

基于HPD算法的中药药对挖掘方法^{*}

薛琪¹, 高博², 温晶¹, 朱彦^{2**}, 孟祥福^{1**}

(1. 辽宁工程技术大学电子与信息工程学院 葫芦岛 125105; 2. 中国中医科学院中医药信息研究所 北京 100700)

摘要:人工智能方法辅助中药药对发现对于中医发展具有重要的现实意义,但当前的药对挖掘方法大多基于现有的单一的机器学习模型,使得药对发现的质量不高。本文旨在找出中医药候选药对,通过分析已知药对中饮片之间的特性和耦合关系,提出了一种改进的机器学习算法(Herb-Pairs Discovering, HPD)。该方法采用药对属性相关度评分方法,找出其中不低于给定阈值的饮片对。在基于朴素贝叶斯对饮片作用分类方法基础上利用决策树算法对饮片组合进行分析进而找出预测药对。在实际中医药测试数据上进行实验,结果分析表明,HPD算法得到的药对召回率达到82.7%,准确率达到80.6%,说明HPD算法可以有效地发现中医药饮片集合中潜在的药对。

关键词:药对挖掘 深度学习 机器学习 作用分类

doi: 10.11842/wst.20210918002 中图分类号: R-058 文献标识码: A

人工智能方法与中医药数据分析相结合,有效提升了中医药数据分析的质量,特别是对于药对发现、经典名方分析具有重要作用。本文以中医药中的药对发现为研究背景,提出一套针对药对发现的有效数据挖掘方法。药对(Couplet medicines),也称对药、对子^[1],两味药成对相配,多有协同增效或减毒作用。药对是方剂配伍的最小单元,在辨证的基础上确立相应治法,根据治法选用特定性能和功效的药物进行组合配对,并经过临床应用被证明行之有效。药对的使用是中医药学家长期医疗实践的经验总结和精华所在,体现了中药应用的基本原则^[2]。通过对药对进行研究,可以找出方剂配伍之间的隐含规律,为中医药的信息化提供技术支持。在人工智能与中医药数据相结合的领域^[3],目前专门针对药对发现的研究还不够深入和系统,药对更多是在临床实践过程中形成的相对固定的饮片搭配,与药对相关的数据也多散落在大量的中药、方剂论述与中医临床处方中,使得数据挖

掘技术在药对发现中面临较大挑战^[4-5]。

本文提出了一种改进的机器学习算法—HPD,该方法在饮片之间的共现基础上,充分考虑饮片的性、味、归经等属性,通过先预测再分类的方法有效提高查找药对的准确率并发现潜在药对。

本文HPD算法分为3个步骤,主要贡献如下:①综合考虑候选药对中饮片的性、味、归经、功效等属性信息,统计各类信息出现的频次,在此基础上将每个饮片信息转换成对应的向量表示。②依据饮片的性、味、归经、功效等属性信息,提出基于朴素贝叶斯算法的药对作用分类方法,获得了较好的分类效果。③按照药对作用分类对药对分别进行预测,并且在预测药对之前,更换对应作用分类的训练集,以增加训练集多元性以及效果准确性。

1 相关工作

近年来,人工智能和机器学习在医学领域引起了

收稿日期:2021-09-18

修回日期:2022-05-08

^{*} 国家自然科学基金委员会面上项目(82174534):广义中医经典名方智能辅助遴选系统关键技术研究,负责人:朱彦;国家科学技术部国家重点研发项目(2019YFC1710400,2019YFC1710401):以疗效为核心构建病证结合数据框架与个体化评价指标,负责人:戴国华。

^{**} 通讯作者:朱彦,副研究员,主要研究方向:中医药知识组织与发现、方剂药品信息标准化与分析挖掘;孟祥福,教授,主要研究方向:空间数据管理、推荐系统和大数据可视化。

极大关注^[6]。医疗数据的快速增长促使研究者们将大数据方法引入医疗领域中。数据挖掘是从大量的数据中,抽取出潜在的、有价值的知识的过程,也称为数据库中的知识发现,融合了数据库、人工智能、机器学习、统计学、知识工程、面向对象方法、信息检索、高性能计算以及数据可视化等最新技术的研究成果,是一个多学科交叉研究领域。相关技术方法包括:遗传算法、粗集方法、决策树、人工神经网络、模糊逻辑、规则归纳、聚类分析、模式识别、频繁集方法、最近邻技术、可视化技术等。

关联规则是在药对挖掘中应用较多的数据挖掘方法。借助于商业挖掘软件如 SAS Enterprise Miner、IBM SPSS Modeler、Insightful Miner 等关联规则分析模块,通过不同途径获取大量样本方剂数据集,通过对药物与药物之间关联关系的挖掘,获得了不同适应症的广义药对。但有时经典的关联规则算法(如 Apriori 算法和 FP-Growth 算法等)并不能很好地满足中医药研究的需要,因此一些研究者自行开发或改进挖掘算法。如黄黎明等^[7]、曾令明等^[8]、引入了关注频繁项概念,提出了基于位图矩阵和双支持度的 Apriori 算法,进行药对挖掘的实践。唐仕欢等^[9]提出了基于属性互信息熵的关联规则挖掘算法,从大量关联关系中挖掘出具有正向关联关系的药对。袁楠等^[10]采用聚类方法自动划分药物剂量的模糊区间,提出药对量效关联规则的挖掘算法。Ung 等^[11]综合运用概率神经网络(probabilistic neural network method, PNN), k 近邻(k nearest neighbor, kNN)、支持向量机(support vector machine, SVM)等方法,对 394 个药对与 2470 个非药对中药的性味属性予以属性分类与模式识别,发现利用人工智能与模式识别方法,可从大量非药对中对约定俗成的药对予以识别确认。尚尔鑫等^[12]分别采用标准关联规则发现 Apriori 算法以及改进多数据库计算方法,对从历代药对文献中收集整理得到的 625 个药对等共 49 个属性形成的数据库进行挖掘研究,并对两种方法得到的结果进行比较。刘子闻等^[13]在 18503 个血瘀证医案药方中,运用关联规则方法通过频繁项集算法找出具有特殊配对关系的既包含已知的,又包含未知的药对。

以上提到的论文对药对的发现有良好的效果,但很少挖掘和分析饮片属性以及候选药对中饮片属性之间的关系,然而这些关系对于药对发现至关重要。

受文献^[14]的启发,本文设计了一种新的药对挖掘 HPD 算法,该算法能利用药对中饮片间属性信息和作用分类对药对进行分类及挖掘。

2 问题定义和解决方案

2.1 问题定义

经过多年的中医药实践,医者认识到用几味药配合起来治病的效果更好。现代药对是指两味药配伍组方^[15],是历代医家长期临证的经验总结,遵循中药配伍相须、相使、相畏、相杀、相恶、相反原理^[16],在临床组方时常常两味药物配伍使用,有良好的协调作用,或对过于猛烈的药物有较好制约作用。如药对半夏和陈皮,陈皮有理气健脾、燥湿化痰的功效,半夏能燥湿化痰、降逆止呕、消痞散结,二者配合使用,以理气促进水行,气顺则水行,湿去则痰消。在经典名方“二陈汤”中,即由半夏、陈皮为主药,达到燥湿化痰、理气和胃的功效,临床用于治疗痰湿停滞,咳嗽痰多、胸脘胀闷、恶心呕吐等。半夏与生姜配伍使用,则通过生姜的温中止呕、化痰止咳功效,加强半夏的燥湿化痰、降逆止呕作用,并通过温热药性促进痰湿运化消除,兼制约半夏的毒性。如化痰散饮、和胃降逆的经典名方“小半夏汤”。找到隐藏药对有利于提高疗效降低副作用,使用更少的饮片发挥更大的作用,并对医者处方用药有启发性的作用^[3]。

近年来,很多学者把关联规则及其改进算法应用到药对发掘工作,也有部分算法运用到饮片中的性、味、归经等属性进一步提高药对发掘的准确性。除此之外,我们还应该挖掘药对中两味饮片性、味、归经等属性之间的相关性,而非仅考虑单个饮片的属性信息。并且,通过对候选药对的作用进行分类,随分类更换训练集也可以在发掘药对方面取得更好的效果。

给定一个中医药饮片数据集合,包含 n 个不同的饮片, $D=\{e_1, \dots, e_n\}$,任意一对饮片的组合都是候选药对(潜在药对)。本文目的是,设计一种快速有效的药对挖掘方法,找出 top- k 个具有最高可能性的药对。

2.2 解决方法

本文提出的模型主要分为以下 3 个阶段:

(1)选择相关度程度高的候选药对。对每个饮片进行两两组合、去重,算出这一对候选药对中两个饮片之间的性、味、归经、功效向量。把四个值相加挑选出数值大于 1.5 的候选药对。

(2)利用朴素贝叶斯的方法,根据步骤一提到的每组候选药对中两个饮片之间的属性向量,通过训练已有药对的属性向量及作用来对候选药对的作用进行分类。

(3)在按照药对作用分类进行预测时,根据数据库中的数据放到训练集,再用数据库的药对进行判断。此种方法的训练集是根据判断出的药对类型而定。

3 数据来源与实现方法

3.1 数据来源

目前与药对相关的数据挖掘研究还存在着一些问题。其中包括,研究需要的基础数据多来自于古今医著或期刊文献,资料收集不全面会直接影响分析结论的可靠性^[4]。本文所用数据来源于中国中医科学院中医药信息研究所开发的“古今方药系统”^[5]。该数据对古今中医方剂文献进行全面系统收集整理的基础上,共收录方剂33938首,包含918种已知药对和其作用分类、配伍机制、用法用量、临床应用等内容和9207个饮片性、味、归经、功效、疾病部门等方面内容,相关数据都参考来源包括《药典》、各级标准、教材等权威资料。

3.2 HPD方法

HPD算法发掘药对总共分为3步,具体算法流程如图1所示:

3.3 药对属性相关度评分方法

3.3.1 饮片药性、药味和归经相关度计算

饮片的药性为寒、热、温、凉中的一种,则把这个值设为1,其余设为0。如果这个饮片的性为平,则四种性每一个分别为0.25。“小寒”这类有程度副词的性,本文参照文献^[4]设为0.8。对于两个饮片之间性的关联,我们用如下公式表示:

$$A_{ij} = \frac{\sum_{k=1}^4 m(X_{ki}, X_{kj})}{\sum_{k=1}^4 m(X_{ki}, X_{kj})}$$

举个例子,如果想计算饮片x和饮片y性的值,则药性的表格如表1所示:

同理可以计算饮片药味 B_{ij} 和归经 C_{ij} 的数值。

3.3.2 功效相关度的计算

对于候选药对中饮片功效相关度的计算,我们通过计算两个饮片相同的功效数量和各自的功效数量,并运用如下公式进行计算,示例如表2:

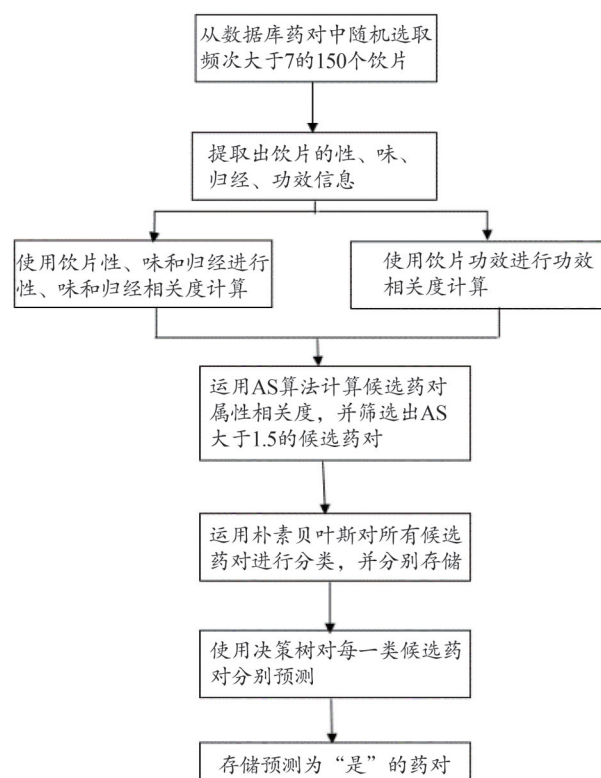


图1 HPD算法流程图

表1 饮片x和饮片y性的值

饮片 药性	寒	热	温	凉
饮片x	1	0	0	0
饮片y	0.25	0.25	0.25	0.25

注:假设饮片x的性为“寒”、饮片y的性为“平”。则性的数值如表1所示。所以 $A_{x,y} = 0.25 / (1 + 0.25 + 0.25 + 0.25) \approx 0.1429$

表2 饮片x和饮片y功效相关度

饮片	功效			
x	1	2	5	6
y	2	4	5	

注:假设饮片x的有四个功效、饮片y有三个功效。功效2和功效5为它们共同的功效。则功效的相关程度 $D_{x,y} = (2 * 2) / (4 + 3) \approx 0.5714$

$$D_{ij} = \frac{2FT_{ij}}{FT_i + FT_j}$$

3.3.3 AS的计算

候选药对属性相关度评分系统(AS)的值有如下公式:

$$AS_{ij} = A_{ij} + B_{ij} + C_{ij} + D_{ij}$$

我们挑选出AS大于1.5的候选药对进入步骤二中。

3.4 朴素贝叶斯预测分类

朴素贝叶斯作为一种简单实用的分类方法,将预测未知样本 Z 属于具有最高后验概率(条件 Z)的类。即将未知的样本分配给类 C_i ,当且仅当 $P(C_i|Z) \geq P(C_j|Z)$,其中 $1 \leq j \leq m$ 且 $j \neq i$ 。这样,我们最大化 $P(C_i|Z)$ 的值,而使得 $P(C_i|Z)$ 最大的类 C_i 则称为最大后验假定^[16]。根据贝叶斯定理^[17],有

$$P(C_i|Z) = \frac{P(Z|C_i)P(C_i)}{P(Z)}。$$

首先用jieba软件包对步骤一中的候选药对的性、味、归经和功效信息进行中文分词,将训练集中每一个作用分类的药对放到一个文件夹中。运用TF-IDF逆文本频率指数用以评估一个词对于一个语料库中一份文件的重要程度。使用贝叶斯分类中的朴素贝叶斯方法对所有候选药对进行训练。在判断好每个候选药对的作用分类后,按照作用分类把候选药对分别存储。

3.5 药对判断

决策树模型是在线数据流分类的最先进方法之一,相对任何其他流或批处理学习算法^[18],能快速地处理大量的数据。作为数据流分类中常用的模型之一,决策树模型常用的属性分裂准则包括信息增益(Information Gain)、增益率(Gain Rate)、基尼指数(Gini Index)^[19]。因为决策树是一种监督学习^[20],所谓监督学习就是给定一堆样本,每个样本都有一组属性和一个类别,这些类别是事先确定的,那么通过学习得到一个分类器,这个分类器能够对新出现的对象给出正确的分类。故本文在预测不同作用分类的候选药对时,会更换训练集,其中训练集与测试集的比例选定为7:3。即,例如,在预测解表类药对时,训练集中只有解表类药对后面的标签为“Yes”,其余均为“No”。这样可使模型充分学习到解表类药对的特性,从而对预测达到一个更好的效果^[21]。

4 实验及结果分析

本节主要通过真实数据集上进行实验来测验本文所提出发现药对算法的性能。

实验数据集:因药对是常用配伍,经常在方剂中一起使用,所以为保证本次实验的实用性,我们从历代方剂结构化数据库中用Python中Collections模块的Counter类进行饮片的筛选。由于药对是由多个饮片

组成的,首先统计出数据库918种药对总共包含的459个饮片。然后再从459个饮片随机挑选出现频次大于7的150个饮片,从10个饮片为一组,5个饮片为递进数量,直至150个饮片为一组。提取出饮片的性、味、归经、功效信息。并取出数据库已知918种药对的作用分类及其中饮片的相关信息。

4.1 实验结果

在使用HPD算法测试完150个饮片之后,找到了119个候选药对。其中包含数据库中药对33个。在剩下的86个药对中,由领域专家根据饮片性味功效的相互作用进行判断,如候选药对中两种药物性味相合、功效上相辅相成,在针对某一类型证候时配合使用可以提高疗效,对临证组方有新的启发,则标注为新的药对。

经过专家人工标注,选出了可判定为药对的饮片组合63组,HPD算法发掘的部分未知药对有{甘草,天麻}、{紫菀,半夏}、{生姜,苍术}、{生姜,白术}。新发现的药对性味相合,功效上相辅相成,符合方剂配伍规律,对医生临证组方有很好的启发作用。

4.2 评价标准

本文的评估采用精确率(Precision)、召回率(Recall)、F值(F-measure),其中,F值为精确率与召回率的调和平均值,统一度量精确率与召回率。记TP为判断正确的匹配,FP为判断错误的匹配,FN为没有判断出来的正确匹配。3个评估标准的计算方法如下^[22]:

$$Precision = \frac{TP}{TP + FP} \times 100\%$$

$$Recall = \frac{TP}{TP + FN} \times 100\%$$

$$F - measure = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

4.3 实验结果的准确性

通过计算,与数据库中药对比对和人工标注后的准确率和召回率分别如下图2和图3。通过折线图可知,使用HPD算法进行测试,准确率和召回率随着饮片数量的增加而逐渐上升,最后趋于稳定。在专家人工标注后,找到药对的召回率高达82.7%,准确率高达80.6%。对比专家标注前后,召回率没有明显上升,稳定在百分之八十左右,说明HPD对已有药对的判断较为准确。相较于标注前,准确率有明显上升,说明HPD算法可有效发掘隐藏药对。

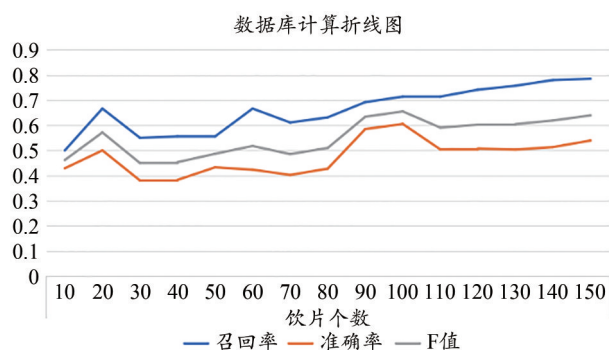


图2 按已知药对计算的准确率、召回率和F值

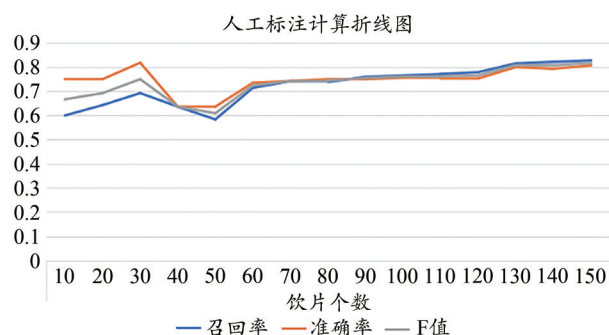


图3 人工标注后计算的准确率和召回率和F值

4.4 实验对比与分析

进行下列实验对比与分析中,选用了相同的数据集,按照分别按照以下论文的算法与参数进行测试。

4.4.1 与《基于关联规则提取的中药药对发现》^[13]的算法对比

在《基于关联规则提取的中药药对发现》论文中,作者采用了用关联规则算法找出方剂中药对的支持度和置信度,定义最小支持度和置信度为0.2。从中筛选出支持度和置信度较高的药对,运用中医理论对所得结果进行分析。我们采用同种方法,先把本文所用到的饮片两两组合,对所有候选药对分别计算每个饮片和共同出现在所有方剂中出现的频次,然后进行支持度和置信度的计算,筛选出支持度和置信度都大于等于0.2的候选饮片进行分析。部分候选药对支持度及置信度如表3。

通过对比候选药对和数据库中的药对,可计算出召回率为0.6428,准确率为0.4688。经过人工标注后,召回率和准确率均为0.7649。从结果可得知,该方法的准确率和召回率并没有展示出很好的效果。关联规则提取药对需要全面多次扫描数据库,会出现时间和空间复杂度急剧上升的现象,十分浪费资源。而且单纯考虑支持度和置信度,会出现在单个方向上的置

表3 部分候选药对支持度及置信度

药对名称	支持度	置信度
川芎,当归	0.271275596	0.690909091
川芎,半夏	0.253644315	0.426470588
白术,茯苓	0.26763285	0.617476594
紫石英,川芎	0.202857143	0.298947368
甘草,人参	0.253539591	0.402246256
当归,白芍	0.246845426	0.387256418
甘草,茯苓	0.244702842	0.393926789
防风,羌活	0.244177019	0.380290206
甘草,当归	0.235791568	0.394342762
白术,陈皮	0.225189961	0.436023183
天麻,全蝎	0.224299065	0.431654676
茯苓,陈皮	0.219793323	0.377216917
紫石英,紫菀	0.219117856	0.321339434
人参,当归	0.219064926	0.470212766
川芎,当归	0.271275596	0.690909091
川芎,半夏	0.253644315	0.426470588
白术,茯苓	0.26763285	0.617476594
紫石英,川芎	0.202857143	0.298947368
甘草,人参	0.253539591	0.402246256
当归,白芍	0.246845426	0.387256418
半夏,陈皮	0.213438735	0.477876106

信度很高反方向置信度却很低的情况,但是因存在一个方向置信度高,所以会被放在潜在药对中。比如本文中,甘草和升麻,甘草和升麻的置信度为0.1204,而升麻和甘草的置信度为0.74613。二者相差0.62573。并且,需要随着总样本数量的变化对置信度和支持度进行实时调整,其合理的选择也是决定算法效果的一个重要因素,若置信度过高,会忽略一些重要因素,若置信度过低又会有很多冗余的信息掺杂其中影响效果,比较浪费精力而且结果很不稳定。

4.4.2 与《基于联合条件概率矩阵的药对提取算法设计及应用》的算法对比

在《基于联合条件概率矩阵的药对提取算法设计及应用》^[23]一文中,构造联合条件概率矩阵,设定最小联合条件概率矩阵(minUnion)作为药对提取阈值,从而形成药物社团发现算法TCM-HPD。实验结果部分如表4所示。通过对比候选药对和数据库中的药对,可计算出召回率为0.6435,准确率为0.5172。经过人工标注后,召回率为0.7824和准确率为0.7738。主要原因是药对所包含的两味药本身还有两味饮片之间的性、味、归经、功效等属性联系,此方法并没有考虑完全。单从饮片共同出现频次上看,如甘草与其他饮片出现频次很高,但是很多带有甘草的组合都不是隐藏药对,如{甘草、细辛}{甘草、生姜}。所以加入饮

表4 部分合条件概率矩阵

名称	香附	紫菀	生姜	杏仁	麻黄	朱砂	牛黄	五味子	苍术	大黄	...	半夏
香附	1	0.28	0.271	0.36	0.405	0.25	0.142	0.22	0.67	0.501	...	0.628
紫菀	0.28	1	0.593	0.82	0.68	0.37	0.208	0.79	0.181	0.62	...	0.69
生姜	0.271	0.593	1	0.51	0.54	0.182	0.09	0.38	0.392	0.503	...	0.71
杏仁	0.36	0.82	0.51	1	0.648	0.257	0.203	0.601	0.26	0.622	...	0.645
麻黄	0.405	0.68	0.54	0.648	1	0.572	0.274	0.526	0.59	0.58	...	0.63
朱砂	0.25	0.37	0.182	0.257	0.572	1	0.643	0.2	0.19	0.48	...	0.269
牛黄	0.162	0.208	0.09	0.203	0.274	0.643	1	0.206	0.16	0.618	...	0.453
五味子	0.22	0.79	0.38	0.601	0.526	0.2	0.206	1	0.21	0.242	...	0.607
苍术	0.67	0.181	0.392	0.26	0.59	0.19	0.16	0.21	1	0.42	...	0.65
大黄	0.501	0.62	0.503	0.622	0.58	0.48	0.618	0.242	0.42	1	...	0.46
...	...	...	...	...	...	...	...	...	...	...	...	...
半夏	0.628	0.69	0.71	0.645	0.63	0.269	0.453	0.607	0.65	0.46	...	1

表5 算法效果展示

算法名称	召回率(标注前)	准确率(标注前)	F值(标注前)	召回率(标注后)	准确率(标注后)	F值(标注后)
HPD算法	0.7857	0.5393	0.6396	0.8276	0.8067	0.8170
关联规则	0.6428	0.4688	0.5421	0.7649	0.7649	0.7649
TCM-HPD算法	0.6435	0.5172	0.5735	0.7824	0.7738	0.7781

片的属性信息更有利于模型分析出药对间的特点、找到药对中饮片之间的相互规律、发现饮片与饮片间性味归经的互补或促进规律,从而提高查找药对的准确率。

三种算法的召回率准确率和F值放在一起,对比统计如表5。

5 讨论

现有药对发现方法,主要采用关联规则及其改进算法,仅关注于饮片共现信息和饮片自身属性,未纳入药对中饮片之间的性、味、归经、功效相关度等属性,造成准确性较低。针对这一现状,本文提出了一种新的HPD算法来实现中药药对的自动发现。实验

结果表明,本文提出的算法可以有效的发掘隐藏药对,相比传统关联规则挖掘方式效果提升10%左右。在一定程度上为临床医生减轻了药物配伍的烦恼,提供更多的组方选择,以便用更精简的药物达到更好疗效。此外对于药对的掌握也是体悟方剂配伍理论的基础,新药对的发现对组方和临床应用有启发作用,可以为新药研发提供参考,寻找新的突破方向,缩短研发周期。同时中医是一门实践性科学,新发现的药对,也必须在临床应用中观察疗效,验证结果。

未来的工作中,将会利用先进的深度学习算法和文本分析算法,对发掘隐藏药对算法进行改进,进一步提高准确率。最终还需要结合具体的应用的病证或功效分类,再进行分析或经过试验和临床的验证。

参考文献

1 中医药学名词审定委员会. 中医药学名词. 北京: 科学出版社, 2004.

2 唐于平, 束晓云, 李伟霞, 等. 药对研究(I)——药对的形成与发展. 中国中药杂志, 2013, 38(24):4185-4190.

3 Ruan C Y, Wu Y P, Yang Y, *et al.* Semantic-aware graph convolutional networks for clinical auxiliary diagnosis and treatment of traditional Chinese medicine. *IEEE Access*, 9:8797-8807.

4 Zhang Q C, Bai C C, Chen Z K, *et al.* Deep learning models for diagnosing spleen and stomach diseases in smart Chinese medicine with cloud computing. *Concurr Comp Pract Exp*, 2019, 33(4):e5252.

5 Cui Y D, Gao B, Liu L H, *et al.* AMFormulaS: An intelligent retrieval

system for traditional Chinese medicine formulas. *BMC Med Inform Decis Mak*, 2021, 21:56.

6 Li C J, Liu D, Yang K X, *et al.* Herb-know: Knowledge enhanced prescription generation for traditional Chinese medicine. 2 *BIBM*, 2020:1560-1567.

7 黄黎明, 刘振宇. 用改进Apriori算法确定药房处方药物的关联规则. 电子设计工程, 2018, 26(24):36-40.

8 曾令明, 唐常杰, 阴小雄, 等. 基于位图矩阵和双支持度的中药配伍挖掘技术. 四川大学学报(自然科学版), 2005, 42(1):57-62.

9 唐仕欢, 陈建新, 杨洪军, 等. 基于复杂系统熵聚类方法的中药新药处方发现研究思路. 世界科学技术(中医药现代化), 2009, 11(2):

- 225–228.
- 10 袁楠, 金晖, 田玲, 等. 基于聚类 and 模糊关联规则的中医药对量效分析. 计算机应用研究, 2009, 26(1):59–61.
 - 11 Ung C Y, Hu L, Cao Z W, *et al.* Are herb-pairs of traditional Chinese medicine distinguishable from others? Pattern analysis and artificial intelligence classification study of traditionally defined herbal properties. *J Ethnopharmacol*, 2007, 111(2):371–377.
 - 12 尚尔鑫, 叶亮, 范欣生, 等. 基于改进关联规则算法的中药药对药味间性味归经功效属性关系的发现研究. 世界科学技术(中医药现代化), 2010, 12(3):377–382.
 - 13 刘子闻, 陈守强, 徐赛, 等. 基于关联规则提取的中药药对发现. 世界最新医学信息文摘, 2019, 19(20):275–276.
 - 14 Fang Z, Zhang M X, Yi Z H, *et al.* Replacements of rare herbs and simplifications of traditional Chinese medicine formulae based on attribute similarities and pathway enrichment analysis. *Evid Based Complement Alternat Med*, 2013, 2013:136732.
 - 15 刘春慧, 耿刚. 基于药对配伍的全蝎组方探讨. 中医杂志, 2019, 60(18):1563–1566.
 - 16 王嘉俊, 李双蕾, 李梦瑶. 中药药对的现代认识与研究. 中医杂志, 2016, 57(8):701–704.
 - 17 Aridas C K, Karlos S, Kanas V G, *et al.* Uncertainty based under-sampling for learning naive Bayes classifiers under imbalanced data sets. *IEEE Access*, 8:2122–2133.
 - 18 Ikononovska E. Online tree-based ensembles and option trees for regression on evolving data streams. *Neurocomputing*, 2015, 150: 458–470.
 - 19 Dan Y B, Dong R Z, Cao Z, *et al.* Computational prediction of critical temperatures of superconductors based on convolutional gradient boosting decision trees. *IEEE Access*, 2020:57868–57878.
 - 20 Messinger A I, Luo G, Deterding R R. The doctor will see You now: How machine learning and artificial intelligence can extend our understanding and treatment of asthma. *J Allergy Clin Immunol*, 2020, 145(2):476–478.
 - 21 Shafaf N, Malek H. Applications of machine learning approaches in emergency medicine; a review article. *Arch Acad Emerg Med*, 2019, 7(1):34.
 - 22 王丰, 王亚沙, 赵俊峰, 等. 一种基于迭代的关系模型到本体模型的模式匹配方法. 软件学报, 2019, 30(5):1510–1521.
 - 23 汪叶群, 杨涛, 李鑫欣, 等. 基于联合条件概率矩阵的药对提取算法设计及应用. 世界科学技术-中医药现代化, 2019, 21(6):1153–1160.

A HPD-based Approach for Discovering the Herb-pairs

Xue Qi¹, Gao Bo², Wen Jing¹, Zhu Yan², Meng Xiangfu¹

(1. School of Electronic and Information Engineering, Liaoning Technical University, Huludao 125105, China;

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

Abstract: The artificial intelligence method to assist Chinese medicine has important practical significance for the development of Chinese medicine, but the current herb pairs discovery methods are mostly based on the existing single machine learning model, which makes the quality of herb pairs discovery not high. This paper is proposed for the discovery of Chinese herb pairs by analyzing the characteristics and coupling relationship between known medicine pairs and Chinese decoction pieces, an improved machine learning algorithm (Herb-Pairs Discovering, referred to as HPD). This method uses the scoring method of herb pairs attribute correlation to find the pairs of decoction pieces that is not lower than a given threshold. On the basis of the classification method based on Naive Bayes on the effect of the decoction pieces, the decision tree algorithm is used to analyze the combination of decoction pieces to find the predicted herb pairs. Experiments on actual Chinese medicine test data showed that the herb pairs recall rate obtained by the HPD algorithm reached 82.7%, and the accuracy rate reached 80.6%, indicating that the HPD algorithm can effectively find the potential herb pairs in the collection of Chinese medicine decoction pieces.

Keywords: Herb-pairs mining, Deep learning, Machine learning, Role classification

(责任编辑: 李青, 审稿人: 王瑀、张志华)