

• 数据挖掘与循证医学 •

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

张迪¹, 牛琪锴², 赵钰文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. 中国中医科学院中药研究所, 北京 100700

摘要: 目的 针对现有中医处方普遍存在的挖掘深度不足、适用性差等问题, 提出一种新的改良中医处方数据挖掘方法。方法 基于复杂网络分析, 融合关联规则与聚类分析, 通过挖掘单味中药在数据集中的多种适用病机与各病机下的高关联配伍药物, 结合多药物对比, 最终实现药物配伍规律挖掘及同效药鉴别分析。为提升方法的准确性与实用性, 提出全集提升度指标与重复说者-听者标签传播算法 (speaker-listener label propagation algorithm, SLPA) 分别作为药物网络边权重与药物类别划分算法。结果 在实证研究中有效分析出《丁甘仁医案》中多种安神药的适用病机与核心配伍的差异; 全集提升度指标相比目前应用广泛的关联规则分析提升度指标, 修正了在计算高频药物关联度时的异常低值问题; 重复 SLPA 算法优化了 SLPA 算法存在的多项缺陷, 相比传统聚类算法, 不仅可实现药物的重叠类别划分, 还可导出药物在类中的重要性, 进而更容易分析出类的整体属性。结论 相比现有药物配伍规律挖掘方法, 该方法可更为准确、客观地挖掘出药物配伍所适用的病机, 具有更高的临床指导价值。全集提升度指标可成为提升度指标的有效替代指标, 重复 SLPA 算法可成为传统聚类算法的有效替代算法。

关键词: 中医药; 用药规律; 复杂网络; 关联规则; 标签传播; 全集提升度指标; 重复 SLPA 算法

中图分类号: R914; R283.21 文献标志码: A 文章编号: 1674-5515(2024)23-8101-10

DOI: 10.7501/j.issn.1674-5515.2024.23.017

Method of traditional Chinese medicine prescription compatibility mining and identification of same effect medicines from perspective of pathogenesis based on universal set lift index and repeated SLPA algorithm

ZHANG Di¹, NIU Qikai², ZHAO Yuwen², ZHANG Weina¹

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

2. Institute of Chinese Materia Medica, China Academy of Chinese Medical Sciences, Beijing 100700, China

Abstract: Objective Existing traditional Chinese medicine (TCM) prescription data mining methods generally have problems such as insufficient mining depth and poor applicability to TCM prescription data. To solve these problems, a new method of TCM prescription data mining is proposed. **Methods** This method is a complex network analysis method. It combines the idea of association rule analysis and cluster analysis. By mining a variety of applicable pathogenesis of single TCM in the data set and the highly associated compatible drugs under each pathogenesis, combined with multi-drug comparison, the drug compatibility rule mining and the identification analysis of the same effect drugs were finally realized. In order to improve the accuracy and practicability of the method, universal set lift index and repeated speaker-listener label propagation (SLPA) algorithm are proposed as the edge weight of the herb network and the herb cluster partition algorithm respectively. **Results** In the empirical study, this method successfully analyzed the different between applicable pathogenesis and core compatibility of various sedative herbs in *Ding Ganren Medical Case Records*. Compared with the currently widely used lift index, universal set lift index corrected the abnormally low value problem in calculating the association degree of high-frequency herbs. Repeated SLPA algorithm optimizes many defects of SLPA algorithm, compared with

收稿日期: 2024-06-21

基金项目: 国家重点研发计划 (2023YFC3502900)

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

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

traditional clustering algorithm, it can not only divide herbs into overlapping clusters, but also get the importance of each herb in the clusters. This makes it easier to analyze the overall characteristics of individual clusters. **Conclusion** Compared with the existing method, this method can more accurately and objectively find out the application conditions of herbal collocations, and has higher clinical guiding value. Universal set lift index can be an effective substitute index for lifting degree index. Repeated SLPA algorithm can be an effective alternative algorithm to traditional clustering algorithm.

Key words: traditional Chinese medicine; medication rule; complex network; association rules; label propagation; universal set lift index; repeated SLPA algorithm

中医药数据挖掘是从大量中医处方数据中提炼名医用药思想、发现中药新方、筛选疾病重要治疗药物的基本方法。但目前的中医处方数据挖掘研究多为单一分析方法的横向堆砌,缺少了递进整合多种分析方法的纵向深入,导致无法挖掘出处方数据更深层次的规律。此外,现有的主流方法多是直接借鉴其他领域的分析方法,少有对中医处方数据的针对性优化,这也导致了部分方法存在适用性差、挖掘效果不佳的情况。

在此背景下,本研究在传统关联规则分析的基础上,结合社区发现算法,提出了一种新的从病机角度分析中医药组方中药物配伍规律的方法,同时对原算法进行了改良。经实证研究验证,该方法能够挖掘出目标中药在处方数据集中的各种使用情境及各情境下的核心药物配伍,通过多药物对比分析,可辅助梳理名医处方中药物的鉴别选择与配伍规律,进

而为提升中医师临床遣方用药的能力提供帮助。

1 研究思路与框架

本研究主要借助复杂网络分析方法对处方数据进行挖掘,具体研究思路如下:①基于全部处方的数据集计算出待分析药物的高关联度药物。本步骤是为了筛选出可能与待分析药物共同对应某种相同病机的药物。②在数据集中抽取包含待分析药物的处方,形成新的数据集,基于新数据集计算高关联度药物间的关联度并构建网络。本步骤是为了挖掘高关联度药物间可能具有的共同病机对应关系,构建新数据集是为了尽可能避免待分析药物不适用于各高关联度药物间建立的共同病机对应关系的情况。③运用社区发现算法计算该网络的社区分布并进行分析。本步骤将基于各种病机对应关系构建的网络划分为多个社区,以此将各类病机以药物社区的形式析出。整体分析框架如图 1 所示。

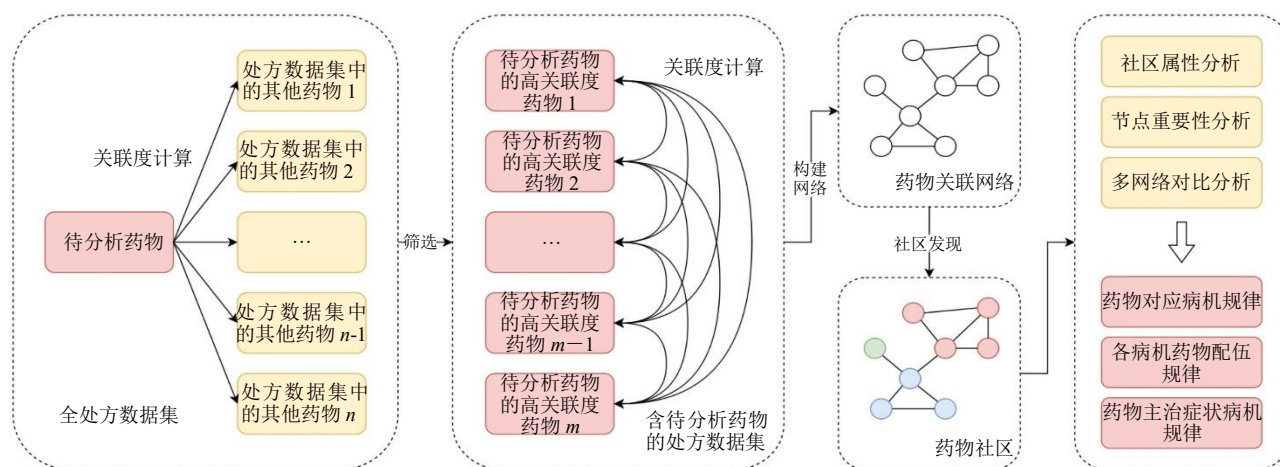


图 1 整体分析框架

Fig. 1 Overall analytical framework

其中,关联度计算基于关联规则分析中的提升度算法,社区发现基于说者-听者标签传播算法(speaker-listener label propagation algorithm, SLPA)^[1]算法,由于 2 种算法均有不足之处,本研究对其进行了改良。

2 算法设计

2.1 用药关联网络的构建

2.1.1 药物关联度与网络边权重的计算

(1) 计算方法选择:在以往的研究中,药物共现频次是衡量药物间关联度的常用指标并普遍被

作为用药关联网络的边权重。但药物共现并不等于药物相关，高共现关系往往只会存在于使用频率较高的药物，使用较少的药物则很难与其他药物具有较高的共现频次。而很多情况下，这种使用较少的药物才是针对某种病机的特异性药物。显然，以往的计算方式难以客观反映药物间的相关性，且会掩盖大量有价值的低频药物用药规律，并不能满足实际需求。因此，寻找一种更为客观、准确的药物间关联度指标，作为高关联药物的筛选标准与用药关联网络的边权重是十分必要的。

关联规则分析是一种广泛应用于中医药处方数据挖掘研究的方法，该方法中的提升度即是一种较好地用于共现数据的关联度指标，且效果优于余弦相似度与 Jaccard 系数等其余常用指标^[2]。该指标表示事件联合概率分布与边缘分布乘积的比值。给定事件 A 和事件 B ，其提升度的计算如公式 (1) 所示。

$$\text{Lift}(A, B) = P(A, B) / [P(A)P(B)] \quad (1)$$

对于中医药处方数据集，式中 $P(A)$ 与 $P(B)$ 分别表示数据集内的处方中含有药物 A 或药物 B 的概率，在实际计算中，其值等于数据集内包含药物 A 或药物 B 的处方数与数据集内的全部处方数之比。与之类似，式中 $P(A, B)$ 指数据集内的处方中同时含有药物 A 与药物 B 的概率。

(2) 基于改良提升度的全集提升度算法：在实际应用中，提升度算法对高频率事件的敏感度较低。当 A 和 B 两个事件的发生频率与共现频率均较高时，由于边缘概率的提升空间较小，会导致计算结果偏低，从而难以有效表达常用药物的配伍关系。目前尚未有研究对此进行改良，为了解决此问题，本研究引入了补集提升度 (complement lift, C-Lift) 与全集提升度 (universal set lift, U-Lift) 的概念。其基本思想是若 2 个事件具有较强的关联性，那么它们不仅同时发生的概率会提升，其同时不发生的概率也会提升，这种事件同时不发生的概率相对于其边缘概率的提升程度称为 C-Lift。若给定事件全集 U ，C-Lift 的计算可表示为公式 (2) 的形式。

$$\text{C-Lift}(A, B) = P(CuA, CuB) / [P(CuA)P(CuB)] \quad (2)$$

对于中医药处方数据集，式中 $P(CuA)$ 与 $P(CuB)$ 分别表示数据集内的处方中不含药物 A 或药物 B 的概率， $P(CuA, CuB)$ 则指数据集内的处方中既不含药物 A 也不含药物 B 的概率。

在此基础上，将 C-Lift 与提升度相乘即可得到

U-Lift，其计算如公式 (3) 所示。

$$\text{U-Lift}(A, B) = P(A, B)P(CuA, CuB) / [P(A)P(B)P(CuA)P(CuB)] \quad (3)$$

当两事件补集为空集或补集的交集为空集时，C-Lift 为 0 或无法计算，此时令 C-Lift 等于提升度，即令 U-Lift 等于提升度的平方。该处理是为了维持 U-Lift 对提升度强化作用的一致性。

(3) U-Lift 的评估：为了直观对比提升度、C-Lift 和 U-Lift，验证 U-Lift 的有效性，现对其取值进行可视化模拟分析。为了模拟所有可能的情况，设置如下策略：将两事件总频次固定为 40，只改变 1 个事件的频次，探究各种事件频次相异的情况；令两事件频次相同，同步变化二者频次，探究各种事件频次相同或相近的情况；令共现频次为两事件中低频事件的频次减 1，探究高共现的情况；令共现频次为 1，探究低共现的情况。对于所有模拟组，将事件总数固定为 50，频次变化范围设置为 10~30，但考虑到频次相同-低共现时，事件频次不可能大于 25，因此将频次相同-低共现组设置为 10~25。

将各策略按频次-共现策略相互组合，3 种提升度随事件频次的变化趋势对比如图 2 所示。可以看到，在频次相同-高共现时，提升度随事件频次的升高迅速下降至较低的水平，而事实上，该组的 2 个事件几乎完全共现且频次远未达到通过随机碰撞就能产生大量共现的数量级别，这显然是不符合预期效果，其真实关联度应远大于此值，而 U-Lift 则对此起到了明显的矫正作用。此外，还可以看出，除频次相同-高共现组外，其余组 U-Lift 与提升度的趋势均一致，在所有情况下，U-Lift 与提升度对相关性的正负判断也均一致，但在数值上表现出强化作用，具有更高的灵敏性。因此，在计算药物间关联度时，使用 U-Lift 可获得更准确的结果，U-Lift 是提升度的有效替代指标。

2.1.2 高关联度药物的筛选与用药关联网络的构建 使用 U-Lift 与共现频率联合筛选高关联度药物，当待计算药物与目标药物的 U-Lift 与共现频率均高于预设阈值时，该药物即被认为是目标药物的高关联度药物。筛选出高关联度药物后，将其构建为用药关联网络，若 2 个药物的 U-Lift 与共现频率均高于预设阈值，便为其建立 1 条边，边的权重为 2 个药物的 U-Lift。

2.1.3 基于中药功效的用药关联网络边权重优化方法 中药功效是药物配伍的最重要依据，因此对

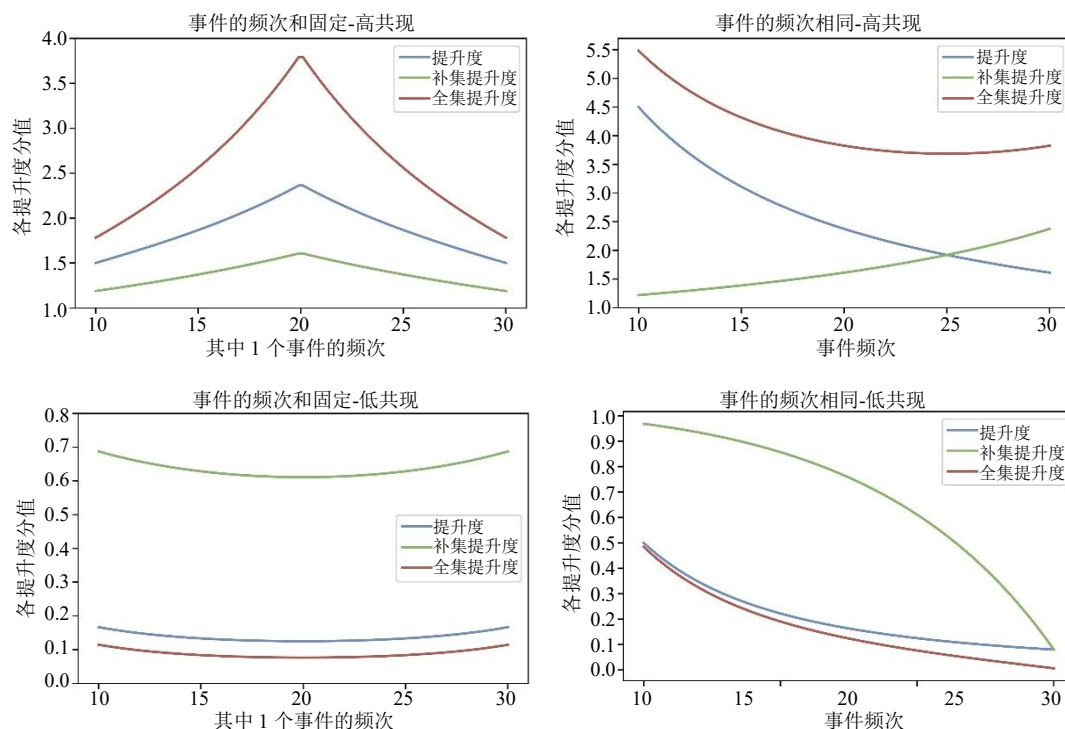


图 2 3 种提升度的趋势对比

Fig. 2 Trend comparison on three kinds of lifting degree

比 2 味药物的功效，便可在一定程度上获知二者在中医药理论层面的关联程度。据此，本研究提出一种基于中药功效的用药关联网络边权重优化方法。

首先，根据《中药学》^[3]的药物分类法构建中药功效树状图，该图共分为 2 层，分别对应《中药学》对药物的一级与二级分类。如在《中药学》中，解表药下分为发散风寒药与发散风热药，则在图中对应解表功效与发散风寒、发散风热功效，2 层功效为从属关系，若一级分类下无二级分类，则不设置第 2 层功效。然后参照中药功效树状图的最末层功效规范化数据集中各药物的功效，完成数据层面的准备工作。

在计算 2 味药物的功效相似性时，参照 Jaccard 系数的计算方式。Jaccard 系数表示 2 个集合交集与并集项数的比值，但在计算功效相似性时，由于不同的功效可能从属于同一更高层面的功效，所以不能简单的定义其是否为交集元素，故本研究扩展了 Jaccard 系数的计算方式，使其允许集合内的相似元素不完全相交。给定集合 A 与集合 B ，其 Jaccard 系数扩展后的计算方式如公式（4）所示。

$$\text{Jaccard}(A, B) = \frac{\sum_i^{n_A} s_{\max}(A_i|B) + \sum_j^{n_B} s_{\max}(B_j|A)}{2(n_A + n_B) - \sum_i^{n_A} s_{\max}(A_i|B) - \sum_j^{n_B} s_{\max}(B_j|A)} \quad (4)$$

式中， n_A 表示集合 A 的元素数； $s_{\max}(A_i|B)$ 表示

集合 A 中的第 i 个元素，与其在集合 B 中最相似元素的相似度。当相似度仅接受 0 或 1 作为结果时，该式的计算结果与原 Jaccard 系数计算方式的结果一致。

若 2 味中药的某 2 个功效完全相同，则令其相似度为 1；若功效不同，但从属于同一更高层面的功效，则令其相似度为 0.5。如生地黄规范后的功效为“清热凉血、清虚热、补阴”，天冬规范后的功效为“补阴、清虚热、清热泻火”，“清热凉血”与“清虚热”“清热泻火”同属“清热”功效，故相似度为 0.5，两药“清虚热”“补阴”功效完全相同，故相似度为 1。则二者的功效相似度计算过程可表示为 $[(0.5 + 1 + 1) + (1 + 1 + 0.5)] / [2 \times (3 + 3) - (0.5 + 1 + 1) - (1 + 1 + 0.5)] \approx 0.714$ 。

最后，将功效相似度映射至 1 个新的区间内，映射方法如公式（5）所示。

$$S_{\text{new}} = I_{\min} + S(I_{\max} - I_{\min}) \quad (5)$$

式中， $I_{\max}$ 与 $I_{\min}$ 分别对应区间的最小值与最大值。将映射后的功效相似度与原边权重相乘，即为优化后的边权重。经调试，将功效相似度映射区间的最小值与最大值分别设置为 0.5 与 2。如前所述，生地黄与天冬映射后的功效相似度约为 1.571，与原边权重相乘后可起到强化作用，而临床中二者确为较常联用的药物，故计算结果与临

床实际相符。

2.2 社区发现

2.2.1 社区发现方法选择 目前主流的社区发现算法可分为传统算法^[4]与基于深度学习^[5]的算法。而由于对研究目的、构建的药物关联网络特点和算法特性的考虑,本研究选择传统算法中的 SLPA 算法实现社区发现,与其他算法相比,该算法具有数据量需求低、无需标注、无需预先设定社区数量、可划分重叠社区、结果信息丰富、可解释性强等特点。

2.2.2 基于蒙特卡罗思想的改良策略 SLPA 算法在计算时存在较多的随机行为,这些行为会导致 SLPA 算法具有很强的不确定性,进而产生多种社区划分结果,干扰最终判断。此外,由于本研究需利用 SLPA 算法的诸多特性进行数据挖掘,故在改进时还需保留其原有算法结构。因此,直接筛选 SLPA 算法产生的社区是最简单有效的方法。

考虑到 SLPA 算法是一种信息动力学算法,可视为对节点间真实传播的模拟,本研究提出了一种重复 SLPA 算法(repeated speaker-listener label propagation algorithm, RSLPA),该算法基于蒙特卡罗思想对 SLPA 算法进行了改良。其具体计算步骤如下:①重复执行指定次数 SLPA 算法并将每次的节点储存器数据汇总;②提取每次执行得到的社区组成与社区中各节点对该社区的权重;③统计各社区出现的频次,计算各社区中节点对该社区权重的均值;④根据频率阈值提取有效社区;⑤根据频率降序排列各有效社区;⑥从首个有效社区开始,根据社区间最大相似度等规则依次纳入社区。

与 SLPA 算法通过节点标签权重判定节点归属社区不同,RSLPA 算法利用社区频率判定节点的归属社区,而将节点标签权重作为节点在社区中的重要性指标。因此 RSLPA 算法相比 SLPA 算法,在用于数据挖掘时可获得更丰富的社区信息,同时由于社区纳入规则的高度可定制化,RSLPA 算法在灵活性上也优于 SLPA 算法。

2.2.3 实证评价

(1)数据来源:选取 Zachary's Karate Club^[6]与 American College Football^[7]数据集对 RSLPA 算法进行评价,两者均为与药物关联网络规模大致相当的有标注的真实数据集且在社区发现算法评价中被广泛使用。在本研究中,前者采用 NetworkX 内置的数据集版本^[8],后者采用数据集创建者提供的版本^[9]。2 个数据集具有不同的规模与性质,可测试算

法在不同网络下的表现情况。

(2)算法稳定性评价:SLPA 算法最主要的缺点是稳定性较差,为对比 RSLPA 算法在此方面的提升,基于调整兰德指数(adjusted Rand index, ARI)^[10]对其稳定性进行评价。ARI 是目前应用最广泛的聚类性能评价指标之一,可反映 2 种类别划分的一致性,取值为-1 到 1,该值越大,则 2 种类别划分的一致性越强。ARI 的计算如公式(6)所示。

$$ARI = \frac{2(TP \cdot TN - FN \cdot FP)}{[(TP + FN)(FN + TN) + (TP + FP)(FP + TN)]} \quad (6)$$

TP 表示 2 个同类样本被正确分为同类的数量,FP 表示 2 个异类样本被错误分为同类的数量,TN 表示 2 个异类样本被正确分为异类的数量,FN 表示 2 个同类样本被错误分为异类的数量

共对算法进行 100 次测试,单次测试中,算法先后执行 2 次运算并计算 2 次结果的 ARI 值,最后取全部测试的平均 ARI 值作为稳定性评价结果。RSLPA 重复次数由 20~500 逐渐增加,100 次测试 2 种算法的平均 ARI 值如表 1 所示。

表 1 2 种算法在测试中的平均 ARI 值
Table 1 Average ARI of two algorithms in test

数据集	SLPA	RSLPA				
		20 次	50 次	100 次	200 次	500 次
Karate	0.607	0.849	0.882	0.899	0.911	0.922
Football	0.718	0.961	0.972	0.976	0.979	0.992

可以看到,RSLPA 算法的稳定性相对于 SLPA 算法有较大的提高,虽然其达到最大稳定性所需的重复次数会随网络规模的扩大迅速上升,但其社区划分结果在重复次数较低时就已趋近于稳定。

(3)社区划分质量评价:由于 Karate 与 Football 数据集均为有标注的真实数据集,本评价中将使用查准率(precision)、ARI 与模块度 3 种不同维度的指标对 RSLPA 的社区划分质量进行评价。其中,查准率表示划分在同一社区的节点,实际也是在同一社区的比例,在分析药物社区时,更细致的社区划分并不会影响人工分析,但若社区中混入不同类药物,则会严重干扰最终结果,因此该指标是影响本研究方法效果的关键指标,其计算公式如式(7)所示。模块度的计算公式如式(8)所示。

$$P = TP / (TP + FP) \quad (7)$$

$$Q = \frac{1}{2m} \sum_{i,j} (A_{ij} - \frac{k_i k_j}{2m}) \delta(c_i, c_j) \quad (8)$$

m 为网络中的总边数, A 为网络的邻接矩阵, k 为节点的度,

当节点 i 和 j 属于同一社区时, δ 函数输出为 1, 否则为 0

由于 Karate 数据集为加权网络, 故应采用模块度的加权形式^[11], 其计算公式如式 (9) 所示。

$$Q=\frac{1}{2w}\sum_{i,j}(W_{ij}-\frac{s_i s_j}{2w})\delta(c_i,c_j) \tag{9}$$

其中, w 为网络中边的总权重, W 为网络的加权邻接矩阵, s 为节点所有边的权重和, 当节点 i 和

j 属于同一社区时, δ 函数输出为 1, 否则为 0。对于无权网络, 公式 (9) 与公式 (8) 等值。

本评价共对算法进行 100 次测试, 同时记录算法在测试过程中各指标的平均值、最高值与最低值。RSLPA 重复次数由 20~500 逐渐增加, 100 次测试 2 种算法的评价指标对比如表 2 所示。

表 2 2 种算法的评价指标对比

Table 2 Comparison on evaluation indicators of two algorithms

数据集	指标	SLPA	RSLPA				
			20 次	50 次	100 次	200 次	500 次
Karate	P_{avg}	0.977	1.000	1.000	1.000	1.000	1.000
	P_{max}	1.000	1.000	1.000	1.000	1.000	1.000
	P_{min}	0.930	1.000	1.000	1.000	1.000	1.000
	ARI_{avg}	0.654	0.682	0.685	0.684	0.680	0.682
	ARI_{max}	0.785	0.741	0.741	0.741	0.741	0.741
	ARI_{min}	0.560	0.632	0.632	0.632	0.676	0.676
	Q_{avg}	0.501	0.504	0.505	0.505	0.506	0.505
	Q_{max}	0.508	0.508	0.508	0.508	0.507	0.507
	Q_{min}	0.474	0.494	0.494	0.494	0.494	0.494
Football	P_{avg}	0.719	0.858	0.862	0.864	0.866	0.872
	P_{max}	0.902	0.899	0.885	0.885	0.885	0.885
	P_{min}	0.318	0.813	0.813	0.813	0.813	0.813
	ARI_{avg}	0.762	0.872	0.876	0.877	0.878	0.882
	ARI_{max}	0.894	0.900	0.889	0.889	0.889	0.889
	ARI_{min}	0.400	0.808	0.847	0.847	0.847	0.847
	Q_{avg}	0.587	0.601	0.602	0.602	0.602	0.601
	Q_{max}	0.605	0.603	0.603	0.603	0.603	0.603
	Q_{min}	0.533	0.581	0.601	0.601	0.601	0.601

P 、 ARI 与 Q 分别代表算法对数据集划分结果的查准率、调整兰德指数与模块度, 其中角标 avg 、 max 、 min 分别代表 100 次测试中各指标的平均值、最大值与最小值。

P , ARI and Q respectively represent accuracy rate, adjusted Rand index and modularity of results of data set partitioning by algorithm, where subscripts avg , max and min respectively represent average, maximum and minimum values of each index in 100 tests.

测试结果显示, 所有重复次数 RSLPA 算法的各指标均值均优于 SLPA 算法, 各指标最低值显著优于 SLPA 算法, 各指标最高值与 SLPA 算法接近。该结果说明, RSLPA 算法可稳定筛选出 SLPA 算法结果中的最优或次优社区划分, 剔除不合理的社区划分, 整体而言, 其社区划分质量优于 SLPA 算法。

综上, RSLPA 算法在各项评价中的整体表现均优于 SLPA 算法, 可有效筛选 SLPA 社区划分结果中的高质量社区。同时, RSLPA 算法还保留并加强了 SLPA 社区划分结果多样、算法规则灵活、社区信息丰富的特性, 是特别适用于数据挖掘的算法。

3 实证研究

丁甘仁是著名的中医临床家与教育家, 是中医

近现代教育的先驱者与孟河医派的代表人物之一, 尤擅治疗急症热病。神志异常是温病热入营分的典型症状, 代表邪热内陷, 病情开始陷入危重, 是较有研究意义的症状, 而安神药则是针对神志异常最常使用的药物类型之一, 有安神定志之功, 可用治心神不宁诸证。故本研究在实证中对丁甘仁使用安神类中药的用药规律进行挖掘, 既可验证本研究提出方法的有效性, 又能得出对临床有一定指导和借鉴意义的挖掘结果。

3.1 数据处理与计算

3.1.1 数据处理 丁甘仁处方数据提取自《丁甘仁医案》^[12], 经过纳排标准与规范后的数据集中共包含 358 首处方, 413 种药物, 11 144 对药物共现关系。

3.1.2 纳入和排除标准 由于同一病例的各诊次处方高度相似,故只纳入首诊处方,要求处方完整,不限制处方剂型;疗效不佳的处方不予纳入。

3.1.3 数据规范 药物名称依据《中国药典》2022 年版规范,《中国药典》未收录的药物,依据《中药学》^[3]、《中华本草》^[13]、《全国中草药汇编》^[14]等进行规范。药物经炮制后性味显著改变,保留其炮制名,如荆芥炭;性味与原药物仍相似的取其本名,如“炒白术”改为“白术”。对于处方中包含的成药,由于内部药物用量极小,医家主观上也将其视为一个整体,故保留其本名。

3.1.4 药物筛选 依据《中药学》中的药物分类,丁甘仁处方数据集中各安神药及各安神药的使用频次为茯神 76 次、远志 51 次、酸枣仁 17 次、龙骨 17 次、龙齿 16 次。其中,龙齿为龙骨的附药,二者来源与成分相似,但在《丁甘仁医案》中用法不

同,故作为不同药物分别分析。

3.1.5 算法参数设置 设置 U-Lift 参数,考虑到药物频次差异,兼顾可信度与网络规模,设置用药关联网络的边最低关联度阈值为 1.5;设置茯神、远志网络的边 2 侧节点最低共现频率为 0.015,酸枣仁、龙骨、龙齿网络的边 2 侧节点最低共现频率为 0.01;强化边权重,设置权重值为关联度的 3 次方;基于药物功效对边权重进行优化。设置 RSLPA 算法的重复次数为 500,内层 SLPA 算法的迭代次数为 20,最低社区频率阈值为 0.3,最小社区长度为 2,不设置社区间最大相似度,允许社区嵌套。

3.2 结果与数据分析

3.2.1 实验结果 5 种目标药物的高关联药物网络输出如图 3 所示,其中,节点大小表示节点的度,边的粗细表示节点间的关联度。导出高关联药物社区信息,结果如表 3 所示。

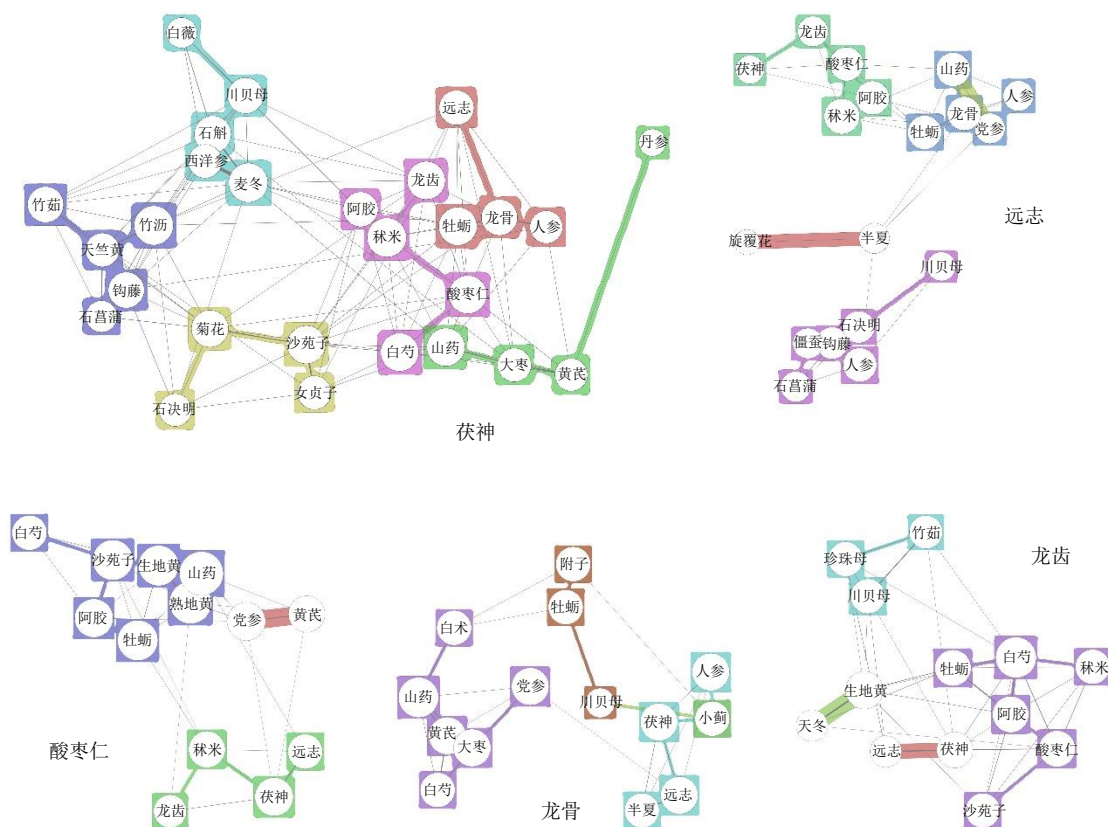


图 3 各目标药物的高关联药物网络
Fig. 3 High correlation drug networks of target drugs

3.2.2 数据分析 表 3 列举了 5 种目标安神药的高关联药物社区划分情况。对于所属同一社区的药物,在医案内往往联用并代表了某种特定的病机,对于不同社区内的药物,在医案内也有联用,但不

具备较稳定的联用关系。表中还列出了各药物在所属社区内的重要性,该值即各药物在算法模拟传播过程中所积累的社区权重,重要性越高说明其与社区的相关程度越高,越能代表社区的整体属性,因

表3 高关联药物社区信息

Table 3 Community information on highly associated drugs

目标药物	高关联药物社区（各药物在所属社区内的重要性）	社区频次
茯神	龙骨（5 629）、人参（4 267）、牡蛎（3 346）、远志（471）	500
	大枣（4 706）、黄芪（3 333）、山药（2 102）、丹参（82）	476
	西洋参（16 920）、麦冬（12 868）、石斛（11 681）、白薇（1 329）、川贝母（853）	405
	天竺黄（6 369）、竹沥（5 049）、石菖蒲（3 430）、竹茹（964）、钩藤（829）	383
	沙苑子（5 959）、女贞子（3 693）、菊花（2 383）、石决明（933）	297
	酸枣仁（1 583）、秫米（836）、白芍（791）、龙齿（765）、阿胶（608）	270
远志	酸枣仁（1 646）、龙齿（956）、秫米（789）、阿胶（783）、茯神（351）	500
	旋覆花（186）、半夏（186）	500
	钩藤（7 119）、僵蚕（4 566）、石决明（3 368）、竹沥（2 764）、石菖蒲（1 127）、川贝母（326）	500
	龙骨（1 394）、牡蛎（1 104）、人参（710）、山药（699）、党参（356）	500
	党参（24）、山药（16）	270
酸枣仁	茯神（1 018）、远志（793）、秫米（231）、龙齿（228）	500
	党参（697）、黄芪（697）	500
	生地黄（2 966）、山药（2 949）、熟地黄（2 936）、牡蛎（1 028）、沙苑子（866）、阿胶（551）、白芍（243）	500
龙骨	远志（1 308）、半夏（1 270）、茯神（877）、人参（388）、小蓟（73）	500
	大枣（2 735）、黄芪（2 666）、山药（1 885）、白芍（1 489）、党参（697）、白术（446）	500
	牡蛎（425）、川贝母（301）、附子（128）	284
	牡蛎（427）、川贝母（301）、附子（133）、小蓟（4）	216
龙齿	竹茹（674）、川贝母（570）、珍珠母（380）	500
	天冬（774）、生地黄（774）	500
	阿胶（1 133）、白芍（1 034）、牡蛎（737）、秫米（685）、酸枣仁（617）、沙苑子（437）	449
	茯神（575）、远志（575）	324

此在分析时，可着重分析重要性高的药物。社区频次反映了各社区划分的可信度，当2个社区高度重叠时，高频次的社区划分往往更加准确，但低频社区的结果也可作为参考。

（1）各安神药的适用病机存在显著差异：根据各社区内药物，可概括出《丁甘仁医案》中应用各安神药时的主要病机如下。

①茯神可用于心神不宁的各种病机：茯神共导出6个无明显交集的药物社区，社区内药物功效依次为收敛固涩、补气、养阴清热、息风清热化痰、养肝平肝、安神补血，其主治几乎可涵盖心神不宁的所有病机。因此对心神不宁的患者，一般可直接选用茯神。

②远志多用于痰证与滑脱证：远志导出的5个药物社区中，除以安神药为主的1个社区外，有2个社区与痰有关，其内药物主治分属寒痰、风痰，因此见寒痰、风痰病机的患者，可考虑选用远志。剩余2个社区完全重叠，以收敛固涩与益气药为主，说明气虚等导致的滑脱诸证，见心神不宁者也可选

用远志。

③酸枣仁多用于虚证：除安神药社区外，酸枣仁还导出2个社区，病机分别对应气虚与阴虚血少精亏。因此对心神不宁伴气血亏虚、肝肾不足等的患者，可考虑选用酸枣仁。但对于阳虚患者，酸枣仁则并非首选药物。

④龙骨可用于滑脱不禁诸证，用治心神不宁时多配伍化痰药：龙骨共导出4个社区，去除高度重叠社区后余3个社区，药物功效分别以化痰安神、补中益气、敛阴潜阳为主。由于龙骨具有收敛固涩的功效，故气虚、阴虚、阳虚等导致的滑脱皆可选用龙骨。化痰安神社区中包含了与龙骨关联密切的全部安神药，且其内部药性偏温，说明龙骨治疗心神不宁患者时，患者多属阳虚夹痰，这提示对于阳虚滑脱的患者，若出现心神不宁的表现，要考虑配伍化痰之品。

⑤龙齿多用于痰热扰神与阴虚阳浮证：龙齿共导出4个社区，功效分别以化痰定惊、滋阴清热、养血、安神为主。因此，对于痰热扰神、阴血亏虚

不能摄阳等致心神不宁者，可选用龙齿。对比龙齿与龙骨的结果可以发现，二者虽然功效相似，但丁甘仁在应用二者时，情境完全不同。整体而言，龙骨更偏于重证、虚证，阴阳皆可选用；龙齿则多用于实证或虚证轻者，且以阴虚证为主。

在临床中，可根据患者的实际病机选用相应的安神药，并参考安神药各病机社区内的药物进行配伍。

(2) 安神药需根据病机差异配伍不同的补气药：表 4 除反映基本的各安神药适用病机及各病机下的核心配伍药物外，若对核心配伍药物进行分析，还可得到应用安神药时选择配伍药物的一般规律。如各安神药社区内均存在补气药，但各药常配伍的补气药却存在差异，此时便可分析此差异，得到丁甘仁在应用安神药时配伍补气药的一般规律，补气药包括人参、黄芪、党参、山药、西洋参、大枣等药。其中，人参多与龙骨、牡蛎等收敛固涩药位于同一社区，说明其多用于危重证；酸枣仁补气社区中仅有党参、黄芪，说明二药善治气虚伴血亏证；山药、西洋参均有补阴功效，但西洋参所处社区内药物药性与山药相比更偏于寒凉，说明山药一般用于虚热不甚或无热者，西洋参则一般用于虚热明显或热盛伤阴者；大枣仅存在于茯神、龙骨二药社区内，说明其适用病机应体现在二药与其余安神药的差异性上，即与阳虚或脱证有关。

(3) 药物组方配伍规律：既往的数据挖掘研究由于缺乏对数据挖掘方法的递进整合，并不能在算法层面实现单味药物病机与药物配伍的联合挖掘，只能通过传统的直接筛选、人工整理等非技术手段辅助其间接实现。本研究结合了改良的关联规则分析与社区发现算法，解决了上述问题，较之既往研究，具有以下优势。

①算法层面：提出的 U-Lift 指标较之既往研究^[15-17]中应用的提升度指标准确性更高。如半夏与远志皆为温化寒痰药，临床中常相须为用。在实证中对龙骨进行分析时，半夏在含龙骨的 17 个处方中出现 9 次，远志出现 11 次，二者共现 9 次，属于高频高共现，但由于传统的提升度指标对高频事件不敏感，故二者的提升度仅约为 1.545，不符合实际情况，而 U-Lift 约为 3.284，对此起到了修正作用。同时，本研究提出的 RSLPA 算法与传统聚类方法^[18-20]相比，可输出划分结果中各药物的重要性，避免对结果的错误概括。实证研究中茯神的补

气社区内含有活血药丹参，传统方法无法判断此药重要性，会将类的整体功效概括为补气活血，但本研究导出社区中提示丹参的重要性远低于其他药物，故分析时可略去，因此可获得更准确的结果。

②对单味药的分析层面：既往研究^[21-23]普遍采取直接分析含目标中药处方的方式。这种分析方式默认处方内的全部药物均与目标药物所针对的病机一致，处方内所总结的病机即是目标中药的适用病机，这显然是不科学的。如酸枣仁作为偏补益性质的安神药，实证研究中若采取此方式分析，则会混入陈皮、附子等与其功效大相径庭的药物，得出酸枣仁适用于气滞、痰证、阳虚等病机的错误结论。而即使只对高频药物进行分析，也会混入甘草、半夏等本就常用的药物，干预最终结果。本研究首先用 U-Lift 对药物进行初步筛选，可避免后续分析时混入不相干的药物，进而得到更准确的分析结果。

③对病机的提取层面：既往研究^[24-26]多采用人工整理或提取原处方记录的方式，将处方病机或治法总结为 4 字证型，再进行后续分析。这种方式虽然符合中医学的记录传统，也易于统计，但 4 字证型或治法往往是处方的整体概括，信息缺失较多，不利于后续分析。此外，处方的病机或治法可能并不能代表其内药物的病机或治法，如归脾汤可治心脾两虚及脾不统血证，但其内木香却主治中焦气滞证，若将前者主治强加于后者，分析结果则难免存在误差。本研究采用社区发现算法，将主治病机相近的药物自然聚为一类，通过以药测证间接分析药物的适用病机，避免了上述问题，结果更为客观。

4 结论

本研究创新应用 U-Lift 与重复 SLPA 算法，提出了一种基于病机的中医药组方配伍规律挖掘及药物鉴别方法。该方法在真实数据集中可清晰呈现多个目标药物间的差异性与各病机下的核心配伍药物信息，经实证评价，数据挖掘结果比传统研究方法更为客观准确，对临床中药物选择、药物配伍、病机诊断等均有较好的应用价值，也可作为现代药理学研究中研究对象的筛选方法。

利益冲突 所有作者均声明不存在利益冲突

参考文献

- [1] 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.
- [2] van Eck N J, Waltman L. How to normalize cooccurrence data? An analysis of some well-known similarity measures [J]. *J Am Soc Inf Sci*, 2009, 60(8): 1635-1651.
- [3] 钟赣生, 杨柏灿. 中药学 [M]. 第 5 版. 北京: 中国中医药出版社, 2021.
- [4] Fortunato S. Community detection in graphs [EB/OL]. arXiv (2010-02-01) [2024-04-01]. <http://0906.0612>. <http://arxiv.org/abs/0906.0612>
- [5] 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.
- [6] Zachary W W. An information flow model for conflict and fission in small groups [J]. *J Anthropol Res*, 1977, 33(4): 452-473.
- [7] Girvan M, Newman M E J. Community structure in social and biological networks [J]. *Proc Natl Acad Sci USA*, 2002, 99(12): 7821-7826.
- [8] NetworkX Developers. Karate club — NetworkX 3.2.1 documentation [EB/OL]. (2023-10-28) [2024-04-04]. https://networkx.org/documentation/stable/auto_examples/graph/plot_karate_club.html.
- [9] Newman M E J. Network data [EB/OL]. (2013-04-19) [2024-04-04]. <https://public.websites.umich.edu/~mejn/netdata/>.
- [10] Hubert L, Arabie P. Comparing partitions [J]. *J Classif*, 1985, 2(1): 193-218.
- [11] NetworkX Developers. Modularity-NetworkX 3.2.1 documentation [EB/OL]. (2023-10-28) [2024-04-04]. <https://networkx.org/documentation/stable/reference/algorithms/generated/networkx.algorithms.community.quality.modularity.html>.
- [12] 丁甘仁. 苏礼, 王怡, 谢晓丽整理. 丁甘仁医案 [M]. 北京: 人民卫生出版社, 2007.
- [13] 国家中医药管理局《中华本草》编委会. 中华本草 [M]. 上海: 上海科学技术出版社, 1999.
- [14] 《全国中草药汇编》编写组. 全国中草药汇编 [M]. 北京: 人民卫生出版社, 1976.
- [15] 刘兴华, 林禧, 方梦冰, 等. 基于多元数据挖掘分析彭胜权治疗岭南地区慢性乙型病毒性肝炎的用药规律 [J]. *时珍国医国药*, 2024, 35(5): 1235-1240.
- [16] 徐玮祯, 王思敏, 方彩珊, 等. 基于数据挖掘方法分析阮岩治疗儿童变应性鼻炎的用药规律 [J]. *中药新药与临床药理*, 2024, 35(6): 896-903.
- [17] 翟桐, 杜佳璐, 任雪阳, 等. 基于 Apriori 和 MIE 算法分析关节炎中药制剂用药规律及核心药对的潜在作用机制 [J]. *中草药*, 2024, 55(9): 3027-3040.
- [18] 石淇允, 许晓彤, 卢敏, 等. 关联规则的中医药防治骨质疏松数据库构建及聚类有效性研究 [J]. *中国骨质疏松杂志*, 2024, 30(4): 474-481.
- [19] 林蓉, 高旅, 肖亚楠, 等. 基于国家专利的中药治疗癫痫组方规律数据挖掘研究 [J]. *中国中医药信息杂志*, 2024, 31(7): 44-49.
- [20] 李芳琳, 王淑敏, 刘晓, 等. 基于数据挖掘的中医药治疗儿童腺样体肥大用药规律及特点 [J]. *世界中医药*, 2024, 19(6): 844-849.
- [21] 耿蓓蕊, 其乐木格, 高洁, 等. 含羌活治疗目疾方剂配伍规律研究 [J]. *中成药*, 2024, 46(4): 1379-1384.
- [22] 梁绿圆, 曹佳蕾, 刘宜杭, 等. 基于“病-药-量”探讨《中华医典》含丹参方剂用药规律的数据挖掘 [J]. *中草药*, 2024, 55(4): 1268-1280.
- [23] 豆改杰, 索南拉毛, 刘盈盈, 等. 基于数据挖掘的藏医药古籍文献中水银(佐太)解毒组方规律探讨 [J]. *世界科学技术—中医药现代化*, 2024, 26(3): 777-786.
- [24] 郭晓雪, 王萍, 李敬华, 等. 基于数据挖掘的慢性萎缩性胃炎中医证治规律分析 [J]. *中国中医药信息杂志*, 2024, 31(4): 38-43.
- [25] 王卉, 丁焕章, 杨勤军, 等. 基于证候-药物的新安医家治疗咳嗽医案数据挖掘研究 [J]. *中国中医药信息杂志*, 2024, 31(3): 34-41.
- [26] 武上雯, 李恒, 郭洪涛, 等. 基于真实世界数据挖掘中医药治疗不同证型系统性红斑狼疮肺间质病变的用药规律研究 [J]. *中国中医基础医学杂志*, 2024, 30(1): 88-94.

[责任编辑 潘明佳]