基于频繁模式挖掘算法的中医问诊策略研究*

李瑞珍[1,2]，夏春明[2,3**]，王忆勤[4]，许朝霞[4]，熊玉洁[3]

（1. 西北工业大学民航学院 西安 710129；2. 华东理工大学机械与动力工程学院 上海 200237；3. 上海工程技术大学电子电气工程学院 上海 201620；4. 上海中医药大学上海市健康辨识与评估重点实验室/中医四诊信息化实验室 上海 201203）

摘 要：目的 研究中医问诊策略，实现快速捕捉患者的关键病情信息，推进中医问诊客观化的发展。方法 采用基于关联分析中频繁模式挖掘算法的症状提问模型，并使用交叉合并的方法建立中医单系统症状提问与多系统综合症状提问的中医症状问诊策略，达到通过最短的时间、最高的效率来获取到患者关键病情信息。结果 实现了从单系统问诊到五系统综合问诊的突破，通过单系统与五系统两种症状提问模式实现了高效获取患者病情信息的过程，且对比传统量表提问方式，系统减少了65%的提问次数就可获取到患者92%的症状信息，大大提高了对患者症状信息获取的效率。结论 在两种不同的症状提问模式下，打破了中医基于量表来询问患者的传统问诊模式，缩短了对患者症状获取的时间，简化了问诊流程，减少了由于经验不足或人为主观造成的差异，能够用于中医临床辅助诊断中。

关键词：中医问诊 频繁模式挖掘算法 症状关联性 问诊策略

doi: 10.11842/wst.20230316001 中图分类号：R-058 文献标识码：A

问诊作为中医诊断中的一个重要环节，医师通常围绕患者的主诉来展开询问，询问内容涉及五脏，有经验的医师不仅可以通过丰富的临床经验进行针对性的问询，还能够根据患者的答复迅速捕捉到有利于诊断的关键信息[1]，但是对于经验不足的医生如何对患者进行针对性的询问一直是中医问诊研究的核心[2-4]。

随着中医问诊客观化的不断发展，逐渐形成了以中医问诊量表为主的询问模式[5-6]，问诊量表虽然规范了问诊的流程，但问诊量表中症状数目过多，依据量表询问患者会造成询问时间过长、问诊效率低等情况[7]。若能够在规范化的中医量表中使用数据挖掘算法[8]，对症状进行关联分析，建立起症状之间的关联性，系统就可以根据患者的症状进行针对性询问，这样的方式不仅可以大大提高问诊的效率，也能够加快中医问诊客观化的进程[9-11]。

在当前的中医科研领域中，大多研究都集中于对某一种特定疾病或是某一特定部位进行细化分析，然而中医学认为人体是一个完整的有机体，通过经络的连接，形成了以心、肝、脾、肺、肾五脏为核心的生理系统[12-13]，因此，停留在对五脏中某一特定系统的研究是远远不够的，从单系统研究过渡到多系统研究也成为了中医客观化发展的必然趋势[14-17]。

本文将从单系统的症状关联性研究出发，由单系提问模式搭建出五系统综合症状提问模式，开发出心、肝、脾、肺、肾五脏综合症状提问系统，既能够实现单个生理系统的症状提问，又能够对患者的五脏进行综合性了解，使中医问诊模式多样化、全面化发展，推进中医问诊客观化的进程[18-21]。

收稿日期：2023-03-16

修回日期：2023-11-04

* 上海市科学技术委员会科技项目（21DZ2203100）：基于机器学习的中医脉诊信息采集与分析国际标准研究，负责人：夏春明；上海市科学技术委员会科技项目（21DZ2271000）：上海市健康辨识与评估重点实验室，负责人：王忆勤；国家自然科学基金委员会面上项目（82074333）：基于舌诊多源信息及临床危险因素的动态变化探讨早发冠心病中医证候演变规律，负责人：许朝霞。

** 通讯作者：夏春明，教授，博士生导师，主要研究方向：工程与生物医学信号分析与处理。

1 资料与方法

1.1 实验数据

1.1.1 单系问诊数据

本文使用的数据是在上海中医药大学的各附属医院收集得到的，收集过程主要根据不同的中医问诊专科量表[22]分别对心、肝、脾、肺、肾5个不同系统的患者进行问诊及信息采集，包括患者的一般情况、既往病史等，此采集过程均由具有中级职称或者博士学位的医师来完成，数据最终经过中医四诊信息化实验室的整理，得到心、肝、脾、肺、肾五5份病例数据集。部分数据集如下表1所示，在某一病例中，若拥有某一症状或证候[23]时，对应位置为1，否则为0。表2为数据集的统计分析结果，在采集到的原始数据中，会删除掉一些频次太低的特征，表中的最终病例样本、最终症状特征为剔除掉频次较低特征后的结果。

1.1.2 五系综合问诊数据

本文使用的数据为相互独立的5份数据集，在构建五系综合症状提问系统时，需要将数据集进行预处理。本文设计了一种交叉合并法来合并数据，可以生成一份完整的五系综合问诊数据，并用此数据集来进行五系综合症状提问系统的模型搭建。在合并之前，为了避免样本数目不均导致的结果错误，将从每个系中抽取相近数量的病例数据进行合并。

交叉合并法的数据处理流程如下图1所示，不同系之间若含有同一症状标签，则将其合并为一个标签，合并完成后，可以得到一份完整的症状集合，将其作为五系统综合数据集的症状标签，在症状标签后加入每个系的证候标签，之后将每个系的样本数目进行累加，得到五系统数据集的样本总数，生成带有症状证候标签且长度为样本总数的全“0”数据集。遍历每个系的病例数据，将每个病例数据中“1”所对应的标签与数据空集中的症状证候标签进行匹配，将匹配到的在全0数据集中改为“1”，其余值依旧为“0”值，最终得到完整的五系综合问诊数据。

数据合并前，5个系的症状总数为353个症状，合并之后的症状总数为166个症状，证候数目不变，合成后的数据示例如下图2所示。

研究如何应用五系数据构建出五系综合症状提问系统，使患者在无法判断自己是属于某系疾病时，可以根据主诉直接进入到五系综合症状提问系统，在系统中全面获取到其可能患有的相关症状，为后续的证候综合诊断奠定基础。

1.2 基于频繁模式挖掘算法的症状提问策略

1.2.1 关联分析

关联分析[24-25]作为数据挖掘中的一个重要分支，主要用来挖掘数据集中特征之间的潜在关系，得出事物之间频繁出现的连接及关联关系，还可以用挖掘出的关联规则来预测事物的发展。

在中医问诊的过程中，医师通常围绕患者的主诉来展开询问，目的是为了获取患者的潜在症状，整个过程存在着很大的主观性，对这个过程使用关联规则分析，通过对临床问诊数据的挖掘，挖掘出症状的频繁项集，计算出症状之间的关联性，不仅可以用来分析症状之间的关联程度，还可以用来预测患者可能患有的潜在症状，以此用于中医临床辅助诊断，可以减少人为主观因素的影响，更有利于提高问诊效率[26-27]。在关联分析中，一般使用以下参数来衡量。

给定一个数据集合D，D中的每条记录都含有n个属性，则$D=\{T_1,T_2,\cdots,T_n\}$。I表示由m个不同的项构成的集合，则$I=\{I_1,I_2,\cdots,I_m\}$，I称为项集[28]，任一个I_i称为项目，长度为k的项集称为k-项集，二者满足关系$T_i\subseteq I$。

关联规则（Association rules）：暗示两种特征之间可能存在着很强的关系。关联规则的定义式为$A\Rightarrow B$，其中$A,B\subseteq I$，$A\neq\varnothing$，$B\neq\varnothing$，且$A\cap B=\varnothing$，其中A为关联规则的先导，B为关联规则的后继。

（1）支持度（Support）[28]：表示项集在整个数据集中

表1 部分数据集样本示例

症状						证候		
病例序号	胃脘痛	泛酸	…	恶心呕吐	呃逆	脾胃气虚	脾胃气滞	脾胃湿热
1	1	1	…	1	0	1	0	0
2	1	1	…	1	1	1	0	0
3	0	0	…	1	0	0	0	0
4	1	1	…	1	0	0	0	1
5	1	1	…	1	1	1	0	0

表2 五系数据统计结果

类型	心系	肝系	脾系	肺系	肾系
原始病例样本	1186	320	500	408	251
男性病例	385	178	228	188	98
女性病例	801	142	272	220	153
症状特征	102	85	207	251	239
最终病例样本	1157	320	475	408	238
最终症状特征	64	75	59	75	72

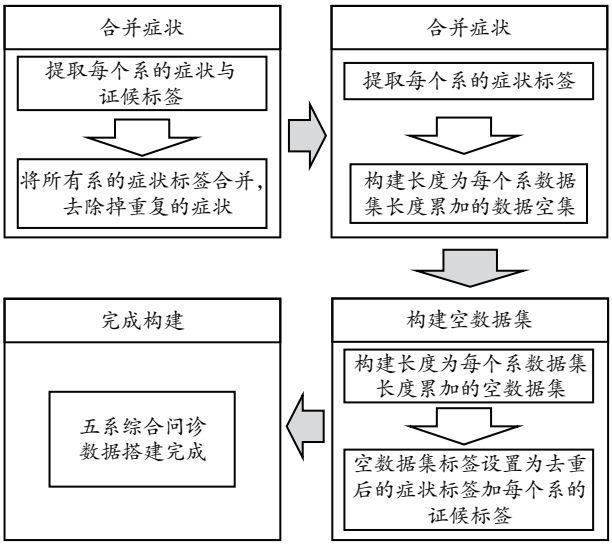

图1 交叉合并法数据处理流程图

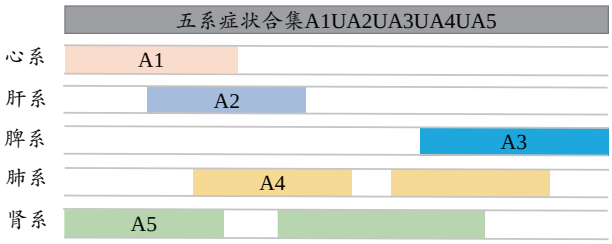

图2 合成后的数据示意图

的重要程度。如果是单个项集，表示为此项集在整个事物中出现的概率；如果是两个及以上项集，则表示这多个项集同时出现的概率。设置的最小支持度[29]为阈值。

$$\text{Support}(A \Rightarrow B) = \frac{count(A \cup B)}{|D|} \quad (1)$$

式中$count(A \cup B)$表示数据集合D中既包含A又包含B的数目，$|D|$表示总数据数。

（2）置信度/可信度（Confidence）：用来反映关联规则的可信程度，表示的是在A项集发生的情况下，B发生的概率。

$$\text{Confidence}(A \Rightarrow B) = \frac{count(A \cup B)}{count(A)} \quad (2)$$

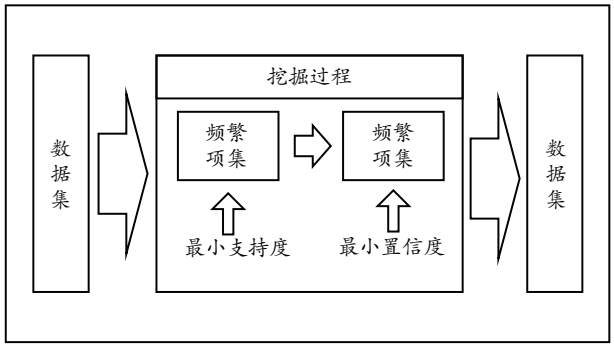

图3 关联规则挖掘流程图

（3）提升度（Lift）：表示关联规则中项集之间的相关性。

$$\text{Lift}(A \Rightarrow B) = \frac{count(A \cup B)*|D|}{count(A)*count(B)} \quad (3)$$

通常情况下，相关性研究的结果分为3种，即：①Lift>1，表示项集之间为正相关关系，且提升度的值越大，正相关性越高；②Lift=1，表示项集之间没有相关关系；③Lift<1，则表示项集之间呈负相关关系，且提升度的值越小，负相关性越高。

（4）频繁项集（Frequency item sets）：又称频繁模式，指支持度大于等于阈值的项集。

在使用关联规则[30]进行数据挖掘的过程中，主要有以下两个步骤：①获取频繁项集：遍历数据集，通过频繁项集算法，获取到支持度大于等于最小支持度的项集；②生成关联规则：通过第一步中获取到的频繁项集，找到置信度大于等于最小置信度的所有关联规则。对数据集进行关联分析数据挖掘的流程如下图3所示。

1.2.2 频繁模式挖掘算法

频繁模式挖掘算法[31]是基于频繁项集的关联分析算法。在FP-Growth算法[32]中，数据结构主要包含3个部分：项头表、FP-Tree、节点链表。项头表用来记录所有的1项频繁集出现的次数，FP-Tree是将原始数据集映射到了内存中的一颗树状结构中，整个数据结构如图4所示。

FP-Growth算法的实现原理如下[28]：①第1次遍历数据，得到包含所有项集的频繁一项集，定义最小支持度（阈值），最小支持度（MinSupport）设置为：MinSupport=测试集长度/20，在频繁一项集中剔除支持度小于阈值的项集，将频繁一项集按照支持度降序排列。②第2次遍历数据，继续剔除非频繁一项集的数据，按照支持度降序排列，通过2次遍历的数据建立

TID	Items
1	{J,L}
2	{K,L,M}
3	{J,L,M,N}
4	{J,M,N}
5	{J,K,L}
6	{J,K,L,M}
7	{K,L}
8	{J,K,L}
9	{J,K,M}
10	{K,L,N}

Item	Frequency
K	8
J	7
L	7
M	5
N	3

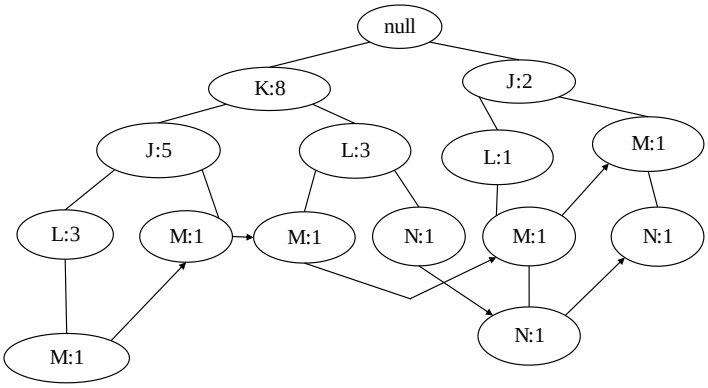

图4　FP-Growth算法数据结构图

项头表。③建立FP-Tree结构，FP-Tree为一种树状结构，按照支持度降序排列，靠前的为父节点，靠后的为子节点，节点用来存储数据，将遍历好的数据插入节点中，若有相同的链表，则公用的父节点计数加一，若没有数据对应的链表，则新建链表插入数据，将全部的数据按照顺序依次插入完成，FP-Tree构建完成。④挖掘频繁项集。按照项头表从底部挖掘项头表项对应的条件模式基，再从条件模式基依次找到项头表项对应的频繁项集，并返回频繁项得到的计数值，此外可以设置频繁项的数值，筛选出符合条件的频繁项集。

FP-Growth算法作为Apriori的改进算法，在进行数据挖掘时主要有2个步骤：①遍历数据，构建FP-Tree；②递归遍历FP-Tree，从FP-Tree中挖掘频繁项集。通过遍历2次数据集，就可以挖掘到频繁项集，提高了运行效率，对于挖掘中医问诊症状的频繁项集有着重要的意义。

1.3　基于频繁模式算法的症状提问系统设计

本文提出的中医智能症状提问系统流程如图5所示，在中医问诊中，通常会围绕患者的主诉症状来展开询问，因此在系统症状提问设计时，也会在开始提问前设置输入主诉症状，症状提问系统会根据主诉症状计算出与之相关联的症状供患者选择，每当患者选择完症状后，系统都会按照患者选择后的症状重新进行计算来提问。若开始没有主诉症状，由于没有患者的相关数据，无法直接为患者提供可能患有的症状，为了解决这一冷启动问题，将此系疾病患者中最常见的前6种症状作为首轮症状来提问，在患者选择了首轮提问的症状之后，系统则继续根据患者选择的症状计算出相关联症状向患者提问，直到所有症状提问完毕。

以心系为例，将症状提问系统在心系疾病中进行运用，搭建出症状提问系统。将整个心系数据的症状数据进行随机划分，其中70%的数据用来做训练，30%的数据用来做测试，具体过程如下：首先遍历心系疾病的数据，将数据按照症状频数的降序排列，并将数据中支持度小于阈值的症状剔除，得到项头表；再遍历整个心系的数据，与项头表中的症状进行对比，将项头表中没有的症状在心系的数据中也进行剔除，剔除之后将数据的顺序按照支持度降序排列，得到处理后的心系疾病数据。如表3所示，在整个心系疾病的数据中，出现频率最高的3个症状分别是：胸闷、心悸、气短/气急/憋气。

构建心系疾病的FP-Tree：将处理后的心系数据中的症状信息逐一读入，将读到的每条数据插入到FP-Tree中，靠前的节点为父节点，靠后的节点为子节点，若读到的数据与之前的数据有公用父节点时，对应的公用节点计数加一。若有新的症状出现时，则创建对应新的节点，直到所有的数据读取完毕，FP-Tree构建完成。

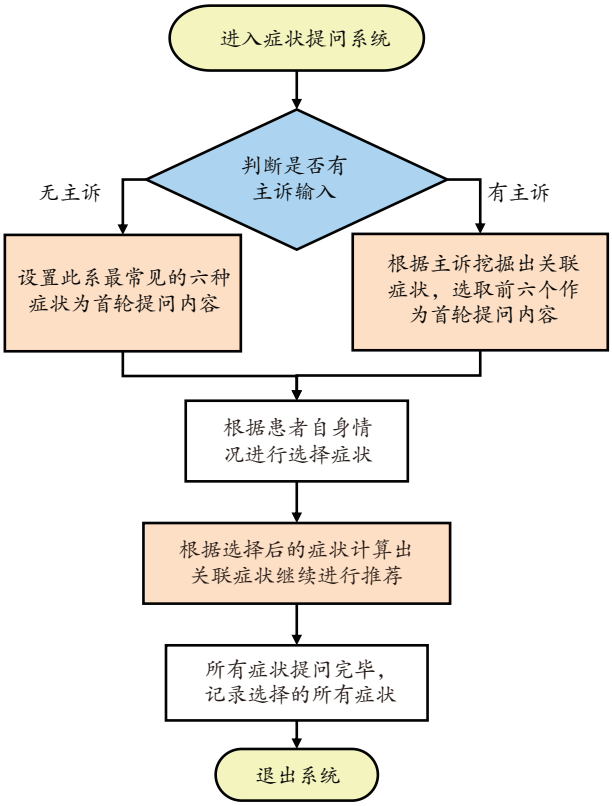

图5 症状提问流程图

表3 心系统部分项头表

症状	支持度
胸闷	0.778338
心悸	0.687657
气短/气急/憋气	0.677582
乏力懒言	0.629723
眩晕/头晕目眩	0.584383
腰膝酸软	0.581864
胸痛	0.445844

FP-Tree的挖掘：要从构建好的FP-Tree中挖掘频繁项集，需要找到条件模式基，条件模式基是指从FP-Tree的最底部的节点开始向上挖掘，寻找与最底部节点相关的节点，找到的节点为频繁项集。如图6所示，最底部头晕目眩的频繁2项集有：{头晕目眩，胸闷}、{头晕目眩，气短/气急/憋气}、{头晕目眩，心悸}，频繁3项集有：{头晕目眩，胸闷，气短/气急/憋气}、{头晕目眩，胸闷，心悸}、{头晕目眩，气短/气急/憋气，心悸}等等，依次递归，得到最大的项集为频繁4项集：{头晕目眩，胸闷，气短/气急/憋气，心悸}。表4是关联规则部分结果，第一列与第二列分别是症状特征，后面依次为症状特征之间的支持度、置信度与提升度。

表4 关联规则部分结果

先导	后继	支持度	置信度	提升度
气短/气急/憋气	胸闷	0.580357	0.888672	1.174905
心悸	胸闷	0.503827	0.812757	1.074539
乏力懒言	胸闷	0.466837	0.815145	1.077696
心悸	气短/气急/憋气	0.452806	0.730453	1.118506
眩晕/头晕目眩	胸闷	0.433673	0.772727	1.021616

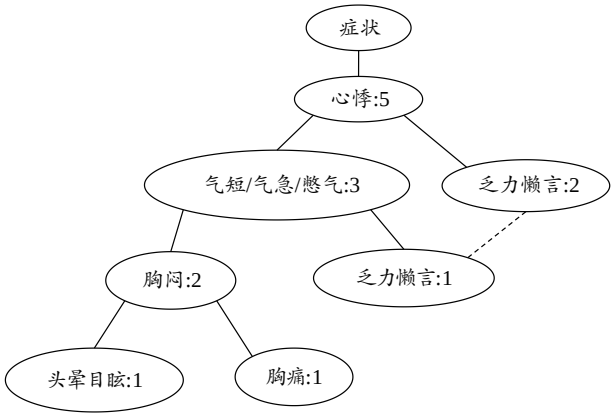

图6 FP-Tree部分树状结构图

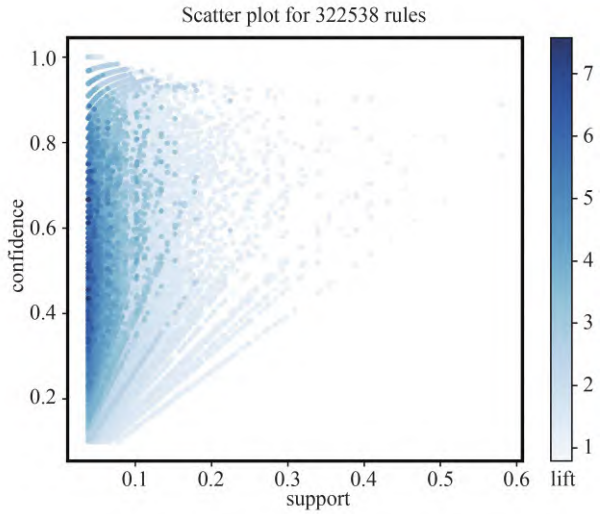

图7 心系统症状间关联规则分析散点图

根据挖掘到的频繁项集，可以找到症状之间的关联性，根据频繁项集中症状的关联性进行提问，如患者主诉是心悸时，症状会生成{头晕目眩，胸闷，气短/气急/憋气……}等相关联症状，依次提供给患者，供患者进行选择，在频繁项集提问完毕之后，将剔除掉的症状作为补充。经过多轮的症状提问后，会返回生成患者患有的症状集合，供后续的研究。

用关联规则分析的结果生成症状之间的关联规则散点图（见图7），横轴为关联规则的支持度，纵轴为

置信度，lift值的高低则由关联规则点的颜色深浅表示，从图中可以看出，心系疾病症状中可以找到30000多个关联规则，通过可视化散点图，可以分析出症状与症状之间关联规则关系，如提升度较高的关联规则的支持度往往较低，支持度与置信度具有明显反相关性，关联规则中，大多都集中在低支持度、高提升度的区域，且高提升度的关联规则大多处于置信度20%-90%。

本文对剩余4个系的数据也进行了关联规则分析，生成了关联规则分析散点图，如图8所示，图中1-4分别表示肝系、脾胃系、肺系、肾系的关联规则散点图，从图中可以看出，肝系与脾胃系生成的关联规则数目较少，肺系与肾系生成的关联规则数目较多，其中，脾胃系的散点图较为分散，其余3个与心系相差不大，这是由于脾胃系的症状数据相比于其他4个系的数据较为分散，容易造成支持度较低，在系统设置最小支持度后，满足条件的关联规则数目减少，最终生成较少的关联规则。

1.4 症状提问系统的评价指标及测试方法

为了查看症状提问系统的效果，本文使用推荐系统中常用的3个指标来评价症状提问模型，分别为：准确率（Precision）、召回率（Recall）、F1值。下面是3个指标的计算方法。

（1）准确率（Precision）：推荐给用户的商品列表A_u中，属于测试集列表L_u的占比，数学公式如下：

$$P(L_u)=\frac{L_u \cap A_u}{L_u} \tag{4}$$

整个测试集的准确率为：

$$P=\frac{1}{n}\sum_{u\in U}P(L_u) \tag{5}$$

（2）召回率（Recall）：测试集中有多少在用户的推荐列表中，公式如下：

$$R(L_u)=\frac{L_u \cap A_u}{B_u} \tag{6}$$

整个测试集的准确率为：

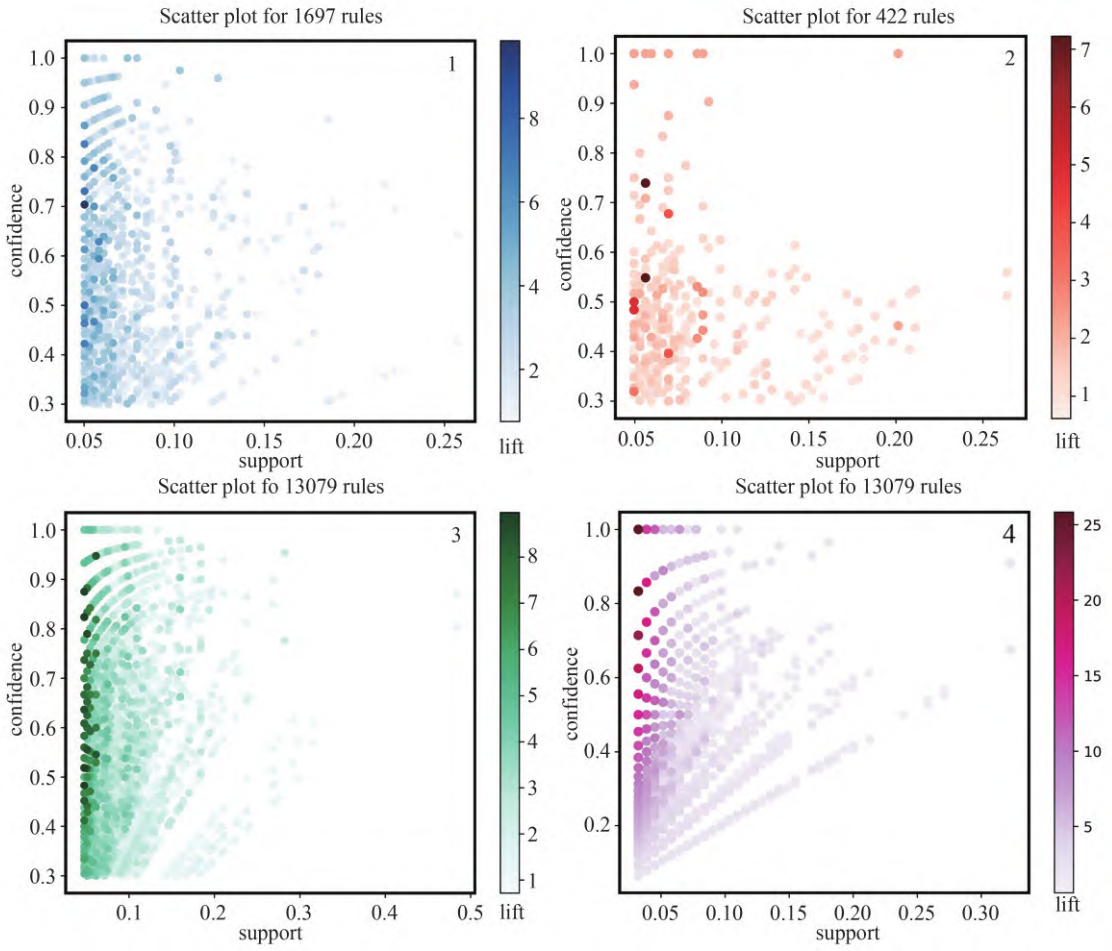

图8 关联规则分析散点图

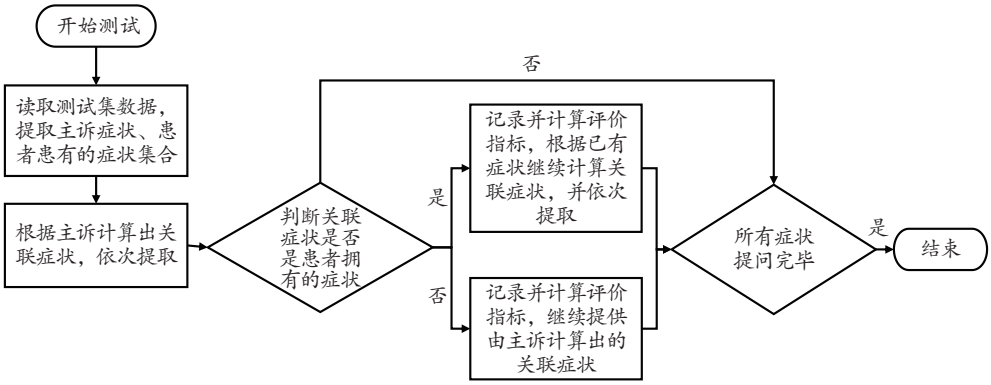

图9 测试问诊症状提问系统流程

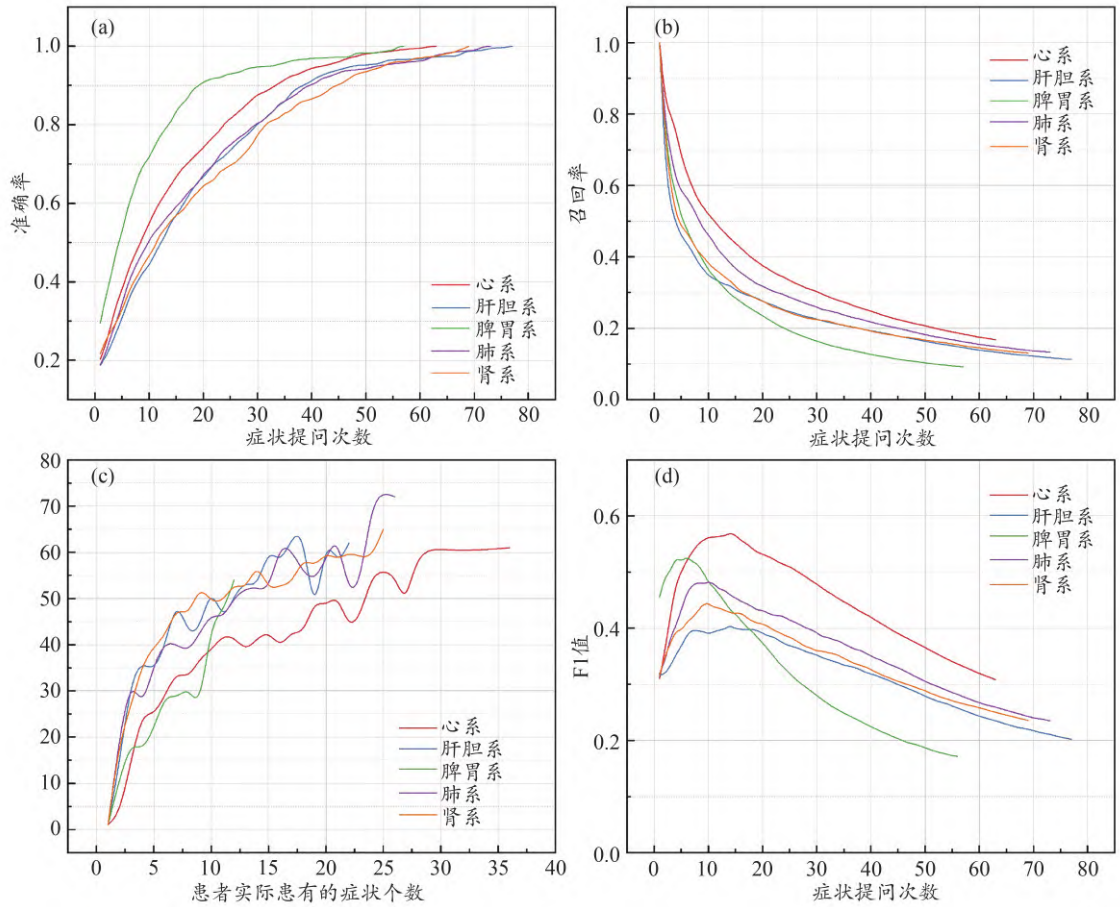

图10 单系症状提问系统测试结果图

$$R = \frac{1}{n}\sum_{u \in U} R(L_u) \tag{7}$$

(3)F1值：当准确率P与召回率R出现矛盾的情况下，需要综合考虑，这时通常使用准确率与召回率的加权调和平均，即F-Measure：

$$F = \frac{(\alpha^2 + 1)P*R}{\alpha^2(P + R)} \tag{8}$$

当参数α=1时，即为F1值，也即：

$$F1 = \frac{2*P*R}{P + R} \tag{9}$$

F1值是对准确率与召回率进行综合计算后得出的结果，当F1值较高时，则表示实验结果是有效的。

在症状提问系统中，准确率起着最为关键的作用。从准确率中可以判断出症状提问次数高效率的范围，召回率用来作为对系统的辅助评估，F1值综合了准确率与召回率的值，在系统中用来验证最终的结

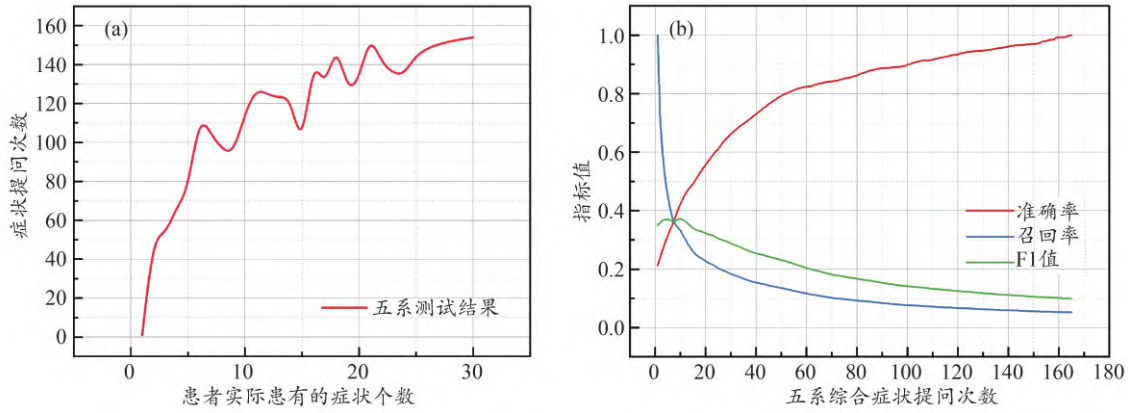

图11　五系症状提问系统测试结果图

果是否在合理的范围中。

在测试过程中，先将测试数据中的症状进行划分，分为主诉症状与询问症状，以此来模仿患者主诉与医生询问的过程，主诉症状按照单系中最常见的主诉症状来提取。将测试数据中每个患者患有的一个主诉症状作为测试过程中的输入，症状提问系统会依次给出患者可能患有的症状，将其与患者的数据进行匹配，若提问系统给出的症状中含有患者实际患有的症状，则将其记录下来，并将症状提问次数进行标记，直到症状提问完毕。测试流程如图9所示。

2　结果

根据上文提出的评价指标与测试方法，分别对心、肝、脾、肺、肾5个系进行测试，研究症状提问次数对症状提问系统效果的影响，并根据测试结果来确定出不同系的最佳症状提问次数范围，以此来提高症状提问的效率，单系症状提问系统测试结果如图10所示。

由单系症状提问系统的测试结果可以看出，系统的准确率在随着症状提问次数的增加不断升高，当提问次数到达一定程度时，曲线上升速率逐渐减慢并开始趋于平缓，在5个系统中，症状提问次数在30次左右时，症状准确率均可达到80%以上，其中脾胃系的症状准确率甚至可以达到94%以上，此方法大大提高了症状的获取效率。

系统的召回率随着症状提问次数的增加在不断下降，F1值也随着症状提问次数的增加先上升后下降，在提问次数为10-30次左右时，指标值较高，此时系统性能表现较好，由患者实际患有的症状个数与症状提问次数关系图可以看出：随着患者患有的症状数目增多，症状提问次数也在不断增加，但在患者症状个数为10-20时，有一段曲线趋于平缓，此结果表明：对于大部分患者，获取其全部症状的提问次数是在一个固定的范围之内。因此，可以根据此结果来确定症状提问次数的范围，其中，心系疾病患者症状个数在10-20个时，系统提问40次就可以获取到患者患有的全部症状，相比中医量表提问64次来获取患者的症状，此方法减少了24次就可以达到相同的效果。

图11是五系综合症状提问系统的测试结果，从图中可以看出，准确率随着症状提问次数的增加而增加，在1-50次的症状提问次数下，准确率上升的速率较快，此结果表明：在当前提问次数的范围内，系统对患者的症状获取率较高，当提问次数达到50次时，平均准确率达到了80%，即系统对患者提问50次就可以获取到患者80%的症状信息，当症状提问次数从51次增加到165次时，系统对症状的获取率仅增加了20%，症状获取率较低；系统的平均召回率随着提问次数的上升而下降，且提问次数在1-70次时，平均召回率都在0.1以上，系统的F1值随着提问次数的增加先上升后保持下降，当提问次数在1-60次时，F1值保持在0.2以上，此结果表明在当前范围内系统的性能较好。由患者实际患有的症状个数与症状提问次数关系图可以看出：患者症状个数在5-10时，症状提问系统提问80-110次就可以获取到患者的全部症状，和初始症状总数相比症状数目减少了一半，此方法大大提高了症状的获取效率。

3　结论与展望

依据中医单系问诊量表的方式能够获取到患者的症状信息，然而获取过程往往效率低、获得信息面

较窄。因此，本文从心、肝、脾、肺、肾五系问诊数据出发，使用频繁模式挖掘算法建立起症状之间的关联性，并设计出交叉合并法来搭建中医单系疾病症状提问系统与五系疾病综合症状提问系统，系统测试结果显示：单系症状提问系统仅需提问30-45次就可以获取到患者90%的症状信息，且对比传统量表的提问方式减少了65%的提问次数，大大提高了对患者症状信息获取的效率；在五系综合症状提问系统中，从原始量表的149次提问次数降低到100次来获取到患者90%的症状信息。此方法打破了中医基于量表来询问患者的传统问诊模式，简化了问诊流程，提升了中医问诊的效率，并创新地实现了由单系疾病到五系疾病的过渡，能够从更为全面的角度来获取患者的症状信息，减少了由于经验不足或人为主观造成的问诊差异，推进了中医问诊客观化的发展。然而本文当前的研究主要针对中医问诊过程中的询问部分，对诊断部分还未提及，在后续的研究中可通过提问系统获取到的症状进行证候诊断研究，从而验证提问系统的有效性、完善中医问诊过程，进一步推动中医问诊客观化的进程。

参考文献

1 董玉舒, 李慧. 中医“治未病”与现代化及客观化结合的挑战与意义. 医学信息, 2019, 32(13):147-149.

2 许朝霞, 王忆勤, 刘国萍, 等. 中医问诊客观化研究进展. 时珍国医国药, 2009, 20(10):2546-2548.

3 杨杰, 牛欣, 徐元景, 等. 中医诊断信息数字化发展. 中医药学刊, 2006, 24(5):810-812.

4 钟有东, 王峰, 邱逸铭, 等. 基于Rasa框架的中医问答系统设计. 电脑知识与技术, 2022, 18(11):74-76.

5 Wong W, Cindy Lam L K, Li R, *et al*. A comparison of the effectiveness between Western medicine and Chinese medicine outpatient consultations in primary care. *Complement Ther Med*, 2011, 19(5):264-275.

6 韩文博, 王凯, 周沪方, 等. 中医风邪客观化研究进展. 中西医结合心脑血管病杂志, 2022, 20(3):455-459.

7 迪盼祺, 夏春明, 王忆勤, 等. 基于协同过滤算法的中医智能问诊系统研究. 世界科学技术-中医药现代化, 2021, 23(1):247-255.

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

9 王忆勤, 郭睿, 颜建军, 等. 基于多标记学习的中医问诊智能系统. 全国第十二次中医诊断学术年会论文集. 银川, 2011:14-19.

10 洪婕, 顾捷飞, 钟臻, 等. 基于大数据挖掘的名老中医智能化传承系统的设计与探索. 中医药管理杂志, 2021, 29(23):337-338.

11 王俊文, 叶壮志. 人工智能技术在中医诊断领域应用述评. 世界科学技术-中医药现代化, 2022, 24(2):810-814.

12 Korngold E, Beinfield H. Chinese medicine and the mind. *Explore*, 2006, 2(4):321-333.

13 何裕民, 刘文龙. 新编中医基础理论. 北京:中国协和医科大学联合出版社, 1996:3-28.

14 王丽婷, 刘长松, 魏玮, 等. 中医学术客观化、智能化传承的思考及初步实现. 中医杂志, 2021, 62(12):1036-1039.

15 宋诗博, 安二匣, 樊西倩, 等. 中医四诊合参客观化研究思考. 中华中医药杂志, 2021, 36(11):6560-6562.

16 樊亚东, 白立鼎, 常军, 等. 心血管疾病中医证候客观化研究进展. 中华中医药学刊, 2021, 39(10):172-176.

17 汪南玥, 刘佳, 宋诗博. 基于人工智能的中医多诊合参技术研究现状与展望. 中华中医药杂志, 2022, 37(1):41-44.

18 黄玮, 余江维. 中医问诊内容及客观化研究探析. 中华中医药杂志, 2019, 34(8):3666-3668.

19 王宗殿. 中医问诊. 合肥: 安徽科学技术出版社 1990:1-53.

20 罗思言, 王心舟, 饶向荣. 人工智能在中医诊断中的应用进展. 中国医学物理学杂志, 2022, 39(5):647-654.

21 刘海婷, 贺习婷. 基于“中医四诊”的高血压病客观化研究概况. 中国民间疗法, 2022, 30(7):120-122.

22 侯春蕾, 许颖, 崔延婕, 等. 2型糖尿病中医问诊量表的研制及临床应用概述. 世界科学技术-中医药现代化, 2021, 23(2):396-401.

23 钟森杰, 李静, 李琳, 等. 脾气虚证的诊断标准及客观化研究述评. 时珍国医国药, 2021, 32(2):421-423.

24 Agarwal S. Data mining: data mining concepts and techniques. 2013 International Conference on Machine Intelligence and Research Advancement. Katra, India. IEEE, 2013:203-207.

25 Earthy J, Jones B S, Bevan N. The improvement of human-centred processes—facing the challenge and reaping the benefit of ISO 13407. *Int J Hum Comput Stud*, 2001, 55(4):553-585.

26 胡恒昶, 莫沙, 苏可, 等. 中医大脑-人工智能在中医临床中的创新性应用. 中国民间疗法, 2021, 29(19):90-93.

27 苏园园, 刘宁宁, 赖优莹, 等. 胃癌中医证型研究进展. 陕西中医, 2023, 44(2):262-266.

28 Han, Jiawei ;Pei, Jian ;Kamber, Micheline. Data Mining: Concepts and Techniques. Burlington:Morgan Kaufmann_RM, 2011:230-522.

29 BING L, HSU W, MA Y. Mining association rules with multiple minimum supports. ACM Special Interest Group on Knowledge Discovery and Data Mining International Conference on Knowledge Discovery and Data Mining. San Diego. CA, 1999:337-341.

30 Agrawal R, Imieliński T, Swami A. Mining association rules between sets of items in large databases. Proceedings of the 1993 ACM SIGMOD international conference on Management of data. Washington D.C. USA. ACM, 1993:207-216.

31 Han J, Pei J, Yin Y, *et al*. Mining frequent patterns without candidate generation: a frequent-pattern tree approach. *Data Min Knowl Discov*,

2004, 8(1):53-87.
32 Totad S G, Geeta R B, Prasad Reddy P V G D. Batch incremental processing for FP-tree construction using FP-Growth algorithm. *Knowl Inf Syst*, 2012, 33(2):475-490.

A Study of Chinese Medicine Consultation Strategies Based on Frequent Pattern Mining Algorithms

LI Ruizhen[1,2], XIA Chunming[2,3], WANG Yiqin[4], XU Zhaoxia[4], XIONG Yujie[3]

(1. Northwestern Polytechnical University, School of Civil Aviation, Xi'an 710129, China; 2. School of Mechanical and Power Engineering, East China University of Science and Technology, Shanghai 200237, China; 3. School of Electrical and Electronic Engineering, Shanghai University of Engineering Science, Shanghai 201620, China; 4. Shanghai University of Traditional Chinese Medicine, Shanghai Key Laboratory of Health Identification and;Assessment/Laboratory of Traditional Chinese Medicine Four Diagnostic Information, Shanghai 201203, China)

Abstract: Objective To study Chinese medicine consultation strategies to achieve rapid capture of key information about patients' conditions and to advance the development of objectification in Chinese medicine consultation. Methods A symptom questioning model based on frequent pattern mining algorithm in correlation analysis was used, and a cross-merging method was used to establish a TCM symptom questioning strategy between single-system symptom questioning and multi-system integrated symptom questioning in TCM, to achieve the shortest time and highest efficiency in capturing key information about the patient's condition. Results A breakthrough from single-system questioning to five-system integrated questioning was achieved, and the process of efficiently obtaining information about the patient's condition was achieved through both single-system and five-system symptom questioning modes, and the system was able to obtain 92% of the patient's symptom information with at most 65% fewer questions than the traditional scale questioning method, greatly improving the efficiency of obtaining information about the patient's symptoms. Conclusion With the two different symptom questioning modes, the traditional TCM questioning mode of asking patients based on scales is broken, the time to obtain symptoms from patients is shortened, the questioning process is simplified, and discrepancies due to inexperience or human subjectivity are reduced, which can be used in clinical aids to diagnosis in TCM.

Keywords: Chinese medicine consultation, Frequent pattern mining algorithm, Symptom correlation, Consultation strategy

（责任编辑：刘玥辰）