Journal of Computer Applications ISSN 1001-9081 2017-05-10
计算机应用,2017,37(5):1424 - 1429,1465 CODEN JYIIDU http://www.joca.cn

文章编号:1001-9081(2017)05-1424-06 DOI:10.11772/j.issn.1001-9081.2017.05.1424

# 最大模糊频繁模式挖掘算法

张海清[1],李代伟[1],刘胤田[1],龚 程[1],于 曦[2*]

(1.成都信息工程大学 软件工程学院,成都 610225; 2.成都大学 信息科学与工程学院,成都 610106)

(* 通信作者电子邮箱 yuxi@cdu.edu.cn)

**摘 要:** 针对有效模式挖掘的组合爆炸及挖掘结果信息如何有效表达的问题,提出了一种基于"核心-牵引"结构的修剪候选模式和考虑项目不确定性的最大模糊模式挖掘算法(MFFP-Tree)。首先,综合分析项目的模糊性,提出模糊支持度,分析项目在事务数据集中的模糊权重,依据模糊修剪策略修剪候选项集;其次,仅扫描数据集一次,就能成功构建模糊模式挖掘树,依据模糊剪枝策略减少模式提取的开销,采用 FFP-array 阵列结构使得搜索方式更精简,从而进一步降低时空开销。根据基准数据集的实验结果,与最大模式挖掘算法 PADS 和 FPMax* 对比分析,MFFP-Tree 挖掘出的最大模糊模式能够更准确地反映项目与项目之间的关系;算法的时间复杂度能减半甚至低 1 个数量级;算法的空间复杂度降低 1 ~ 2 个数量级。

**关键词:** 高级模式挖掘;最大模糊模式;模糊支持度;核心-牵引模式结构;模糊修剪策略

**中图分类号:** TP311.1 **文献标志码:** A

## Mining algorithm of maximal fuzzy frequent patterns

ZHANG Haiqing[1], LI Daiwei[1], LIU Yintian[1], GONG Cheng[1], YU Xi[2*]

(1. *College of Software Engineering, Chengdu University of Information Technology, Chengdu Sichuan* 610225, *China*;
2. *College of Information Science and Engineering, Chengdu University, Chengdu Sichuan* 610106, *China*)

**Abstract:** Combinatorial explosion and the effectiveness of mining results are the essential challenges of meaningful pattern extraction, a Maximal Fuzzy Frequent Pattern Tree Algorithm (MFFP-Tree) based on base-(second-order-effect) pattern structure and uncertainty consideration of items was proposed. Firstly, the fuzziness of items was analyzed comprehensively, the fuzzy support was given, and the fuzzy weight of items in the transaction data set was analyzed, the candidate item set was trimmed according to the fuzzy pruning strategy. Secondly, the database was scanned once to build FFP-Tree, and the overhead of pattern extraction was reduced based on fuzzy pruning strategy. The FFP-array structure was used to streamline the search method to further reduce the space and time complexity. The experimental results gained from the benchmark datasets reveal that the proposed MFFP-Tree has outstanding performance by comparing with PADS and FPMax* algorithms: the time complexity of the proposed algorithm is optimized by twice to one order of magnitude for different datasets, and the spatial complexity of the proposed algorithm is optimized by one order of magnitude to two orders of magnitude, respectively.

**Key words:** advanced pattern mining; maximum fuzzy pattern; fuzzy support; base-(second-order-effect) pattern structure; fuzzy pruning strategy

## 0 引言

大规模数据集中挖掘潜在有用但隐藏的信息是模式挖掘的主要目标。传统的模式挖掘方法,主要包括 Apriori[1] 和 FP-growth[2] 算法,并且这两种算法的特征和性质已经被广泛应用到其他改进的关联规则的研究中[3-5]。随着数据集的大规模增长,具有更高算法性能和满足更多目标需求的算法不断被提出,其中包括挖掘序列频繁模式[6-7]、基于(无)阈值约束的 Top-$K$ 频繁模式[8-9]、基于数据流的频繁模式[10-11]和基于加权的频繁模式[12]等。上述频繁模式挖掘方法均基于传统的频繁模式的先验性质,频繁项集的所有非空子集也一定是频繁的,并且挖掘出的模式遵守约束条件,项目出现的频度必须要大于指定阈值;然而,根据分析医疗大规模数据的实践经验得知,具有实践指导意义的模式通常是相对频繁的项目和出现频率相对较低的项目的组合。例如,针对一个病人的诊断项目,病人所患的疾病通常涉及多个不同的科室,并且单个病人的患病特征集合一般由常见病特征和该病人"个性化"的疾病特征组成。因此,为了阐述大规模数据集所隐含的模式的复杂性,对频繁项目和相对不频繁的项目应该综合分析。本文主要目的是为了发现与该疾病密切相关的其他疾病或者由该疾病最易诱发的其他疾病,而不仅仅是给出常见疾病之间的关联性。

本文旨在规避挖掘具有欺骗性和误导性的传统模式的基

收稿日期:2016-10-08;修回日期:2016-12-23。

基金项目:国家自然科学基金青年基金资助项目(61602064,61502059);成都信息工程大学科研基金资助项目(KYTZ201615)。

作者简介:张海清(1986—),女,山东聊城人,讲师,博士研究生,主要研究方向:大数据分析; 李代伟(1976—),男,四川达县人,副教授,硕士研究生,主要研究方向:数据集成与可视化、机器学习; 刘胤田(1972—),男,四川隆昌人,教授,博士研究生,主要研究方向:数据挖掘; 于曦(1973—)男,吉林长春人,副教授,博士研究生,主要研究方向:决策系统、神经网络。

础上,在大规模动态数据集中提取最具代表性的模式。根据对文献[1-11]的分析,如何提取最有效的最具代表性的模式并没有得到良好论证,并且依据传统强关联规则所生成的一些频繁模式、闭模式以及极大模式具有欺骗性和误导性。相对传统的频繁模式挖掘,本文模式具有以下特性: 本文中的项目的权重呈现模糊化而非精确值; 本文挖掘的项目的独立性以及提取的规则的形式不相同; 本文所关注的项目规则对应的网络和层次结构不相同。根据文献[13-14],项目的不确定性因素分析、数据的模糊化处理目前较为成功的解决方案为粗糙集理论和基于模糊集的模糊操作。而本文关注的不确定性问题是项目的模糊程度而非不可分辨关系,因此,采用模糊隶属度函数描述项目与项目之间在单条事务中的关系、项目在整个集合中的相对关系、事务与整个集合的关系,从而,在理论和实验上研究面向大规模医疗信息数据中隐藏的有价值的基础模式,以此为依据引入核心项、牵引项以及模糊加权模式的概念,构建模糊最佳频繁模式挖掘模型。

## 1　模糊模式结构定义

根据医疗数据集的特征分析,患者在一段时间内通常具有若干项主干疾病(核心项)和若干项由核心项所牵引的二阶效应的项(牵引项)。例如,老年患者的疾病项目是〈慢性咽炎,淋巴细胞百分数升高,消化不良,慢性支气管炎〉,根据治疗数据,该患者的慢性咽炎具有较高的危险等级,其他项目均为该项目的作用下所产生的二阶效应项目。因此,本文挖掘的模糊模式定义为核心项(base pattern, bp)和牵引项(second order effect pattern, sop)的组合。

根据核心项和牵引项之间的关系,挖掘的模糊模式的结构主要包含两类: 1) 所有特定的核心项目和全部(或者部分)牵引项一起出现。核心项目具有很高的模糊权重,从而具备较强吸附能力来吸附具有较低模糊权重的牵引项。2) 部分特定的核心项和全部(或者部分)牵引项一起出现。核心项中某些项不具有较高的模糊权重,只有部分的核心项具有吸附牵引项的能力,但是规则模式挖掘还是应该考虑不发生的核心项对整个核心项和整个事务的影响,因为不发生的核心项可能会减少或者改变核心项目的吸附能力以及吸附其他项目的活跃性。例如,在诊断患者出现严重流感现象时,即使在一段时间内病人并未出现发热的情况,医疗记录中还是要求必须标记病人的体温状况,同时该体温项目也对其他的核心项有重要的影响。基于模糊模式的两类结构,本文给出模糊模式(Fuzzy Frequent Pattern, FFP)的定义。

**定义 1**　模糊模式(FFP)。根据以上分析,本文挖掘的模糊模式可以定义为两类: 核心项(base pattern, bp)全部出现和其所吸附的牵引项(second order effect pattern, sop); 核心项部分出现(bp)、核心项中未出现的项(¬ bp),以及出现的核心项所牵引的项(sop)。模糊模式因此可被定义为式(1):

$$FFP = \begin{cases} \left(\bigcup_{i=1}^{n} bp_i\right) \cup \left(\bigcup_{i=1}^{m} sop_i\right) \\ \left(\bigcup_{i=1}^{x} bp_i\right) \cup \left(\bigcup_{i=1}^{n-x} \neg\, bp_i\right) \cup \left(\bigcup_{i=1}^{m} sop_i\right) \end{cases} \tag{1}$$

**定义 2**　模糊模式的模糊支持度($SUP_P$)。定义模式 $P = \{i_1, i_2, \cdots, i_i, \cdots i_n\}$,那么对于事务集 $T_i$ 中每一个项目 $i_i$ 在模式 $P$ 中的权重定义为: $\tilde{W}(T_i) = \{\tilde{w}_1(T_1), \tilde{w}_2(T_2), \cdots, \tilde{w}_i(T_i), \cdots, \tilde{w}_n(T_n)\}$,对于项目 $i_i$ 在总的项目集 $I$ 中的权重记为 $\tilde{V} = \{\tilde{v}_1, \tilde{v}_2, \cdots, \tilde{v}_i, \cdots, \tilde{v}_n\}$,其中 $\tilde{W}$ 和 $\tilde{V}$ 的取值属于模糊值。模糊模式的模糊支持度($SUP_P$)计算如式(2) ~ (3):

$$\tilde{w}_i = \left[\sum_{j=1}^{|T_i|} \tilde{w}_j(T_j)\right] / |T_j| \tag{2}$$

$$SUP_P = \left[\sum_{i=1}^{n} \tilde{w}_i \otimes \tilde{v}_i\right] / n \tag{3}$$

模糊模式的模糊支持度 $SUP(FFP)$ 是一个三角隶属度函数,被描述为: $SUP(FFP) = (SUP^{L}(FFP), SUP^{M}(FFP), SUP^{U}(FFP))$。其中,上标"L,M,U"是模糊函数所对应的上界、中间值和下界。标示"¬"指的是该项目不与其他的项目同时出现。例如,$\left(\bigcup_{i=1}^{x} bp_i\right) \cup \left(\bigcup_{i=1}^{n-x} \neg\, bp_i\right) \cup \left(\bigcup_{i=1}^{m} sop_i\right)$ 表示所有的在集合 $\left(\bigcup_{i=1}^{n-x} \neg\, bp_i\right)$ 中的元素不和集合 $\left(\bigcup_{i=1}^{x} bp_i\right) \cup \left(\bigcup_{i=1}^{m} sop_i\right)$ 中的元素在同一个事务中同时发生。模式 FFP 的出现必须满足式(4)所示的约束条件。

$$\begin{cases} 1)\ \theta \leqslant \sigma \leqslant SUP^{L}\left(\bigcup_{i=1}^{n} bp_i\right) \\ 2)\ \theta \leqslant SUP^{L}\left(\bigcup_{i=1}^{n} sop_i\right) \leqslant SUP^{U}\left(\bigcup_{i=1}^{n} sop_i\right) \leqslant \sigma \\ 3)\ \theta \leqslant SUP^{L}\left(\left(\bigcup_{i=1}^{n} bp_i\right) \cup \left(\bigcup_{i=1}^{m} sop_i\right)\right) \leqslant \\ \qquad SUP^{U}\left(\left(\bigcup_{i=1}^{n} bp_i\right) \cup \left(\bigcup_{i=1}^{m} sop_i\right)\right) \leqslant \sigma \\ 4)\ \varepsilon \leqslant SUP^{L}\left(\left(\bigcup_{i=1}^{x} bp_i\right) \cup \left(\bigcup_{i=1}^{n-x} \neg\, bp_i\right) \cup \left(\bigcup_{i=1}^{m} sop_i\right)\right) \leqslant \\ \qquad SUP^{U}\left(\left(\bigcup_{i=1}^{x} bp_i\right) \cup \left(\bigcup_{i=1}^{n-x} \neg\, bp_i\right) \cup \left(\bigcup_{i=1}^{m} sop_i\right)\right) \leqslant \\ \qquad \sigma\ (x \leqslant n) \end{cases} \tag{4}$$

其中: 核心项 $bp$ 满足的最小支持度阈值为 $minsup$,核心项需要满足的最小模糊权重阈值为 $\sigma$,参数 $min\_connect\_sup$ 用来定义核心项和二阶效应项目之间的边界,$\theta(\theta \leqslant \sigma)$ 表示 $sop$ 项目集的最小模糊权重阈值,$\varepsilon$ 定义为调节参数以根据挖掘模式数量的需要来个性化的设置变量变化范围。其中,$minsup$、$\theta$、$min\_connect\_sup$、$\sigma$ 的取值与传统的支持度取值方式相同,针对不同的数据集,无法理论证明最佳的参与阈值,参数阈值均根据实验数据来确定,本文的参数阈值分析见第 3 章。

最大模糊模式是模糊模式中没有超集的项。挖掘最大模糊模式需要首先定义模糊模式挖掘树。模糊模式挖掘树反映事务数据集内部之间的关联关系,并且将其挖掘结果记为 FFP 模式。FFP-Tree 管理和维护数据库中不断增加的事务集以及与之相关的项目综合权重信息。本文将保留所有的核心项目,如果核心项的模糊权重小于阈值,那么该项目将会采用"¬"项的方式插入到 FFP-Tree 中,保留模糊权重小于阈值的项目的原因是这些项目仍然对核心所吸附的项目有影响,不出现的核心项目仍然会影响核心所吸附的项或者影响核心的吸附能力。除核心项以外的项目将会采用项目的模糊权重来判断项目是否出现以及项目出现的强度。综上,FFP-Tree 的结构定义见定义 3。FFP-Tree 的构建流程见图 1。

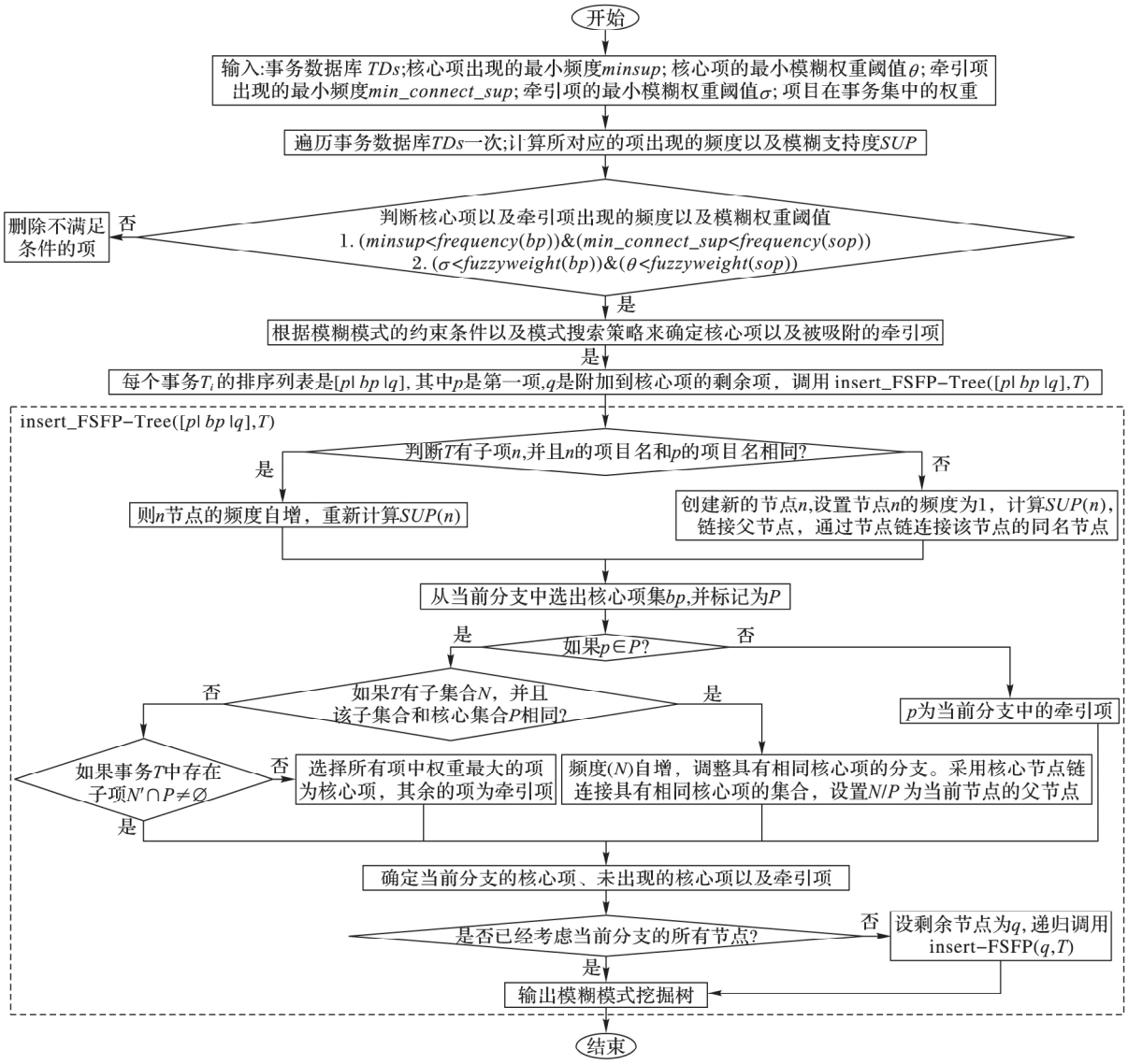



**图 1 FFP-Tree 的构建流程**

Fig. 1 FFP-Tree construction process

**定义 3** FFP-Tree(**模糊模式挖掘树**)。**模糊模式挖掘树的结构包含以下 4 个部分:**

1) **头节点,标记为**“Root”。

2) **每个节点包含 7 个字段: 项目名**(item-name)、**当前分支**(branch-level)、**父节点**(parent)、**子节点**(children)、**节点链**(node-link)、**模糊支持度**(fuzzy support)、**出现频度**(count number) **和核心节点链接**(node-link-base)。**所有共享同一个节点名的节点用节点链连接,凡是包含相同核心项的分支采用自底向上的方式由核心节点链连接,并且事务项的综合模糊度来自于所有节点的综合模糊度和出现频度的组合计算。为了表示每个项目的出现频度,频度数**(count number) **也作为一个字段。特别地,头表当中的出现频度表示了每一个项目在树中出现的总频数,在** FFP-Tree **中节点出现的频数是该节点在当前路径上的出现频数。**

3) **核心节点项目集**(baseItems)。**该字段主要用来记录当前核心项目的信息,包含: 当前核心项目名、当前未发生的核心项目、核心项目的频数、模糊支持度以及核心节点链**(node-link-base) **的头表。**

4) **项目的头表**(header table)。**头表主要放置项目集并且依据项目的模糊度值来降序排列。头表主要包含两个字段: 头表名**(item-name) **和节点链的头节点**(head of node-link)**,并且该节点链由同一个节点名的链接来连接。**

## 2 最大模糊频繁模式挖掘算法

**最大模糊频繁模式**(Maximal Fuzzy Frequent Pattern, MFFP) **挖掘算法见算法 1。最大模糊模式挖掘应该提供的参数有: 模糊支持度值**(fuzzy support value)、**核心项**(base patterns)、FFP-Tree **和基于** FFP-Tree **的阵列结构**(FFP-array)。FFP-Tree **的结构定义、核心项集的选择策略、项目的模糊度值、以及项目的出现频率阈值均作为最大模糊模式挖掘树的优化剪枝策略。其中,最大模糊模式核心项和牵引项的出现需要以下约束条件同时成立:** $(minsup < frequency(bp))$、$(min\_connect\_sup < frequency(sop))$、$(\sigma < fuzzyweight(bp))$、$(\theta < fuzzyweight(sop))$**,并且该条件在** FFP-Tree **的构建时就被应用。因此,算法 1 不需要再次筛选核心项和牵引项。**

**算法 1 最大模糊模式挖掘**(MFFP-Tree Mining) **算法。**

Input: 事务数据集 $TDs$, 允许出现的最小频度 $minmum\_count\_number$,项目的最小模糊支持度 $\theta$
Output: MFFPs
BEGIN
1) 计算模糊支持度 $SUP(i)$ 并且重新对项目集合按照降序排列;
2) 采用动态模糊修剪策略来确定项目的核心项集 $bp$;
3) 构建基于 $TDs$ 数据集和核心项的 FFP-Tree;
4) 构建基于阵列结构和条件模式基的 FFP-array;
5 if 路径 $p_i$ 是单路径 then
6) 生成新的模式 $np_i$( 通过检查当前路径上的 $bp_i$( all of the items $\{i\}$ in path $p_i$)
7) if( $SUP(np_i) \geqslant \theta$ and superset_check( $np_i$) is false)
8) $MFFP = MFFP \cup np_i$;
9) else
10) 记录更新 $MFFP = MFFP \cup bp_i$;
11) else
12) for each item $a_i$ in $TDs.header$
13) 基于 FFP-array 结构,在 $a_i$' s 的条件模式树上生成新的频繁项集 $sfi$;
14) 基于项目的模糊度对生成的 $sfi$ 按照降序排序
15) Call MFFP Mining( $sfi$, $minmum\_count\_number$, $\theta$) ;
16) endfor
17) endif
18) END

根据算法 1,如果当前路径是单路径( 算法 1 中的第 5) 行),则通过对当前路径的项目超集检测和当前项目的模糊支持度最小阈值检测以确定新的 $np_i$ 模式。如果计算的模糊支持度大于等于最小阈值并且当前求取的模式并无超集,那么此时产生的MFFP模式即为求取的最大模糊模式( 算法1中的第 6) ~8) 行);否则,当前求取的 MFFP 模式并不满足最大模糊模式的求取条件,算法则选取具有强吸附力的核心项集作为当前路径的最大模糊模式( 算法 1 中的第 10) 行)。对于多路径,基于 FFP-array 结构生成条件模式树,并基于模糊度值对项目进行降序排列,然后依据项目头表对新产生的核心项设置模糊度值,并递归调用该函数直到产生单路径( 算法 1 中的第 12) ~16) 行)。

给定事务数据表 1,依据模糊模式的构建流程( 图 1),构建 FFP-Tree( 见图 2)。

| Item | Crsip | count number | head of node-link |
|---|---|---|---|
| j | 1.655 | 2 | |
| m | 1.545 | 3 | |
| h | 1.314 | 3 | |
| b | 0.915 | 4 | |
| p | 0.876 | 2 | |
| i | 0.288 | 2 | |
| o | 0.204 | 2 | |
| {m,b,o} | 1.495 | 2 | |
| {h,b,o} | 1.290 | 1 | |
| {h,b,¬o} | 1.290 | 1 | |
| {m} | 1.545 | 1 | |

Base patterns

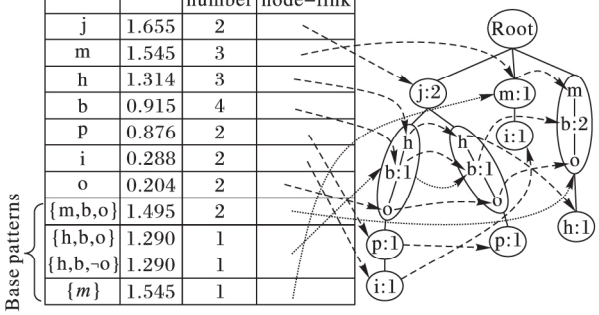



图 2　基于核心项集的 FFP-Tree
Fig. 2　FFP-Tree construction based on base patterns

基于算法 1,得到最大模糊模式为 〈j,( h,b,o) 〉、〈 m,b,o) 〉, 其中,( h,b,o) 、( m,b,o) 为分支核具有较强的吸附力,且对其他项具有较强的吸附力。然而,依据传统的最大频繁模式挖掘算法[1-5,15] ,仅能够挖掘得到最大频繁模式 〈j〉、〈m,b,o〉,并且挖掘得到的最大频繁模式也不能够反映项目与项目之间的重要关系。

表 1　在 $\theta = 0.2$ 和 $minmum\_count\_number = 2$ 之下的样例数据集
Tab. 1　Sample Database under $\theta = 0.2$ and $minmum\_count\_number = 2$

| TID | Transaction | Item sorted based on Definition 2( 模糊支持度) | count number |
|---|---|---|---|
| 1 | b,h,i,o,p,j | j,h,b,p,i,o<br>1.655,1.314,0.915,0.876,0.288,0.204 | b,h,o,i,j,p<br>4,3,3,2,2,2 |
| 2 | c,d,g,i,m,n,e,s | m,d,e i,n,c,s<br>1.545,0.765,0.669,0.288,0.205,0.136,0.136 | m,c,i,s,d,e,n<br>3,2,2,2,1,1,1 |
| 3 | a,b,l,m,o | m,b,a,o,l<br>1.545, 0.915, 0.578, 0.204, 0.068 | b,m,o,a,l<br>4,3,3,1,1 |
| 4 | b,h,m,o | m,h,b,o<br>1.545, 1.314, 0.915, 0.204 | b,h,m,o<br>4,3,3,3 |
| 5 | b,c,h,p,s,r,x,y,j | j,h,y,b,r,x,p,c,s<br>1.655, 1.314, 1.04, 0.915, 0.89, 0.89, 0.876, 0.136, 0.136 | b,h,c,j,p,s,r,x,y<br>4,3,2,2,2,1,1,1 |

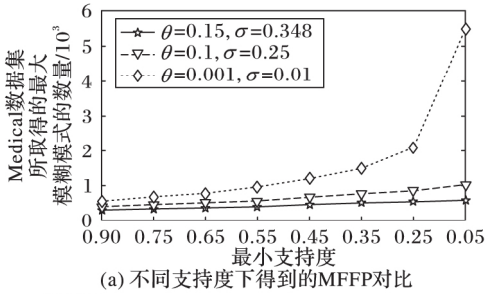
## 3　实验结果分析

为了验证本文算法的有效性,本章对比分析 MFFP 算法与传统最大频繁模式挖掘 PADS 和 FPMax* 算法的时间复杂度和空间复杂度。由于频繁模式算法挖掘的时空复杂度通常由几部分组成,而每个算法的组成部分均不相同,故通常对比算法的关键部分。例如,FP-Tree 算法的时间复杂度包含: 条件模式基、构造头表、构造 FP-Tree 和 FP-growth 挖掘,而 FP-growth 是整个算法的核心,故进行算法性能分析时分析的是 FP-growth 的性能。同样地,本文对比最大模糊模式算法、FPMax* ,以及 PADS 的核心部分。对比的数据集包括真实数据集: Chess、Mushroom,以及人工数据集 T10I4D100K 和 T40I10D100K( http: //fimi. ua. ac. be/data/) ,同时本文提供了一个新的 Medical 数据集( http: //medical. witaction. com: 808/medical) ,该数据属于稀疏数据集并且包含真实病人检测出的疾病事务项。表 2 给出了数据集的特征。算法实验平台均采用 2.20 GHz Pentium i7-3632QM 处理器,8 GB 内存,700 GB 硬盘,操作系统为 Windows 7, 所有算法均是采用 C++ 语言实现并在 Microsoft Visual Studio 2010 下面编码实现。

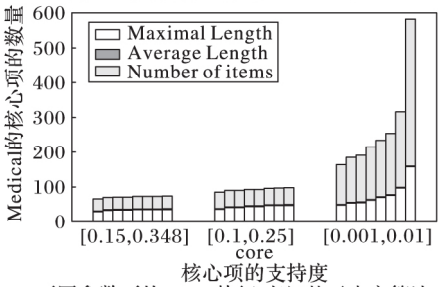
表 2　事务数据集的特征描述
Tab. 2　Transaction dataset description

| 数据集分类 | 数据集 | 事务数量 | 项数量 | 平均长度 | 最大长度 |
|---|---|---|---|---|---|
| 稠密 | Chess | 3 196 | 76 | 37 | 37 |
| | Mushroom | 8 124 | 120 | 23 | 23 |
| 稀疏 | T10I4D100K | 100 000 | 1 000 | 10 | 29 |
| | T40I10D100K | 100 000 | 943 | 39 | 77 |
| | Medical | 6 341 | 698 | 6 | 24 |

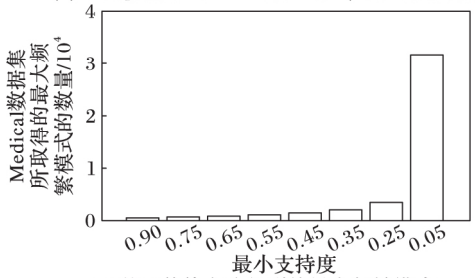
首先对 Medical 数据集挖掘结果进行分析(见图3),当允许出现的核心项的最小模糊支持度($\sigma$)和允许出现项的支持度($\theta$)间距增大时,那么挖掘出的医疗数据集会大幅度增加。



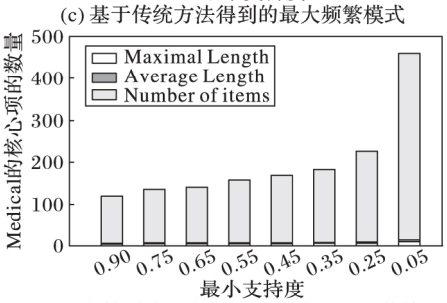

(a) 不同支持度下得到的MFFP对比




(b) 不同参数下的MFFP特征对比(基于本文算法)




(c) 基于传统方法得到的最大频繁模式



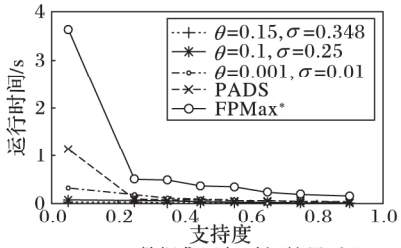

(d) 不同参数下的最大频繁模式对比(基于传统方法)

**图3 最大模糊模式与传统最大频繁模式挖掘实验结果对比**

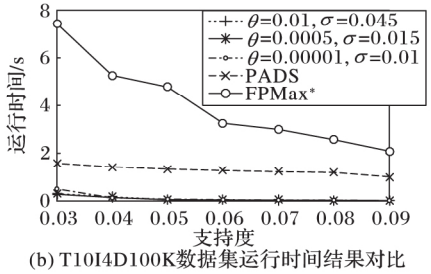
Fig. 3 Comparision of the experiment results between obtained MFFPs and obtained patterns from the traditional frequent pattern mining

当 $\sigma = 0.6$ 和 $\theta = 0.15$ 时,挖掘出的最大模糊模式的数量将会达到最高点,但是此时会产生一定数量的"假"模糊模式,因此,需要修改约束条件以提高挖掘模糊模式的有效性。同时,在参数 $\sigma = 0.348$ 和 $\theta = 0.05$ 时,挖掘出的最大模糊模式的数量和质量是最佳的。同样,通过探测修改参数阈值,该算法能够探测到最佳挖掘点并且为其他的数据集挖掘到最大模糊模式。根据实验结果分析,可以得出一般结论,对于稠密型数据集、核心项集和最大模糊模式的出现相对集中(特别是 Chess 数据集)。该研究发现稠密数据集所隐藏的规律是较为集中和稳定的,且稠密数据集的挖掘跟项目出现的频度有强相关性,更改项目的模糊权重对稠密数据集影响不大。然而,稀疏数据集的最大模糊模式相对离散,需要多次探测才能够确定最终的模糊模式。此外,通过实验结果相比,传统的频繁模式挖掘、最频繁模式挖掘以及闭频繁模式的挖掘结果离实际需求有较大的差距,说明新型模式挖掘需要增加考虑数据不确定性和离散性。

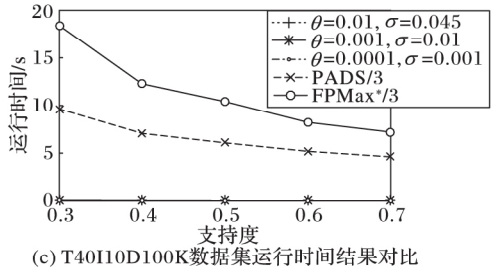
根据算法时间复杂度分析,提出的 MFFP-Tree 模糊模式挖掘算法比 FPMax* 和 PADS 算法具有最好的时间性能。由于模糊修剪策略的提出,即使参数模糊权重和项目的出现频度骤增时,本文提出的最大模糊模式挖掘算法仍具有较低的运行时间增量。同时,在事务数据集的规模增大和项目出现的频度阈值设定较小时,本文算法的优越性更加显著。对所有的数据集,算法 FPMax* 具有最差的时间性能,且当项目的出现频度下降时,该算法的时间复杂度将会骤增。综上,本文算法对实验数据集具有最好的时间性能,针对不同的数据集分别降低时间复杂度一半甚至更低。算法的时间复杂度对比结果见图4。



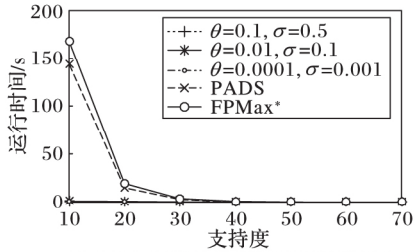

(a) Medical数据集运行时间结果对比




(b) T10I4D100K数据集运行时间结果对比




(c) T40I10D100K数据集运行时间结果对比



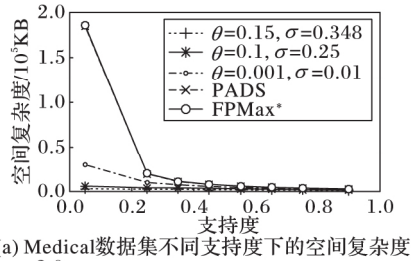

(d) CHESS5数据集运行时间结果对比

**图4 最大模糊模式与传统最大频繁模式挖掘时间复杂度对比**

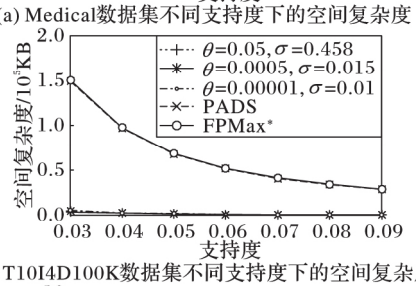
Fig. 4 Comparison of runtime complexity between obtained MFFPs and obtained patterns from traditional frequent pattern mining

算法的空间复杂度的对比结果见图5。从图5可以看出,本文算法所采用的模式搜索策略和阵列技术大大降低了空间复杂度。根据空间复杂度结果分析,本文算法具有显著的性能。算法 FPMax* 和 PADS 的空间复杂度情况非常相似,因为这两种算法均采用了类 FP-tree 结构。但是,这两种算法

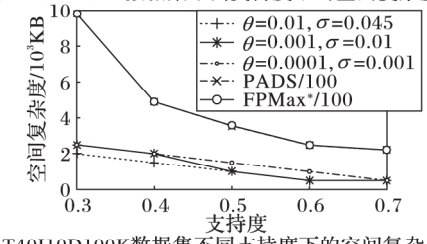
与本文的算法性能具有巨大的差距。因此,为了能良好地显示 3 种算法的空间复杂度对比,本文按照不同比例缩小了 FPMax* 和 PADS 算法的空间复杂度结果。根据图 5 所反映的空间复杂度挖掘结果,相对稠密型数据集,本文提出的最大模糊模式挖掘与 PADS 和 FPMax* 算法在挖掘稀疏数据方面具有更大的优越性。针对不同的数据集,本文算法可以优化空间复杂度从一个数量级到两个数量级不等。最大模糊模式挖掘耗费较少的空间复杂度是因为提出了修剪子树剪枝策略以确保更好地调度候选模式从而进行较少的子模式检测。在提出相应的剪枝策略和模糊约束的基础上,在已有的算法需要检测的一些子模式并不需要在最大模糊模式算法中检测。



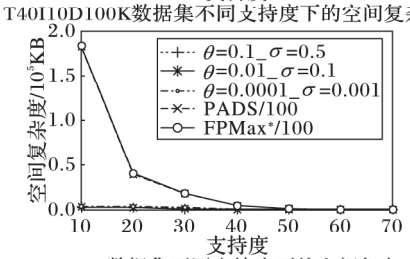
(a) Medical数据集不同支持度下的空间复杂度



(b) T10I4D100K数据集不同支持度下的空间复杂度



(c) T40I10D100K数据集不同支持度下的空间复杂度



(d) CHESS5数据集不同支持度下的空间复杂度

**图 5　最大模糊模式挖掘与传统最大频繁模式挖掘空间复杂度对比**

Fig. 5　Comparison of memory usage between obtained MFFPs and obtained patterns from the traditional frequent pattern mining

## 4　结语

高级模式挖掘对潜在的隐藏信息发现和有用信息的恰当表达至关重要。本研究创新性地提出了模糊模式结构: 核心项和相应的牵引项的组合,并且提出了模糊支持度以及基于模糊支持度的剪枝策略来分析和挖掘隐藏在项目集中的有用信息。为了分析最大模糊模式挖掘算法的有效性,本文对挖掘结果、时间和空间复杂度进行了对比分析,相对于 PADS 和 FPMax* 算法。结果表明,最大模糊模式考虑模糊权重来分析项目的不确定性从而更加准确地反映了项目与项目之间的关系。在时间复杂度方面,最大模糊模式挖掘算法比 PADS 和 FPMax* 算法快 2 倍至一个数量级。在空间复杂度方面,最大模糊模式挖掘算法比 PADS 和 FPMax* 算法优越一个数量级至两个数量级。根据挖掘的有效信息的数量和质量分析,该算法更适合处理频繁项和出现次数较低的项目的组合。

在今后的工作中,从医学的角度,将会对比分析相对频繁的疾病和相对较低的并发症疾病的临床资料,从而从医学的角度验证提出的最大模糊模式对医疗疾病发现的有效性; 从大数据知识发现的角度,将会探究核心-牵引项的模式结构在高级知识挖掘中的作用,从而挖掘更优的新结构和发现更有效的新特征。

**参考文献** ( References)

[1] AGRAWAL R, IMIELINSKI T, SWAMI A. Mining association rules between sets of items in large databases[J]. ACM SIGMOD Record, 1993, 22(2): 207 – 216.

[2] HAN J, PEI J, YIN Y, et al. Mining frequent patterns without candidate generation: a frequent-pattern tree approach[J]. Data Mining and Knowledge Discovery, 2004, 8(1): 53 – 87.

[3] TSENG V S, SHIE B E, WU C W, et al. Efficient algorithms for mining high utility itemsets from transactional databases[J]. IEEE Transactions on Knowledge and Data Engineering, 2013, 25(8): 1772 – 1786.

[4] GRAHNE G, ZHU J. Fast algorithms for frequent itemset mining using FP-trees[J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(10): 1347 – 1362.

[5] ZENG X, PEI J, WANG K, et al. PADS: a simple yet effective pattern-aware dynamic search method for fast maximal frequent pattern mining [J]. Knowledge and Information Systems, 2009, 20(3): 375 – 391.

[6] MUZAMMAL M, RAMAN R. Mining sequential patterns from probabilistic databases[C]// Proceedings of the 2011 Pacific-Asia Conference on Knowledge Discovery and Data Mining. Berlin: Springer, 2011: 210 – 221.

[7] AGGARWAL C, HAN J. Frequent Pattern Mining [M]. Berlin: Springer, 2014: 19 – 61.

[8] ZHANG X, ZHANG Y. Sliding-window top-$k$ pattern mining on uncertain streams[J]. Journal of Computational Information Systems, 2011, 7(3): 984 – 992.

[9] **杨皓, 段磊, 胡斌, 等. 带间隔约束的** Top-$k$ **对比序列模式挖掘**[J]. **软件学报**, 2015, 26(11): 2994 – 3009. ( YANG H, DUAN L, HU B, et al. Mining Top-$k$ distinguishing sequential patterns with gap constraint[J]. Journal of Software, 2015, 26(11): 2994 – 3009. )

[10] CHEN H. Mining top-$k$ frequent patterns over data streams sliding window[J]. Journal of Intelligent Information Systems, 2014, 42(1): 111 – 131.

[11] ZIHAYAT M, AN A. Mining top-$k$ high utility patterns over data streams[J]. Information Sciences, 2014, 285(1): 138 – 161.

[12] YUN U, LEE G. Sliding window based weighted erasable stream pattern mining for stream data applications[J]. Future Generation Computer Systems, 2016, 59(C): 1 – 20.

[13] LI T. Fuzziness in systems modelling[J]. International Journal of General Systems, 2013, 42(1): 1 – 2.

[14] CHEN H, LI T, LUO C, et al. A decision-theoretic rough set approach for dynamic data mining[J]. IEEE Transactions on Fuzzy Systems, 2015, 23(6): 1958 – 1970.　(**下转第** 1465 **页**)

## 5　结语

本文针对视频在时间域上包含多频率分布的情况,提出了基于 S 变换动态滤波的自动检测及放大视频中非平稳微小运动的方法。通过 S 变换,对不同时刻呈现不同频率的运动信号进行分析处理,得出随时间变化的动态频率值,以此设计出动态带通滤波器,最后实现运动放大处理,动态地展现视频中微小运动的运动情况,并在一定程度上抑制噪声干扰。此外,基于现有的一些图像信噪比评价算法,自定义一种视频的 SNR 来分析本文方法的抗噪性能。实验结果表明,本文方法在实际视频放大处理中得出很好的运动放大效果,不仅适用于单一频率或者频率变化不大的视频,而且对多频率的视频放大效果也比较理想,具有更好的实用性。后续将继续研究非平稳信号运动放大的有关问题,并从空域多分辨率角度进行讨论。

**参考文献** (References)


[1] RUBINSTEIN M, WADHWA N, DURAND F, et al. Revealing invisible changes in the world[J]. Science, 2013, 339(6119): 519 – 519.

[2] PARK S, KIM D. Subtle facial expression recognition using motion magnification [J]. Pattern Recognition Letters, 2009, 30(7): 708 – 716.

[3] BALAKRISHNAN G, DURAND F, GUTTAG J. Detecting pulse from head motions in video[J]. Computer Vision and Pattern Recognition, 2013, 9(4): 3430 – 3437.

[4] POH M Z, MCDUFF D J, PICARD R W. Non-contact, automated cardiac pulse measurements using video imaging and blind source separation[J]. Optics Express, 2010, 18(10): 10762 – 10774.

[5] LIU C, TORRALBA A, FREEMAN W T, et al. Motion magnification[J]. ACM Transactions on Graphics, 2005, 24(7): 519 – 526.

[6] WU H, RUBINSTEIN M, SHIH E, et al. Eulerian video magnification for revealing subtle changes in the world[J]. ACM Transactions on Graphics, 2012, 31(4): Article No. 65.

[7] WADHWA N, RUBINSTEIN M, DURAND F, et al. Phase-based video motion processing[J]. ACM Transactions on Graphics, 2013, 32(4): Article No. 80.

[8] SUSHMA M, GUPTA A, SIVASWAMY J. Semi-automated magnification of small motions in videos[C]// PReMI 2013: Pattern Recognition and Machine Intelligence, LNCS 8251. Berlin: Springer, 2013: 417 – 422.

[9] 雷林,李乐鹏,李准,等. 自动检测及放大视频中的微小运动[J]. 小型微型计算机系统,2016,37(9):2120 – 2124. (LEI L, LI L P, LI Z, et al. Automated detection and magnification of small motion in videos [J]. Journal of Chinese Computer Systems, 2016, 37(9): 2120 – 2124.)

[10] STOCKWELL R G, MANSINHA L, LOWE R P. Localization of the complex spectrum: the S transform [J]. IEEE Transactions on Signal Processing, 1996, 44(4): 998 – 1001.

[11] 张建辉. *K*-means 聚类算法研究及应用[D]. 武汉: 武汉理工大学,2007. (ZHANG J H. Research an application of *K*-means clustering algorithm [D]. Wuhan: Wuhan University of Technology, 2007.)

[12] 赵淑红,朱光明. S 变换时频滤波去噪方法[J]. 石油地球物理勘探, 2007,42(4): 402 – 406. (ZHAO S H, ZHU G M. Time-frequency filtering to denoise by S transform[J]. Oil Geophysical Prospecting, 2007, 42(4): 402 – 406.)

[13] RUBINSTEIN M. Analysis and visualization of temporal variations in video[D]. Cambridge: Massachusetts Institute of Technology, 2014.

[14] TURAGA D S, CHEN Y, CAVIEDES J. No reference PSNR estimation for compressed pictures [C]// Proceedings of the 2002 International Conference on Image Processing. Piscataway, NJ: IEEE, 2004: 173 – 184.

[15] 杨柱中,周激流,郎方年. 用噪声检测算法改进理想低通滤波器[J]. 计算机应用,2014,10:2971 – 2975. (YANG Z Z, ZHOU J L, LANG F N. Improving ideal low-pass filter with noise detection algorithm [J]. Journal of Computer Applications, 2014, 34(10): 2971 – 2975.)

[16] 李乐鹏,雷林,孙水发,等. 视频微小运动放大的加速方法[J]. 计算机工程与应用, 2015, 51(24): 195 – 200. (LI L P, LEI L, SUN S F, et al. Improved video small motion magnification processing [J]. Computer Engineering and Applications, 2015, 51(24): 195 – 200.)


This work is partially supported by the National Natural Science Foundation of China (61272237, 61402259), the Hubei Natural Science Fund for Innovative Research Groups (2015CFA025), the Major Program of Educational Commission of Hubei Province of China (D20151204), the Opening Fund of Hubei Key Laboratory of Intelligent Vision Based Monitoring for Hydroelectric Engineering (2014KLA04).

**LEI Lin**, born in 1990, M. S. His research interests include video processing.

**LI Lepeng**, born in 1990, M. S. His research interests include video processing.

**YANG Min**, born in 1992, M. S. candidate. His research interests include video processing.

**DONG Fangmin**, born in 1965, Ph. D., professor. His research interests include intelligent information processing.

**SUN Shuifa**, born in 1977, Ph. D., professor. His research interests include image processing, computer vision.

(上接第 1429 页)

[15] 牛新征,余堃. 基于 FPMAX 的最大频繁项目集挖掘改进算法[J]. 计算机科学,2013,40(12): 223 – 228. (NIU X Z, SHE K. Mining maximal frequent item sets with improved algorithm of FPMAX [J]. Computer Science, 2013, 40(12): 223 – 228.)

This work is partially supported by the National Natural Science Foundation of China (61602064, 61502059), the Scientific Research Foundation of Chengdu University of Information Technology (KYTZ201615).

**ZHANG Haiqing**, born in 1986, Ph. D., lecturer. Her research interests include fuzzy set, decision making, data mining.

**LI Daiwei**, born in 1976, M. S. associate professor. His research interests include data integration and visualization, machine learning.

**LIU Yintian**, born in 1972, Ph. D., professor. His research interests include data integration and visualization, machine learning, data mining.

**YU Xi**, born in 1973, Ph. D., associate professor. His research interests include neural networks, decision making.