doi: 10.3969/j.issn.0490-6756.2018.06.009

利用稀疏表达学习挖掘中医方剂功效配伍

张思原[1,4], 刘兴隆[2], 姚 攀[1], 于中华[1], 陈 黎[1], 廖 强[3]
(1.四川大学计算机学院, 成都 610065; 2.成都中医药大学, 成都 610075;
3.四川大学外国语学院, 成都 610065; 4. 中国核动力研究设计院核反应堆系统设计技术重点实验室, 成都 610213)

摘 要: 中医方剂是中医药学的重要组成部分,也是中医临床治病的主要形式和手段.为了"辨证论治",需要从配伍功效出发,研究药组的配伍规则.多味药组成的方剂的功效不是其组成药物功效的简单叠加,而是由它们之间相互作用的结果.目前利用数据挖掘技术挖掘研究方剂的配伍,主要利用方剂中药物的频率,进行浅层分析,但这种方法并不能很好的揭示药物之间的相互联系.为此,本文提出了一种利用稀疏表达学习,自动挖掘古方中的功效配伍规律.稀疏表达学习结合 L1 正则化和逻辑斯蒂判别式,将不起作用或作用很小的药物视为是噪声过滤掉,起主导作用的药物则为被挖掘的功效配伍药组.最后,将提出的方法在 14 种功效的古方数据集中进行实验和验证,并以 Dice 系数和平均查准率作为评估参数,实验结果证明,稀疏表达学习方法相比目前的主流方法在配伍规则的挖掘上更准确、有效.

关键词: 功效配伍;方剂;稀疏表达学习;逻辑斯蒂;L1 正则化
中图分类号: TP391 **文献标识码**: A **文章编号**: 0490-6756(2018)06-1180-09

Utilizingsparse representation learning to mine oriented-efficacy compatibility in traditional chinese medicine prescriptions

ZHANG Si-Yuan[1,4], *LIU Xing-Long*[2], *YAO Pan*[1],
YU Zhong-Hua[1], *CHEN Li*[1], *LIAO Qiang*[3]
(1. College of Computer Science, Sichuan University, Chengdu 610065, China;
2. Chengdu University of Traditional Chinese Medicine, Chengdu 610075, China;
3. College of Foreign Languages and Cultures, Sichuan University, Chengdu 610065, China; 4. Science and Technology on Reactor System Design Technology Laboratory Nuclear Power Institute of China, Chengdu 610213, China)

Abstract: Traditional Chinese medicine (TCM) prescription, as an important part of TCM theory, is one of the main manifestation forms and ways of clinical treatment. We need to study oriented-efficacy compatibility for treatment based on syndrome differentiation. A prescription is composed of several or a dozen drugs, which efficacies are not simply the composition of all individual effect. In fact, its efficacies are the results of the interactions among drugs inside the prescription. At present, most researches focus on exploiting the frequencies of drugs in prescriptions by utilizing data mining technologies, which cannot catch the interactions among drugs. Therefore, this paper proposes a novel algorithm utilizing sparse representation learning to mine oriented-efficacy compatibility in TCM ancient prescriptions, which takes low weight drugs as noise and makes up an oriented-efficacy drug group with high weight drugs.

收稿日期: 2018-05-21
基金项目: 四川省科技支撑项目(2014GZ0063); 四川省重点研发项目(2018GZ0182)
作者简介: 张思原(1992-), 女, 四川宜宾人, 硕士生, 研究方向为自然语言处理、医学信息学.
通讯作者: 陈黎. E-mail: cl@scu.edu.cn

We combine the logistics and L1-norm based regularization to mine the oriented-efficacy compatibility. Lastly, 14 prescription datasets with different efficacies are used to validate our approach as well as dice index and the average retrieval rate are taken as metrics. Experimental results show that our approach is more effective and accurate than those of the state-of-the-art research.

Keywords: Oriented-efficacy compatibility; Prescription; Sparse representation learning; Logistics; L1-regularization

1　引　言

中医是我国各代名家智慧和经验总结提炼出来的宝贵财富，中医方剂是中医药学的重要组成部分，也是中医临床治病的主要形式和手段. 几千年来，中医药领域的无数临床实践与理论研究积累了海量的中医方剂. 中医方剂是中医药学“理-法-方-药”的一个重要组成部分，其配伍规律有着深刻的科学内涵[1]. 目前中医的配伍研究主要是从“辨病论治”的目的出发，收集治疗某一疾病的大量方剂进行配伍的研究，从而获取治疗某疾病的配伍规律，这些配伍规律对于临床治疗是有着重要的作用.

而方剂学是为了阐明方剂与病证之间治法的关系，揭示构成方剂的诸要素与功效之间的关系. 方剂学的研究范围主要是以古人经典方剂的制方原理为主线，研究方中药物配伍的主次关系和功效与主治病证病机相关的配伍原理. 因此，方剂学的研究者“辨证论治”，研究面向功效的配伍规则，这些规则对于方剂学的理论研究以及方剂的规范化工作的推进都有重要意义.

一首方剂往往由几味至十几味药物组成，其中每一味药都有自身的特定功效集合，但是方剂功效不是其组成药物功效的简单叠加[2]，而是通过方剂配伍理论中“君臣佐使”、“七情”的相互作用下产生. 同时一个方剂可以具有多种功效，但通常只有部分药组对特定功效起作用，如表 1 是一首中医名方的中药组成和功效.

表 1　方剂示例

Tab. 1　An example of a prescription

方剂	五苓散
中药	茯苓 180g、泽泻 300g、猪苓 180g、肉桂 120g、白术(炒)180g
方剂功效	利水渗湿，温阳化气

表 1 中，五苓散具有利水渗湿，温阳化气两种功效，其中利水渗湿功效主要由泽泻、茯苓、猪苓、白术四种中药共同作用形成的，我们把这四味药称为具有利水渗湿功效的配伍药组. 面向功效的配伍规则挖掘就是需要从大量的古方中获取某个功效的药组.

目前中医方剂配伍规律的理论研究取得了不少成果，对中医药学和方剂学都起到了巨大的推动作用，主要围绕药物配伍规律、病症-复方、核心药组[4]以及配伍禁忌等任务展开. 这些配伍规则的挖掘的方法可以应用在面向功效的配伍任务中，但是目前挖掘的方法主要都是利用了词频以及药自身的属性. 利用词频的方法，如关联规则，主要从方剂中共现的中药进行浅层的显式统计，对一些高频出现的调和药，如甘草，会被误认为是配伍中的药物，同时方剂的功效不是药物功效的叠加，而是药物之间相互作用下产生的，目前的功效配伍研究无法捕捉到药物之间的相互联系. 利用中药自身属性的方法，如聚类，会出现挖掘出的面向功效的药组都具有相同的功效，而在功效配伍中存在一些药自身不具有某功效，而是经过配伍以后和其他药一起相互作用而具有某功效的. 对于大量具有某个特定功效的方剂只有部分药物对这个功效起作用，其他药物对这个特定功效的贡献度很小或者不起作用. 为此，本文提出了一种利用稀疏表达学习的方法去自动挖掘古方中的面向功效的配伍规律. 利用稀疏表达学习将不起作用或作用很小的药物看作是噪声过滤掉，权重大的药物就是要被挖掘的功效配伍药组. 本文采用 L_1 正则化和逻辑斯蒂判别式结合的方式实现稀疏表达过程，使用坐标轴下降法训练模型并对训练好的模型参数进行降序排序同时解码成对应药物，通过取前 n 味药得到特定功效下药味数为 n 的一个配伍药组.

本文后续内容组织如下：第二节介绍相关工作，分析当前研究的优缺点；第三节对研究任务进行了阐述；第四节介绍了研究所使用的数据集；第五节详细阐述了利用稀疏表达学习挖掘功效配伍的算法及过程；第六节介绍实验方法和实验结果，并对实验结果进行了详细的分析；最后，第七节对全文的工作进行了总结，提出了工作的后续改进方向.

2 相关工作

目前方剂配伍的研究主要使用四类方法:频数统计[5,6]、关联规则[1,7-10]、复杂网络[11-14]和聚类方法进行配伍挖掘[15,16]. 频数分析采用频数统计对中医配伍规律进行研究分析,对方剂中药物共现的次数、药物类别等进行统计,然后取频次较高的药组或药物根据中医理论进行分析. 这类方法适用于发现显性经验,不足之处在于结果一般取高频药组,而中医中甘草、牛膝等作为使药(调节作用,调和诸药)常出现在方剂中与其他药物搭配会被误认为起主要作用,但是有时它们也可能是某功效(如活血祛瘀)或疾病的核心药组的组成,因此完全去除这些药物进行统计又会造成重要信息损失. 目前使用较多的是关联规则[7,10]的方法,如文献[7]发现肺纤维化中药方剂中的核心药物组合;文献[8]挖掘外用生肌中药处方的配伍规律;文献[9]利用正负双支持度的关联规则挖掘算法探究古方中的全局配伍规律等. 关联规则的方法需要设置最小支持度或置信度等阈值,若设置过小则会产生大量的频繁模式不宜于人工分析,若设定过大则可能漏掉大量有意义的关联规则. 同时这种方法也存在不能有效过滤甘草、牛膝等使药的缺点. 复杂网络的方法主要从构建中药药性网络、中药功效网络、中药方剂网络的角度对配伍规律进行研究[11-14]. 文献[14]考虑了药物的性味归经、药物功效等属性,又结合了方剂配伍共现共同计算药物之间的相似性并构造复杂网络,采用随机游走和标签传播的方法发现网络中具有一定联系的药组. 文章主要研究方剂中的常用药对,但是并没有对整个药组的配伍功效进一步研究说明. 文献[15,16]主要采用聚类的方法将药物两两进行分析,最后得到核心药组. 而在配伍中是由多个药组的共同作用,并不一定是两两药对进行作用.

综上所述,目前对方剂的配伍研究并未考虑到药物之间的相互关系,面向功效的配伍研究很少. 为此本文提出了一种基于稀疏表达学习的方剂功效配伍挖掘算法,能够有效的发现方剂中面向功效的配伍规律.

3 任务的描述

功效配伍挖掘是中医配伍规律研究的一个重要任务,目的是通过对大量方剂的分析,找出特定功效的配伍药组. 本文从大量已知功效的方剂中挖掘出具有特定功效的配伍药组,如图 1 所示. 算法的输入是一组具有某种特定功效的方剂(方剂中的中药,不考虑剂量)和方剂的功效,通过稀疏表达学习算法得到具有特定功效的配伍药组.

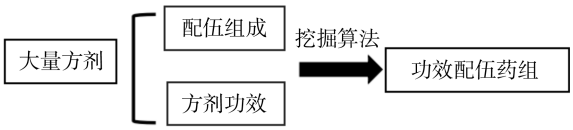

图 1 功效配伍挖掘任务

Fig. 1 The task of oriented-efficacy compatibility mining

4 数据集

本文从中医中药网(https://www.zhzyw.com/),包括《圣济总录》、《本草纲目》、《中国药典》等中医名书记录的名方、古方收集了 51,083 剂古方. 根据方剂中的主治信息对方剂功效进行自动标注,由于未进行方剂功效名的规范化处理,为此我们利用规则进行了功效名的简单规范化处理,如功效"强筋骨"和"壮筋骨"可以统一为"壮筋骨","活血化瘀"和"活血祛瘀"合并为活血化瘀.

本文的研究未考虑药剂量对配伍的影响,因此在预处理过程中,我们去掉了中药的剂量,并对中药的药名进行了规范化处理. 把同物异名的中药统一使用规范化的中药名称. 最后我们选择其中 14 种功效方剂构成 14 个数据集进行模型的学习及验证,数据集情况,如表 2 所示.

5 稀疏表达学习的方剂功效配伍挖掘算法

目前中医方剂配伍规律研究主要使用的关联规则是对同一症状或疾病的方剂根据药物共现频率进行统计分析. 而在方剂中频繁出现的甘草在大

表 2 数据集分布情况

Tab. 2 Distribution of dataset

功效	活血化瘀	化痰	壮筋骨	止血	消肿	利湿	明目	健脾	止咳	安神	活血止痛	止痒	通络	清热解毒
方剂数	362	1530	387	518	941	474	635	1245	419	566	164	348	355	820
中药数	355	687	347	359	549	375	413	596	322	396	209	314	327	584

多数情况下并不是方剂中的核心药物,仅仅作为使药起到调和诸药的作用,但是有时候这些药在方剂中对某些功效(如清热解毒,活血化瘀等)也会起主要作用,针对这种情况关联规则无法有效地处理.因此,本文根据方剂中存在的与功效配伍相关的信息即"对于方剂中某个功效,方中不是所有的药物都对该功效起主要作用",把从具有特定功效的方剂集中挖掘功效配伍药组问题看成是一个稀疏表达学习的问题.通过稀疏表达学习能够将方剂中药物之间的系数权重作为重要的鉴别信息引入模型,通过对输入药物进行稀疏约束,使之变得稀疏,这样可以去除与特定功效不相关的药物,使得对方剂功效具有重要预示作用的药物能够被保存下来.

5.1　稀疏表达学习

5.1.1　*稀疏表达模型*　传统的逻辑斯蒂判别式可用于两类或多类的非线性分类问题,通过计算样本所属类别的后验概率进行分类[18].对于二分类问题,若每次观测的 p 个预测变量值表示成向量 $\boldsymbol{x}=[x_1,x_2,\cdots,x_p]$ 且响应变量为 $y\in\{0,1\}$,那么相应的响应变量后验概率如式(1)所示.

$$\pi(\boldsymbol{x})=P(y=1\mid\boldsymbol{x})=\frac{1}{1+\exp[-(\boldsymbol{\omega}^T\boldsymbol{x}+\omega_0)]} \tag{1}$$

其中,$\boldsymbol{\omega},\omega_0$ 为模型待估计参数;$\boldsymbol{\omega}$ 是 p 维的特征参数;ω_0 是偏置项.若给定 q 个样本 $\chi=\{(x_t,y_t)\}_{t=1}^{q}$,则可以根据训练样本及其类标去估计模型参数,优化问题的损失函数为负的对数似然函数,如式(2)所示.

$$E(\boldsymbol{\omega},\omega_0\mid\chi)=-\sum_{t=1}^{q}[y_t\log(\pi(\boldsymbol{x}_t))+(1-y_t)\log(1-\pi(x_t))] \tag{2}$$

根据压缩感知和稀疏表示理论,L_0 范数能较好地刻画稀疏性,但是 L_0 最小化问题是非凸的,求解困难,因此常用 L_1 范数代替 L_0,把问题近似转化为 L_1 凸优化问题.L_1 范数正则化也称为绝对缩减和变量选择算子[16],一般是在模型的经验风险上加上模型参数向量中各个元素绝对值之和,如式(3)所示.

$$L(\boldsymbol{\omega})=\frac{1}{N}\sum_{i=1}^{N}(f(x_i;\boldsymbol{\omega})-y_i)^2+\lambda\|\boldsymbol{\omega}\|_1 \tag{3}$$

其中,N 为训练样例个数;$f(x_i;w)$ 为样例 x_i 的模型预测结果;y_i 为其黄金标注,第一项为模型的平方损失函数,第二项中 $\|\boldsymbol{w}\|_1$ 表示参数向量的 L_1 范数,$\lambda\geqslant 0$ 为正则化系数用于调整两项之间的关系.L_1 正则化可以在进行模型参数估计的同时使得部分对模型影响较小的变量的对应参数变为 0,通过让模型参数变得稀疏来实现特征的自动选择,也使得模型可解释性更强.

5.1.2　*方剂配伍挖掘的稀疏表达学习模型*　本文提出了利用稀疏表达学习(Efficacy Compatibility Mining with Sparse Representation,ECMSR)对特定功效的古方数据集进行挖掘,以此获得功效配伍规律,模型如图 2 所示.

ECMSR 模型是通过在 Logistic 判别模型中引入 L_1 正则化所实现.图 2 展示了模型的挖掘过程,首先以方剂集中出现的所有药物作为特征,每个方剂以 one-hot 的形式对其中的药物进行编码得到方剂的特征表示,同时逻辑斯蒂判别模型中的特征权重 w 与药物一一对应,然后使用坐标轴下降法训练模型使其从大量方剂中学习特定功效下的药物稀疏表达,并对训练好的模型参数进行降序排序同时解码成对应药物,最终通过取前 N 味药得到该功效下药味数为 N 的一个配伍药组.

假设对于方剂功效 f,给定方剂样本集为 $D=\{(\boldsymbol{x}_i,y_i)\}_{i=1}^{m}$,其中 x_i 为第 i 个方剂的 n 维的药物特征向量 $\boldsymbol{x}_i=(x_i^1,x_i^2,\cdots,x_i^n)$,$n$ 为该方剂集中不同药物的味数,每一维对应一种药物.对于在方剂 i 中出现的药物则对应特征维为 1,否则为 0.y_i 是功效类标且 $y_i\in\{0,1\}$,若方剂 i 含有功效 f,则 $y_i=1$,否则 $y_i=0$.根据式(1)~式(3)引入 L_1 正则化的 Logistic 优化目标函数为

$$\Phi(\boldsymbol{\omega},\omega_0\mid D)=-\sum_{i=1}^{m}[y_i\log(\pi(\boldsymbol{x}_i))+(1-y_i)\log(1-\pi(\boldsymbol{x}_i))]+\lambda\sum_{\omega}|\boldsymbol{\omega}| \tag{4}$$

式(4)中,第一项为 Logistic 损失函数,第二项为 L_1 正则化项.其中 m 为方剂个数,$\boldsymbol{\omega}$ 为 n 维参数向量,每一维对应表示一种药物的重要性.

虽然 L_1 正则化可以得出期望的稀疏性,但由于 L_1 范数是不可导的,从而 L_1 正则化问题很难得到对偶形式,传统的基于梯度的算法(最速下降法、牛顿法等)都无法对该问题加以解,坐标轴下降法是求解 L_1 范数最小化问题的常用方法,顾名思义即沿着坐标轴的方向去下降,通过启发式方法一步

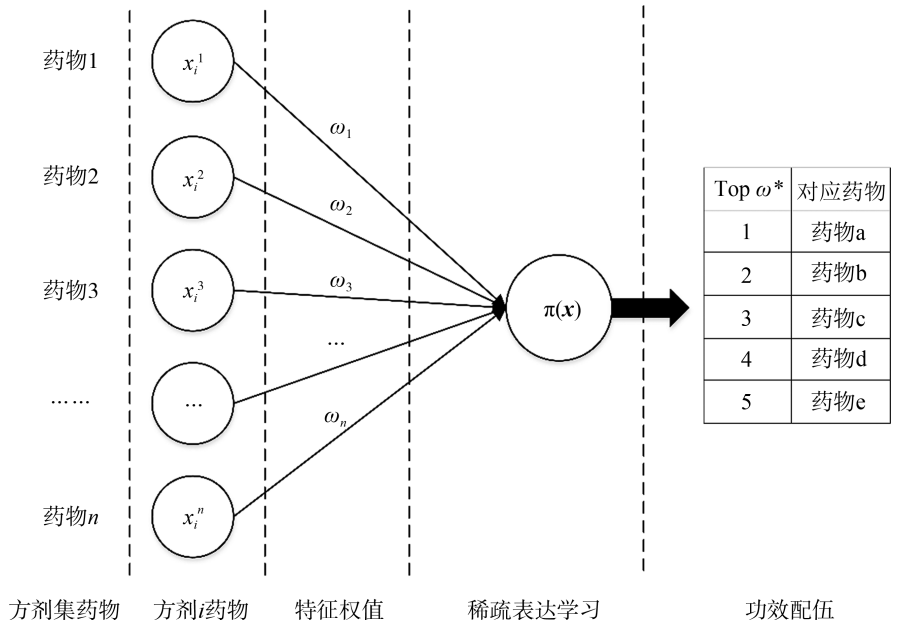

图 2 稀疏表达学习的方剂功效配伍模型

Fig. 2 Oriented-efficacy compatibility model based on sparse representation learning

步迭代求解函数的最小值.

因此,本文采用坐标轴下降法对优化目标函数式(4)进行迭代优化直到设定的最大迭代次数,可以得到训练好的具有稀疏性的权重 $\boldsymbol{\omega}^*$. ECMSR 算法伪代码如下:

输入 1) 训练数据集 $X\in \mathrm{R}^{n*d}$;
2) 训练数据标签 $Y\in \mathrm{R}_n$
3) 正则化参数 λ;
4) 最大迭代次数 maxiter;

输出 训练完成后并降序排序的模型参数w^*;
1) 随机初始化参数 w;
2) for $i=0,1,\cdots,d$
$w_i:=\arg\min_{wi}L(w_0,w_1,w_2,\ldots,w_d)$
3) 重复第 2) 步直到收敛;
4) 对训练得到的参数进行降序排序得到w^*.

5.2 功效配伍获取

根据训练好的模型得到$\boldsymbol{\omega}^*=[\omega_1^*,\omega_2^*,\cdots,\omega_n^*]$,对参数向量中元素的值进行降序排序,并把每个参数解码成它所对应的药物. 若要得到功效 f 的配伍药组,取 top N 可分别得到不同的功效配伍结果. 本文的实验中,$N\leqslant 5$.

6 实验结果与分析

6.1 实验设置

实验从收集的古方中,整理了 14 种功效的方剂数据. 为了解决每种功效只有分类模型中正例的情况,我们从其他功效中按正例数目的 1.3 倍随机收集负例.

为了对挖掘结果的准确性进行评估,本文收集了实验功效下中医专家总结的常用功效配伍作为标准配伍集. 实验的目的是希望利用稀疏表达的方法能够将中医专家总结的功效挖掘出来. 表 3 展示了各功效下标准配伍集的数据集的信息.

实验中,选择"清热解毒"功效作为开发集以此来确定模型的超参数即正则化因子 λ 为 4.0,模型的迭代终止条件为达到迭代次数 200 时,分类性能

表 3 配伍评估集

Tab. 3 The evaluation dataset

功效(f)	活血化瘀	化痰	壮筋骨	止血	消肿	利湿	明目	健脾	止咳	安神	活血止痛	止痒	通络
评估集配伍数(j)	10	19	13	31	19	29	15	30	15	11	11	4	16
配伍集药味数	17	38	15	47	32	49	23	39	24	31	20	11	31
对应实验功效数据没有的药味数	1	0	2	10	3	3	0	3	1	2	3	0	3

在开发集上获得最好的结果.

本文把 Apriori 算法和 Logistics 算法作为 baseline，将 ECMSR 和 baseline 进行比较.

6.2　评价方法

本文以 Dice 系数和平均查准率作为评估指标[17]，评估功效配伍结果的准确性和稀疏表达的有效性. 假设功效 f 有 j 个已知的标准功效配伍药组 A_u，$u=1,2,\cdots,j$，每个 A_u 一般为 2～5 味药的集合. 由 ECMSR 模型得到的特定味数的功效配伍药组结果为 B.

(1) Dice 系数.

通过 Dice 系数计算药物集合 A_u 和 B 的相似度表示如式(5)和式(6)所示.

$$\mathrm{sim}_u=\frac{2\mid A^u\cap B\mid}{\mid A^u\mid+\mid B\mid},u=1,2,\cdots,j \tag{5}$$

$$\mathrm{sim}=\max \mathrm{sim}_u \tag{6}$$

其中，$|A^u|$、$|B|$ 分别表示 A^u 和 B 中药物的个数；$|A^u\cap B|$ 表示 A^u 和 B 集合中相同药物的个数. sim_u 表示模型得到的功效配伍 B 与功效 f 的第 u 个标准配伍的相似得分，每个 A^u 是一个独立的配伍并且本文把每个 A^u 视为同等重要，在式(6)中选择最大的相似度得分作为功效配伍 B 的得分. sim 值越大代表模型准确性越高.

(2) 平均查准率.

稀疏表达模型输出的是排序过后的结果，我们希望与功效越相关的药物越靠前，因此采取信息检索中的平均查准率作为评价指标. 假设 R_f：$<h_1^f$，$h_2^f,h_3^f,\cdots,h_n^f>$ 是模型输出的药物排序，h_1^f 表示与功效 f 最相关的药物，而 h_n^f 表示与功效 f 最不相关的药物. 对于每个标准配伍 A^u 里所包含的药物是与功效 f 实际相关的药物，依次计算 R^f 排序中第 v 个药物 h_v^f 的查准率，如式(7)所示.

$$P(v)=\frac{s_v}{v},\ v=1,2,\cdots,n \tag{7}$$

其中，s_v 是 R_f 中从 h_1^f 至 h_v^f 与功效 f 实际相关的药物个数，v 为当前药物排名. 计算平均查准率(Average Precision)如式(8)，式(9)所示.

$$P_{\mathrm{avg}}{}^u=\frac{\sum_{h_v^f\in A^u}P(v)}{|A^u|},u=1,2,\cdots,j \tag{8}$$

$$P_{\mathrm{avg}}=\max P_{\mathrm{avg}}{}^u \tag{9}$$

其中，P_{avg} 越大不仅能体现找出了尽可能多的相关药物，也反映出排序越靠前越重要.

6.3　结果及分析

本文首先采用 Dice 系数作为评价指标对 ECMSR 模型和 Apriori 频繁项挖掘算法以及基本的 Logistic 模型进行评估，我们的模型总体上取得了最好的效果.

为了能进行实验对比，Apriori 算法挖掘结果中对于长度为 l 的功效配伍我们选取其支持度最大的 l-频繁项集. Logistic 模型，对式(2)中的损失函数采用梯度下降法训练模型，在表 4 的验证集数据中性能稳定最好时取迭代次数 $d=50$，最后采用与 ECMSR 模型同样的方式把参数解码为药物作为输出结果进行评估.

ECMSR 与 Apriori 和 Logistics 的 Dice 系数评估结果如表 4 所示.

从表 4 中可以发现，Apriori 算法的评估结果，各功效下不同长度药组的配伍总平均 Dice 系数为 40.48%，Logistic 模型的评估结果其平均 Dice 系数为 41.07%，而本文 ECMSR 模型的评估结果得到平均 Dice 系数为 42.35%，相比基本的 Logistic 模型提高了 1.28%，比 Apriori 算法提高了 1.87%. 说明 ECMSR 模型找到的配伍药组更准确. 表 4 中“消肿”功效三种模型 dice 系数为 0，是由于都没有找到与标准配伍集一致的药物. 总结发现“消肿”的标准配伍集主要以围绕蒲公英配伍居多，而对应实验方剂集中围绕金银花配伍居多. 尽管标准配伍集是实验前收集的专家总结的常见功效配伍，覆盖面比较全足以对三个模型进行评价，但仍存在少数配伍遗漏的情况，“消肿”功效也因此没有找到能够匹配的标准配伍药组.

为了验证稀疏表达的有效性，采用平均查准率作为评价指标对 ECMSR 模型和 Logistic 模型进行对比实验，如表 5 所示.

从表 5 可以看出，Logistic 模型的平均查准率总体平均值为 35.19%，而 ECMSR 模型基于平均查准率的评估结果，总体平均值为 36.71%，比基本的 Logistic 模型高 1.52%. 由此可见，稀疏性使得与功效越相关的药物排得越靠前.

通过与 Apriori 算法和基本的 Logistic 模型对比可知，ECMSR 模型在 Dice 系数和平均查准率上都有所提高，说明 ECMSR 模型对功效配伍挖掘是较准确有效的. 而“活血止痛”功效 dice 系数以及平均查准率实验结果比较差的原因是由于模型参数过于稀疏而欠拟合，考虑到超参数只在一组功

表 4 各模型 Dice 系数评估结果

Tab. 4 The evaluation results for Dice index on different models

	Logistic 模型					Apriori 算法					ECMSR 模型				
	单味	2 味	3 味	4 味	5 味	单味	2 味	3 味	4 味	5 味	单味	2 味	3 味	4 味	5 味
活血化瘀	0	0.4	0.67	0.57	0.75	0	0.4	0.33	0.57	0.25	0.5	0.8	0.67	0.57	0.75
化痰	0.5	0.4	0.33	0.29	0.25	0.5	0.4	0.67	0.86	0.75	0.5	0.4	0.33	0.29	0.25
壮筋骨	0.5	0.8	0.67	0.57	0.5	0.5	0.8	0.667	0.57	0.5	0.5	0.4	0.33	0.29	0.25
止血	0.5	0.4	0.67	0.57	0.5	0.5	0.4	0.33	0.29	0.25	0.5	0.4	0.67	0.57	0.5
消肿	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
利湿	0.5	0.4	0.33	0.29	0.25	0.5	0.4	0.33	0.29	0.25	0.5	0.4	0.67	0.57	0.5
明目	0.5	0.4	0.33	0.29	0.25	0.5	0.4	0.33	0.29	0.25	0.5	0.4	0.67	0.57	0.5
健脾	0.5	0.4	0.33	0.57	0.5	0.5	0.8	0.67	0.57	0.75	0.5	0.4	0.33	0.57	0.5
止咳	0.5	0.4	0.67	0.57	0.5	0.5	0.4	0.33	0.57	0.5	0.5	0.8	0.67	0.57	0.5
安神	0.5	0.8	0.67	0.57	0.5	0.5	0.4	0.33	0.57	0.5	0.5	0.8	0.67	0.57	0.5
活血止痛	0	0.4	0.33	0.29	0.25	0	0.4	0.33	0.29	0	0	0	0.33	0.29	0.25
止痒	0.5	0.4	0.33	0.29	0.25	0.5	0.4	0	0.29	0.5	0.5	0.4	0.33	0.29	0.25
通络	0.5	0.4	0.33	0.57	0.5	0.5	0.4	0.67	0.57	0.5	0.5	0.4	0.33	0.29	0.25
平均值	0.38	0.43	0.38	0.44	0.38	0.38	0.43	0.44	0.42	0.38	0.42	0.43	0.46	0.42	0.38
总体结果			0.40					0.41					0.42		

表 5 Logistic 与 ECMSR 模型平均查准率对比结果

Tab. 5 The average precision results between Logistic and ECMSR

	Logistic 模型					ECMSR 模型				
	单味	2 味	3 味	4 味	5 味	单味	2 味	3 味	4 味	5 味
活血化瘀	0	0.167	0.389	0.389	0.589	0.3333	0.67	0.67	0.67	0.67
化痰	0.3333	0.3333	0.3333	0.3333	0.3333	0.3333	0.3333	0.3333	0.3333	0.3333
壮筋骨	0.3333	0.67	0.67	0.67	0.67	0.3333	0.3333	0.3333	0.3333	0.3333
止血	0.3333	0.3333	0.389	0.389	0.389	0.3333	0.3333	0.389	0.389	0.389
消肿	0	0	0	0	0	0	0	0	0	0
利湿	0.333	0.333	0.389	0.389	0.389	0.3333	0.3333	0.389	0.389	0.389
明目	0.3333	0.3333	0.3333	0.3333	0.3333	0.3333	0.3333	0.389	0.5	0.5
健脾	0.3333	0.3333	0.3333	0.5	0.5	0.3333	0.3333	0.3333	0.5	0.5
止咳	0.3333	0.3333	0.556	0.556	0.556	0.3333	0.667	0.667	0.667	0.667
安神	0.3333	0.667	0.667	0.667	0.667	0.3333	0.667	0.667	0.667	0.667
活血止痛	0	0.167	0.167	0.167	0.167	0	0	0.111	0.111	0.111
止痒	0.3333	0.3333	0.3333	0.3333	0.3333	0.3333	0.3333	0.3333	0.3333	0.3333
通络	0.3333	0.3333	0.3333	0.3333	0.3333	0.3333	0.3333	0.3333	0.3333	0.3333
平均值	0.2564	0.3334	0.3762	0.3890	0.4044	0.2820	0.3590	0.3911	0.4018	0.4018
总体情况			0.3519					0.3671		

效验证集上确定，而实际每个功效的样本数目和药物数都有所不同造成. 虽然现在的超参数已适用于大多数功效集，但以后可以尝试在多组功效下确定更稳定的超参数.

表 6 为 5 味药组合的功效配伍挖掘结果，第二列从左到右按对应药物权重参数降序排列.

表 6　ECMSR 模型的功效配伍结果

Tab. 6　Compatibility rules of ECMSR

功效	功效配伍(五味药)
活血化瘀	桃仁，红花，丹参，没药，玄胡
化痰	贝母，半夏，天竺黄，天南星，天麻
壮筋骨	牛膝，覆盆子，枸杞，茴香，菟丝子
止血	三七，地榆，槐花，白芍，蒲黄
消肿	冰片，芙蓉叶，没药，金银花，连翘
利湿	茵陈，薏仁，滑石，泽泻，猪苓
明目	菊花，菟丝子，枸杞，石决明，粳米
健脾	白术，扁豆，神曲，党参，山药
止咳	贝母，杏仁，紫菀，半夏，粳米
安神	枣仁，茯神，朱砂，远志，琥珀
活血止痛	乳香，红花，没药
止痒	蛇床子，苦参，轻粉，白鲜皮，丹皮
通络	地龙，赤芍，桂枝，桃仁，丹参

以“活血化瘀”功效为例，图 3 表示在 ECMSR 模型训练得到的权重参数结果. 为了更清晰的展示，参数训练结果为 0 的药物被省略. 活血化瘀功效数据集预处理后共包含 355 种药物，最后通过稀疏表达的作用筛选出了 6 味药对该功效具有正向作用，338 味药受到稀疏性约束对应参数为 0.

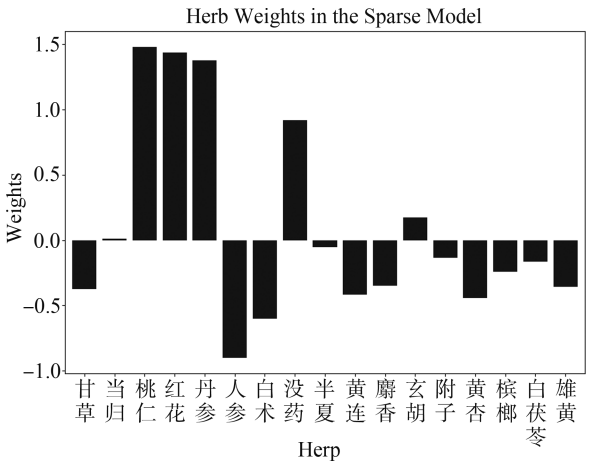

图 3　活血化瘀功效稀疏模型药物对应权重

Fig. 3　The weights of medicines on promoting circulation and removing stasis

7　结　论

本文根据中医方剂的功效和功效配伍之间的关系提出了一种利用稀疏表达学习自动挖掘方剂的功效配伍算法. 在 14 中功效的古方中进行了实验，使用 Dice 系数和平均查准率作为评估指标和 Apriori 算法进行了比较，实验结果说明本文提出的模型能更好的捕捉方剂中药物之间的关联，能更有效的发现功效配伍的规律.

本文在功效配伍中只考虑了药物的组成，而方剂中药物的剂量或者炮制方法都会影响到方剂的功效从而影响配伍的结果，在将来的研究功效配伍中尝试把剂量这些因素考虑进去，设计更有效合理的挖掘模型.

参考文献：

[1]　刘兴隆，贾波，张丰华，等. 基于能力培养的中药学专业《方剂学》课程教学改革探讨[J]. 时珍国医国药，2016，2016：1997.

[2]　彭京，唐常杰，曾涛，等. 基于神经网络和属性距离矩阵的中药方剂功效归约算法[J]. 四川大学学报：工程科学版，2006，38：92.

[3]　买买提依力·努尔买提，吐尔洪·阿西木，阿布都热依木·阿不都克里木，等. 基于高频药对及功效配伍数据分析探讨治疗冠心病的组方规律[J]. 中国实验方剂学杂志，2016，22：200.

[4]　窦立君，张金凤，刘爱利. 频繁闭图挖掘算法在中医方剂中的应用[J]. 吉林大学学报：理学版，2012，50：1223.

[5]　胡波，谭工. 基于关联规则的中医治疗乳腺增生病用药规律研究[J]. 中国实验方剂学杂志，2012，18：12.

[6]　张天嵩，张素，李秀娟，等. 治疗肺纤维化中药复方用药规律的数据挖掘[J]. 中国中医药信息杂志，2011，18：31.

[7]　潘定举，王丽颖，崔昊，等. 基于关联规则挖掘技术探索外用生肌中药处方配伍规律[J]. 贵阳中医学院学报，2017，39：41.

[8]　韩楠，乔少杰，宫兴伟，等. 面向正负关联规则的方剂配伍规律挖掘算法[J]. 小型微型计算机系统，2017，38：1538.

[9]　王倩，金卫，宋欣霞. 基于优化 Apriori 算法的中风病证治规律研究[J]. 医学信息学杂志，2017，38：62.

[10]　周雪忠，刘保延，王映辉，等. 复方药物配伍的复杂网络方法研究[J]. 中国中医药信息杂志，2008，15：98.

[11] 王映辉，周雪忠，张润顺，等.利用复杂网络与点式互信息法分析挖掘名老中医用药经验研究[J]. 中国数字医学，2011，6：76.

[12] Liu K，Sun Y，Zhang D. An intelligent drug matching method for traditional Chinese medicine[C]// International Conference on Cloud Computing and Intelligence Systems. Beijing：IEEE，2016.

[13] Wang L，Zhang Y，Xu X. Anovel group detection method for finding related chinese herbs[J]. J Inf Sci Eng，2015，31：1387.

[14] 吴嘉瑞，张冰，杨冰，等. 基于关联规则和复杂系统熵聚类研究颜正华治疗泄泻用药规律[J]. 中华中医药杂志，2013，28：2274.

[15] 李健，卢朋，唐仕欢，等. 基于中医传承辅助系统的治疗肺痈方剂组方规律分析[J]. 中国实验方剂学杂志，2012，18：254.

[16] Trofimov I，Genkin A. Distributed coordinate descent for L1-regularized logistic regression[C]// International Conference on Analysis of Images，Social Networks and Texts. Cham，Springer，2015.

[17] 吴胜利，谭延之，施化吉. 搜索引擎指标综合特性的评价[J]. 江苏大学学报：自然科学版，2015，36：181

[18] Tibshirani R. Regression shrinkage and selection via the lasso[J]. J R Stat Soc B，1996，58：267.

引用本文格式：

中　文：张思原，刘兴隆，姚攀，等. 利用稀疏表达学习挖掘中医方剂功效配伍[J]. 四川大学学报：自然科学版，2018，55：1180.

英　文：Zhang S Y，Liu X L，Yao P，*et al*. Utilizing sparse representationlearning to mine oriented-efficacy compatibility intraditionalchinese medicine prescriptions [J]. J Sichuan Univ：Nat Sci Ed，2018，55：1180.