逐步形成具有中医药特色的数据挖掘方法与研究范式。

利益冲突 所有作者均声明不存在利益冲突

参考文献

[1] 李蕙质,周小玲,杨玉杰,等.数据挖掘算法在中药方剂研究中的应用现状[J].中国药房,2024,35(1):112-118.

[2] 刘梦玲,章新友,丁亮,等.数据挖掘方法在中药配伍规律研究中的应用与进展[J].中国中药杂志,2021,46(20):5233-5239.

[3] 潘晔,娄静,潘玉颖,等.中医药数据挖掘现状分析与创新探索[J].中国中医药信息杂志,2022,29(5):5-9.

[4] Agrawal R, Srikant R. Fast algorithms for mining association rules [A] // Proc of the 20th International Conference on Very Large Data Bases [C]. Santiago: Morgan Kaufmann, 1994: 487-499.

[5] Han J W, Pei J, Yin Y W. Mining frequent patterns without candidate generation [J]. SIGMOD Rec, 2000, 29(2): 1-12.

[6] 赵乙川,杨硕,郑洪新.基于数据可视化分析郑洪新教授治疗不寐兼郁证的核心组方与用药特色[J].世界科学技术—中医药现代化,2024,26(9):2242-2253.

[7] 代俊泽,刘毅,廖翠平,等.中药组方治疗糖尿病合并骨质疏松症用药规律挖掘及其作用机制的网络药理学与分子对接研究[J].中华中医药学刊,2024,42(12):158-163.

[8] 唐仕欢,陈建新,杨洪军,等.证-熵-方:中医药研究

- 的新领域 [J]. 中国中医基础医学杂志, 2010, 16(4): 267-270.
- [9] 王嘉欣, 姚鉴玲, 马嘉慕, 等. 基于 AHP-SOM 聚类-TOPSIS 和中医传承辅助平台研究中医药治疗围绝经期抑郁症组方规律 [J]. 中草药, 2022, 53(22): 7153-7163.
- [10] 石若玉, 李东东. 基于数据挖掘探讨中医药治疗 H 型高血压病的用药规律 [J]. 中国中药杂志, 2024, 49(9): 2522-2531.
- [11] 但文超, 赵国栋, 何庆勇, 等. 中医药处方数据挖掘的常见问题辨析与展望 [J]. 中国中药杂志, 2023, 48(17): 4812-4818.
- [12] 曹峻文, 叶晖, 黄秋月, 等. 岐黄学者张学智培土畅中法治疗幽门螺杆菌根除后化生性慢性萎缩性胃炎的处方用药规律 [J]. 中国中西医结合消化杂志, 2024, 32(7): 568-574.
- [13] 吴小欣, 李敏, 梁滨, 等. 穴位贴敷治疗脑卒中优势病症的选穴用药规律 [J]. 广州中医药大学学报, 2024, 41(4): 1074-1082.
- [14] 何菊, 杨涛, 戴彩艳, 等. 基于复杂网络的名老中医治疗肺癌用药规律研究 [J]. 世界科学技术—中医药现代化, 2020, 22(4): 1020-1024.
- [15] 张迪, 牛琪锴, 赵钰文, 等. 基于全集提升度与重复 SLPA 算法的病机视角下中医药组方配伍规律挖掘及同效药鉴别方法 [J]. 中草药, 2024, 55(23): 8101-8110.
- [16] Scipy.stats.hypergeom—SciPy v1.15.2 Manual [EB/OL]. [2025-04-14]. <https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.hypergeom.html>.
- [17] 于佳佳, 刘梦, 王晓强, 等. 基于中医古籍数据挖掘探讨食管反流病的用药规律及有毒中药的应用 [J]. 中华全科医学, 2025, 23(2): 309-313.
- [18] Bron C, Kerbosch J. Algorithm 457: Finding all cliques of an undirected graph [J]. *Commun ACM*, 1973, 16(9): 575-577.
- [19] 邱磊, 杨铭, 郭晓燕, 等. 基于数据挖掘的肾综合征出血热中医证治组方规律研究 [J]. 中国中医药信息杂志, 2021, 28(4): 56-62.
- [20] Indel—RapidFuzz 3.10.1 documentation [EB/OL]. [2024-12-18]. https://rapidfuzz.github.io/RapidFuzz/Usage/distance/Indel.html#rapidfuzz.distance.Indel.normalized_similarity.
- [21] 万玛措, 斗周才让, 周则加, 等. 国医大师尼玛治疗肺热病的用药组方规律研究 [J]. 中成药, 2023, 45(5): 1735-1740.
- [22] 毛永馨, 周莉, 秦菲, 等. 基于数据挖掘、网络药理学、分子模拟和实验验证探究中药治疗糖尿病足溃疡用药规律及作用机制 [J]. 中草药, 2024, 55(16): 5559-5572.
- [23] 公培强, 沈元良, 蒋元烨. 基于中医传承辅助平台结合绍派伤寒学术思想挖掘治疗胁痛用药组方规律 [J/OL]. 辽宁中医杂志 (2024-07-11). <https://kns.cnki.net/kcms/detail/21.1128.R.20240708.1708.146.html>.
- [24] 王凯悦, 常惠, 王宇, 等. 基于数据挖掘探讨中药治疗妊娠恶阻的用药规律 [J]. 世界中医药, 2024, 19(3): 403-409.
- [25] 严如根, 何静, 薛晔佳, 等. 中药灌肠方治疗盆腔炎性疾病后遗症的用药规律研究 [J]. 中成药, 2024, 46(10): 3492-3496.
- [26] 曹灿, 徐向楠, 吴美晶, 等. 基于数据挖掘的半夏-乌头类反药组合临床同方使用规律分析 [J]. 中草药, 2024, 55(22): 7746-7755.
- [27] 张庆, 姬文帅, 孔欣欣, 等. 基于关联规则和隐结构模型的《普济方》中治疗喘证方剂的用药规律分析 [J]. 中草药, 2023, 54(5): 1517-1525.
- [28] 辛静, 蒋士卿, 张云慧, 等. 中医药治疗肺癌的用药规律 [J]. 世界中医药, 2024, 19(9): 1316-1323.
- [29] Njah H, Jamoussi S, Mahdi W, et al. A new equilibrium criterion for learning the cardinality of latent variables [A] // 2015 IEEE 27th International Conference on Tools with Artificial Intelligence (ICTAI) [C]. Vietri sul Mare: IEEE, 2015: 958-965.
- [30] 王家悦, 李萍, 周冬梅, 等. 对 1 049 例寻常型银屑病患者病程证候、证素演变规律的回溯性研究 [J]. 北京中医药大学学报, 2024, 47(10): 1438-1448.
- [31] 孙春全, 黎元元, 谢雁鸣, 等. 真实世界中银杏叶滴丸治疗冠心病与脑梗死的合并用药方案研究 [J]. 世界科学技术—中医药现代化, 2021, 23(3): 807-814.
- [32] Jin D, Yu Z Z, Jiao P F, et al. A survey of community detection approaches: From statistical modeling to deep learning [J]. *IEEE Trans Knowl Data Eng*, 2023, 35(2): 1149-1170.
- [33] Xie J R, Szymanski B K, Liu X M. SLPA: Uncovering overlapping communities in social networks via a speaker-listener interaction dynamic process [A] // 2011 IEEE 11th International Conference on Data Mining Workshops [C]. Vancouver: IEEE, 2011: 344-349.
- [34] Gregory S. Finding overlapping communities in networks by label propagation [J]. *New J Phys*, 2010, 12(10): 103018.
- [35] 赵艳青, 李青松, 项敏泓, 等. 中医药数据挖掘中常见问题的思考与策略: 以中医药术语及用药规律总结为例 [J]. 中华中医药杂志, 2018, 33(4): 1220-1225.
- [36] 仲芳, 杨巍, 赵翀, 等. 数据挖掘技术在中医医案的应用研究 [J]. 中国中医药信息杂志, 2020, 27(2): 141-144.
- [37] 朱彦, 朱玲, 崔蒙. 论中医药数据挖掘 [J]. 中华中医药杂志, 2016, 31(8): 2932-2935.
- [38] 崔蒙, 吴朝晖, 乔延江. 中医药信息学 [M]. 北京: 科学出版社, 2015.
- [39] 张文学, 刘敬霞, 梁希森, 等. 中医药方剂数据挖掘方法与应用 [M]. 北京: 中国中医药出版社, 2019.

[责任编辑 潘明佳]