doi:10.3969/j.issn.1672-4348.2017.06.018

概念层次架构的挖掘关联规则及其实证应用

王建骅, 陈兆芳, 郑莉

(福建工程学院 管理学院, 福建 福州 350118)

摘要: 在概念层次里进行关联规则的挖掘,并考虑到用户感知与主观判断所产生的认知不确定性;结合模糊分割法与 FP-Growth 方法,应用于概念层次架构中找出关联规则方法,主要分为两个阶段:层级架构的顺序将数据项做抽象化,找出高频模糊格;由高频模糊格来产生多层次模糊关规则。最后通过比较验证所提方法可提高算法的执行效率、缩短计算时间。

关键词: 数据挖掘; 模糊分割法; FP-Growth; 关联规则

中图分类号: TP301.6 **文献标志码**: A **文章编号**: 1672-4348(2017)06-0597-09

A mining association rule based on conceptual hierarchy and its applications

Wang Chien-Hua, Chen Zhaofang, Zheng Li

(School of Management, Fujian University of Technology, Fuzhou 350118, China)

Abstract: An association rule was explored on the conceptual hierarchy and the cognitive uncertainty caused by users' perception and subjective judgment was considered. The fuzzy partition method and FP-Growth were combined to mine the association rules on the conceptual hierarchy. It mainly consisted of two phases: the first was to find the high frequency fuzzy patterns by abstracting data items in the order of the hierarchical structure and the second was to generate multiple-level fuzzy association rules from those frequent patterns. Experiments and comparisons with other methods show that the proposed method could improve the efficiency of the algorithm and shorten the computational time.

Keywords: data mining; fuzzy partition; FP-Growth; association rule

关联规则是数据挖掘中主要的技术之一,其所挖掘的频繁项集主要是以 1994 年的 Agrawal 等人所提出的 Apriori 算法最具代表性[1]。然而 Apriori 算法在执行时会产生庞大数量的候选项目集合以及多次重复扫描数据库等缺点,进而造成算法效能不佳。因此许多学者都提出了许多改进的方式,其目的主要是减少数据库的扫描与增加执行上的效率。例如 DHP 算法可减少候选项目集,来生成关联规则[2];DIC 算法可同时扫描多个阶段,并降低扫描数据库的次数,提高整体效率[3];FP-Growth 算法是利用 FP-tree 的树状结构来将事务数据压缩在内,且在整个挖掘的过程中,只须扫描数据库两次。在过程中不须产生候选项目集,除了大幅的加快挖掘的速度外,还可节省大量的储存空间,因而整体的效率相当不错[4]。

其次,对于一个决策问题而言,需要考虑到使用者的认知与主观判断所产生的认知不确定性。Zadeh 所提出的模糊理论可用来处理包含含糊性与意义不明确的认知不确定性[5]。再加上语意变数与语意值[6-8]所描述的模糊概念较符合决策者在主观上的认知,有助于决策分析的进行,因此近来的模糊数据挖掘则成为一项重要议题。

收稿日期: 2017-09-15
基金项目: 福建工程学院教育科学研究项目(GB-K-17-31)
通讯作者: 王建骅(1979-),男,金门人,讲师,博士,研究方向:数据挖掘与柔计算。

一般在对交易项目描述时,除了可用实体名称来表示外,交易项目之间还能透过抽象化的阶层方式来进行分类。也就是说,从较高阶层上来寻找项目间的关联性是个重要的研究。Han 与 Fu 提出类似图 1 关于“食品”的树状结构的概念阶层结构图[9],该结构模型有 4 个递阶层次,阶层编号由最上层开始编为层次 0,直到产品最底层为层次 3,而阶层越高名称越是广泛化(General)。例如在层次 1 的”Bread”在层次 2 里具有”White”与”Wheat”的广泛化概念。此外,由根节点到最低层次的某一节点则形成一条具有父子关系的路径。例如”Black tea”是”Beverage”的继承者,而”Linton”是”Black tea”的继承者。

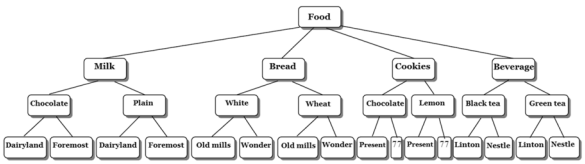

图 1 关于“食品”的概念层次

Fig.1 Conceptual hierarchy for “Food”

基于概念层次的挖掘,本文提出一项模糊数据挖掘方法。首先将架构的每个节点视为一个语意变量并以适当的数量来加以分割,接着采用 FP-growth 方式在多层次间挖掘模糊关联规则;并透过表格结构的方式来储存高频模糊格,这做法比传统交集方式更加快速,且在挖掘的过程中无需产生过多的候选项目集,即可完成整体挖掘任务。

1 相关文献研究

1.1 产生编码表

若一个阶层架构有 h 个层次,则每个节点能被编成 $h-1$ 的序列。当一个节点编在层次 j 时,则节点将被转换成 $c_1c_2\cdots c_{j-1}c_j\cdots *$,其 $j\leqslant h-1$ 且 $*$ 的数量为 $h-j-1$。其次,$c_u(u=1,2,\cdots,j)$ 为一个整数时,其值会由父节点的分支点来决定。以图 1 为例,阶层为 4,故 $h=4$ 且节点的序列长为 3。而“Bread”与“Plain”分别为根节点与“Milk”的第 2 个分支,因此“Bread”与“Plain”可被编成“2 * *”与“12 *”。因此,通过此方法来进行编码后的结果如表 1-表 3 所示。

表 1 在阶层 3 所进行的编码转换

Tab.1 Transcoding on Level 3

TID	Items
1	(112,5),(211,5),(212,7),(221,2),(311,9),(422,8)
2	(111,5),(121,6),(122,3),(222,7),(321,10),(322,3)
3	(212,7),(222,3),(311,5),(322,6)
4	(111,6),(112,1),(211,5),(212,6),(311,7),(321,3),(412,3),(421,1)
5	(111,3),(112,10),(211,2),(222,4),(411,3),(412,9)

表 2 在阶层 2 所进行的编码转换

Tab.2 Transcoding on Level 2

TID	Items
1	(11 * ,5),(21 * ,12),(22 * ,2)(31 * ,9),(42 * ,8)
2	(11 * ,5),(12 * ,9),(22 * ,7),(32 * ,13)
3	(21 * ,7),(22 * ,3),(31 * ,5),(32 * ,6)
4	(11 * ,7),(21 * ,11),(31 * ,7),(32 * ,3),(41 * ,3),(42 * ,1)
5	(11 * ,13),(21 * ,2),(22 * ,4),(41 * ,12)

表 3 在阶层 1 所进行的编码转换

Tab.3 Transcoding on Level 1

TID	Items
1	(1 * * ,5),(2 * * ,14),(3 * * ,9),(4 * * ,8)
2	(1 * * ,14),(2 * * ,7),(3 * * ,13)
3	(2 * * ,10),(3 * * ,11)
4	(1 * * ,7),(2 * * ,11),(3 * * ,10),(4 * * ,4)
5	(1 * * ,13),(2 * * ,6),(4 * * , 12)

1.2 模糊切割法

语意变量的概念是由 Zadeh 所提出[5],而一个语意变量的值可由自然语言的形式来表达[6-8]。因此,模糊分割系就是把每个语意变数以其所给予的语意值加以切割。例如{Small, Medium, Large}是一个定义在数量上的语意值集

合。如图 2 所示，在 2 个二维空间上分别在 x_1 与 x_2 上定义了 3 个语意值，因此共有 9 个模糊格产生，其中的灰色阴影区则是所对应到的模糊格 (A_{11}, A_{23})。图 2 亦为使用模糊切割后的结果。

而其中，模糊格 $\mu^K_{x_k,l_i}(x)$ 被定义为：

$$\mu^K_{x_k,l_i}(x) = \max\{1 - | x - a^K_{l_i} | b^K, 0\} \quad (1)$$

$a^K_{l_i} = m_i + ((m_a - m_i)(i-1)/(K-1))$，$b^K = (m_a - m_i)/(K-1)$，且 m_a 与 m_i 是 x_k 范围区间内的最大值与最小值；l_i 是 K 语意值被定义在语意变量 x_k 上的第 i 个语意值。而 (A_{11}, A_{23}) 被称为候选 2 维模糊格，其是由 A_{11} 与 A_{23} 所产生的。候选 1 维模糊格能进一步产生其他候选或高频模糊格。

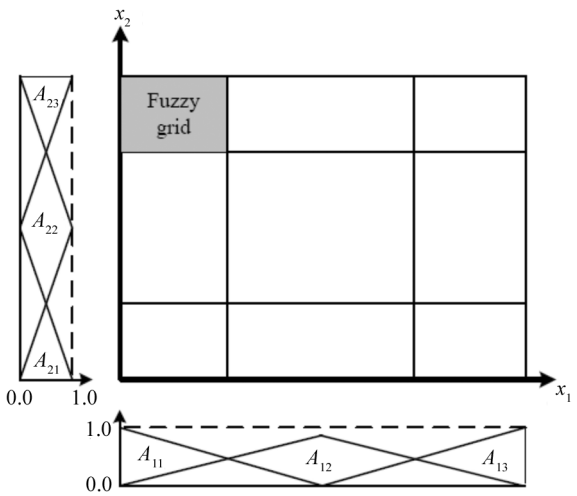

图 2　关于 x_1 与 x_2 的模糊分割法

Fig.2　Fuzzy partitions for x_1 and x_2

2　算法的提出及其原理步骤

首先简要介绍算法所使用到的符号表示。而表 4 是用来说明使用算法的交易事务数据集。图 3 则是此例用到的隶属函数。

2.1　符号表示

n：　交易事务数据的数量；

m：　使用来描述每一笔交易事务项的数量，其中 $1 \leqslant m$；

x_k：　第 k 个项，其中 $1 \leqslant k \leqslant m$；

K：　语意值在交易数据库里的每个量化项的语意值，其中 $K \geqslant 2$；

t_p：　第 p 笔的交易数据，其中 $1 \leqslant p \leqslant n$；

$A^{K,h}_{x_k,l_i}$：　K 语意值的第 i 个语意值在 h 个阶层上被定义在语意值 x_k，其中 $1 \leqslant k \leqslant m$，$1 \leqslant i \leqslant k$；

$\mu^{K,h}_{x_k,l_i}(\cdot)$：　$A^{K,h}_{x_k,l_i}$ 的隶属函数；

$q^{x_k}_{t_p}$：　第 p 笔交易数据在项 x_k 的量化值；

$FS(A^{K,h}_{x_k,l_i})$：　在 h 阶层上的每个项 x_k 的模糊格 $A^{K,h}_{x_k,l_i}$ 支持度；

$G^{\max,h}_{x_k}$：　$A^{K,h}_{x_k,l_i}$ 的模糊格有着最大模糊支持度值；

α：　预先指定的最小模糊支持度值；

β：　预先指定的最小模糊信赖度值；

h：　概念阶层的总层次，其中 $h \geqslant 1$；

SP^h_d：　h 阶层里 d 维的高频样式，其中 $d \geqslant 1$。

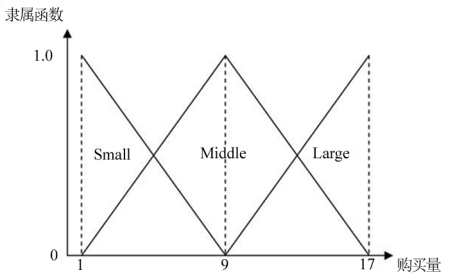

图 3　此例所用到的量化隶属函数

Fig.3　Quantization membership function used in this example

2.2　算法的提出

输入值：

(1) n 笔的交易事务数据；

(2) 每个语意变数有着 K 个语意值；

(3) 概念层次的总层次；

(4) 预先指定的最小模糊支持度值，α。

(5) 预先指定的最小模糊信赖度值，β。

输出值：

模糊关联规则集。

步骤 1：概念层次编码。概念层次有 h 层，则每个节点则被编成 $h-1$ 个序列。如果编码的类型相同，数量则被合并。在表 4 的范例里以层次 3 进行处理，其结果如表 1 所示。

步骤 2：使用模糊切割法将每笔交易数据里每个项 x_k 在 $h-1$ 层的量化项转化成模糊格 $A^{K,h-1}_{x_k,l_i}$ 来呈现，如表 5 所示。

步骤 3：建构表格 FGTTFS 并以由下步骤来生成高频模糊格。其范例格式如表 6 所示。

(1) 模糊格 (Fuzzy grid, FG)：每列表示一个模糊格与每行表示着语意值。

(2) 交易事务 (Transaction table, TT)：每行代

表 4 此例所使用的交易事务数据集

Tab.4 Transaction data set used in this example

TID	Items
1	(Milk-Chocolate-Dairland,5),(Bread-White-Old mills,5),(Bread-White-Wonder,7),(Bread-Wheat-Old mills,2),(Cookies-Chocoate-Present,9),(Beverage-Green tea-Nestle,8)
2	(Milk-Chocolate-Dairland,5),(Milk-Plain-Dairland,6),(Milk-Plain-Foremost,3),(Bread-Wheat-Wonder,7),(Cookies-Lemon-Present,10),(Cookies-Lemon-77,3)
3	(Bread-White-Wonder,7),(Bread-Wheat-Wonder,3),(Cookies-Chocolate-Present,5),(Cookies-Lemon-77,6)
4	(Milk-Chocolate-Dairland,6),(Milk-Chcoclate-Foremost,1),(Bread-White-Old mills,5),(Bread-White-Wonder,6),(Cookies-Chocolate-Present,7),(Cookies-Lemon-Present,3),(Beverage-Black tea-Nestle,3),(Beverage-Green tea-Linton,1)
5	(Milk-Chcoclate-Dairland,3),(Milk-Chocolate-Foremost,10),(Bread-White-Old mills,2),(Bread-Wheat-Wonder,4),(Beverage-Black tea-Linton,3),(Beverage-Black tea-Nestle,9)

表 5 将交易事务数据集转换成模糊集

Tab.5 Fuzzy sets transformed from the transaction data set

TID	Items
1	(0.500/112.Small+0.500/112.Middle),(0.500/211.Small+0.500/211.Middle),(0.250/212.Small+0.750/212.Middle),(0.875/221.Small+0.125/221.Middle),(1.000/311.Middle)+(0.125/422.Small+0.875/422.Middle)
2	(0.500/111.Small+0.500/111.Middle),(0.375/121.Small+0.625/121.Middle),(0.750/122.Small+0.250/122.Middle),(0.250/222.Small+0.750/222.Middle),(0.875/321.Middle+0.125/321.Large),(0.750/322.Small+0.250/322.Middle)
3	(0.250/212.Small+0.750/212.Middle),(0.750/222.Small+0.250/222.Middle),(0.500/311.Small+0.500/311.Middle),(0.375,322.Small+0.625/322.Middle)
4	(0.375/111.Small+0.625/111.Middle),(1.000/112.Small),(0.500/211.Small+0.500/211.Middle),(0.375/212.Small+0.625/212.Middle),(0.250/311.Small+0.750/311.Middle),(0.750/321.Small+0.250/321.Middle),(0.750/412.Small+0.250/412.Middle),(1.000/421.Small)
5	(0.750/111.Small+0.250/111.Middle)+(0.875/112.Middle+0.125.112.Large),(0.875/211.Small+0.125/211.Middle)+(0.625/222.Small+0.375/222.Middle),(0.750/411.Small+0.250/411.Middle)+(1.000/412.Middle)

表 6 建构于阶层 3 的表格 FGTTFS

Tab.6 Table FGTTFS for items on Level 3

Fuzzy grid	FG				TT				FS
	111.Small	111.Middle	…	422.Large	t_1	t_2	…	t_6	
111.Small	1	0	…	0	0	0.500	…	0.750	0.325
111.Middle	0	1	…	0	0	0.500	…	0.250	0.275
⋮	⋮	⋮		⋮	⋮	⋮		⋮	⋮
422.Large	0	0	…	1	0	0	…	0	0

表 t_p 且每元素记录着所符合的模糊格 $A_{x_k,l_i}^{K,h-1}$ 的隶属程度。

(3)模糊支持度(Fuzzy support, FS):其存放所对应模糊格的模糊支持度,其计算方式为下式。

$$FS(A_{x_k,l_i}^{K,h-1}) = [\sum_{p=1}^{n} \mu_{x_k,l_i}^{K,h-1}(q_{t_p}^{x_k})]/n \qquad (2)$$

步骤 4：检查每一个模糊格 $FS(A_{x_k,l_i}^{K,h-1})$ 是否大于或等于预先指定的最小支持度 α。如果 $FS(A_{x_k,l_i}^{K,h-1})$ 满足此条件，则放进高频 1 维模糊格里。在此范例里，α 设为 0.275，因此符合此条件有的糊格有 111.Small、111.Middle、112.Small、112.Small、112.Middle、211.Small、212Middle、222.Small、222.Middle、311.Middle 等 10 个高频 1 维模糊格，此时 SP_1^3 = {111.Small、111.Middle、112.Small、112.Middle、211.Small、212Middle、222.Small、222.Middle、311.Middle}。

步骤 5：当 SP_d^{h-1} 不为空时，就继续执行以下步骤。

步骤 6：寻找 $G_{x_k}^{\max,h-1}$。主要是在相同区域的高频模糊格里，找出项 x_k 有着最大的模糊支持度。在此例子里，结果会产生出 111.Small、112.Small、211.Small、212.Middle、222.Small 与 311.Middle 等 6 个 1 维高频模糊格。

步骤 7：计算每个 1 维高频模糊格值的总和并构建一个递减名为 Header Table 的表格，如表 7 所示。接着就进行模糊 FP-growth。

表 7　阶层 3 的

Tab.7　Header table of Level 3

1-dim fuzzy grids	Count
311.Middle	2.250
212.Middle	2.125
211.Small	1.875
111.Small	1.625
222.Small	1.625
112.Small	1.500

步骤 8：在扫描模糊集时首先要重新建构模糊集表，如图 4 所示，其模糊集是按照 Header Table 来排序。

表 8　阶层 3 的新模糊集

Tab.8　New fuzzy sets of Level 3

TID	Items
1	(1.000/311.Middle),(0.750/212.Middle),(0.500/211.Small),(0.500/112.Small)
2	(0.500/111.Small),(0.25/222.Small)
3	(0.500/311.Middle),(0.750/212.Middle),(0.750/222.Small)
4	(0.750/311.Middle),(0.625.212.Middle),(0.500/211.Small),(0.375/111.Small),(1.000/112.Small)
5	(0.875./211.Small),(0.750/111.Small),(0.625/222.Small)

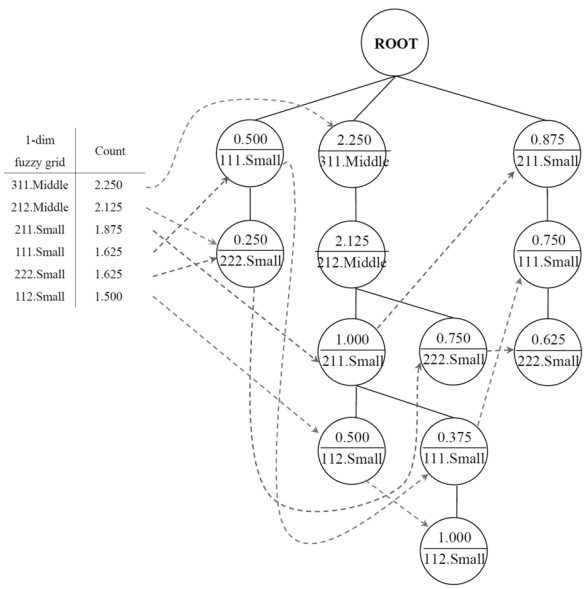

图 4　阶层 3 的模糊 FP-tree

Fig.4　Fuzzy FP-tree of Level 3

步骤 9：在模糊 FP-tree（FFP-tree）的根节点上设置{ROOT}。在第 2 次扫描时新的模糊集就能在树里的模糊格基于相同的交易事务来连接节点，以建构出模糊 FP-tree。其结果如图 4 所示。

步骤 10：在模糊 FP-tree 里挖掘可能的高频样式。从 Header Table 最底端开始扫描并从中取得模糊 FP-tree 从每个项所建构的路径里的条件式模式基底。其次，每个项的条件隶属函数 FP-tree 能由条件模式基础来建构起来。则每个项可能的全部集将从条件模糊 FP-tree 里生成出来。在此例子里，其结果如表 9 所示。

步骤 11：使用下列步骤来找出所符合的高频样式。

（1）把所生成的高频样式放入表格 FGTTFS 计算其模糊支持度，其结果如表 10 所示。

(2)检查每个高频样式的模糊支持度是否大于或等于预先指定的最小模糊支持度值。

在此范例里,阶层 3 的高频样式只有(212.Middle, 311.Middle)的值是大于 α,符合此条件,故被放进 SP_2^3 里。

步骤 12:令 $h-2$,则进行第 2 阶层架构的模糊关联规则挖掘,其编码后的结果如表 2 所示,之后反复步骤 2-11,其最后挖掘的结果为(21 * .Middle, 31 * .Middle)分别放进 SP_2^2 里,第 2 阶层的挖掘完成。接着,令 $h-3$,则进行第一阶层架构的模糊关联规则挖掘,其编码后的结果如表 3 所示,之后反复步骤 2-11,其最后挖掘的结果为(2 * * .Middle, 3 * * .Middle)则被放进 SP_2^1 里,第 1 阶层的挖掘完成。令 $h-4$,由于等于 0,已无阶层可进行挖掘,故完成整体挖掘的动作。

步骤 13:对于 SP_d^h 里的可能的高频样式(C_d^h)建构出候选关联规则:$C_{d_1}^h, C_{d_2}^h, \cdots, C_{d_r}^h \rightarrow C_{d_{r+1}}$,则所有候选关联规则的模糊信赖度是由下式计算的:

表 9 阶层 3 所生成的所有可能样式

Tab.9 All possible generated patterns of Level 3

1 维高频模糊格	所生成的所有可能样式
112.Small	{(211.Small, 112.Small), (212.Middle, 112.Small), (311.Middle, 112.Middle), (111.Small, 112.Small), (212.Middle, 211.Small, 112.Small), (212.Middle, 111.Small, 112.Small), (311.Middle, 212.Middle, 112.Small), (311.Middle, 211.Small, 112.Small), (311.Middle, 111.Small, 112.Small), (311.Middle, 212.Middle, 211.Small, 112.Small), (311.Middle, 212.Middle, 111.Small, 112.Small), (311.Middle, 211.Small, 111.Small, 112.Small), (311.Small, 212.Middle, 211.Small, 111.Small, 112.Small)}
222.Small	{(111.Small, 222.Small), (311.Middle, 222.Small), (212.Middle, 222.Small), (311.Middle, 212.Middle, 222.Small)}
111.Small	{(211.Small, 111.Small), (212.Middle, 111.Small), (311.Middle, 111.Small), (311.Middle, 212.Middle, 111.Small), (311.Middle, 211.Small, 111.Small), (212.Middle, 211.Small, 111.Small), (311.Middle, 212.Middle, 211.Small, 111.Small)}
211.Small	{(212.Middle, 211.Small), (311.Middle, 211.Small), (311.Middle, 212.Middle, 211.Middle)}
212.Middle	{(311.Middle, 212.Middle)}
311.Middle	*

表 10 阶层 3 的所有路径的表格 FGTTFS

Tab.10 Table FGTTFS for all paths on Level 3

Fuzzy grid	FG				TT				FS
	111.Small	112.Small	…	311.Middle	t_1	t_2	…	t_6	
111.Small×112.Small	1	1	…	0	0	0	…	0	0.075
111.Small×211.Small	0	1	…	0	0	0	…	0.656	0.169
⋮	⋮	⋮		⋮	⋮	⋮		⋮	⋮
222.Small ×311.Middle	1	0	…	1	0	0	…	0	0.075

$$FC(C_d^h) = \frac{FS(C_d^h)}{FS(C_{d_1}^h, C_{d_2}^h, \cdots, C_{d_r}^h)} \qquad (3)$$

在此范例里,候选关联规则是:

212.Middle→311.Middle;311.Middle→212.Middle;

21 * .Middle→31 * .Middle;31 * .Middle→21 * .Middle;

2 * * .Middle→3 * * .Middle;3 * * .Middle→2 * * .Middle。

它们的模糊信赖度值分别为 0.750, 0.708,

0.694，0.694，0.611 与 0.660。

步骤 14：检查 $FC(C_d^h)$ 是否大于或等于预先指定的最小模糊信赖度值，β。在此范例里，β 设为 0.65，只要符合 β 值的条件，模糊关联规则就被输出。模糊关联规则如下所示：

(1)若 Bread White Wonder 的购买量为中等量时，则 Cookies Chocolate Present 的购买量为中等量。

(2)若 Cookies Chocolate Present 的购买量为中等量时，则 Bread White Wonder 的购买量为中等量。

(3)若 Bread White 的购买量为中等量时，则 Cookies Chocolate 的购买量为中等量。

(4)若 Cookies Chocolate 的购买量为中等量时，则 Bread White 的购买量为中等量。

(5)若 Cookies 的购买量为中等量时，则 Bread 的购买量为中等量。

3　算例实验及其结果分析

本实验就所提出的方法在不同的数据库大小(包含随机产生的 10 000 与 20 000 笔交易事务)与最小模糊支持度的设定下来探讨运行时间与关联规则的生成所造成的影响。算法是以 C#.Net 设计，并在配备为 3.07GHz 的 Intel Core i3 的个人计算机上执行，其使用 Windows Server　2003 R2 的作业系统与 SQL Server 2008 的数据库上执行。所使用的概念阶层如图 1 所定义，而在每笔交易事务中，所购买的项及其数量由随机产生，但购买数量不超过 17 单位，且购买项不会重复产生。所使用的模糊集如图 3 所示。

将最小模糊支持度设为 0 的情况下，所得到的结果如表 11 与表 12 所示。在表 11 与表 12 的实验结果则包含了 Hong 等的方法[10]与 Hu 的方法[11]得到的数据来比较。从中能发现到最小模糊支持度设定得越小，所产生的关联规则就越多。在运行时间上，所提的方法明显优于 Hang 等与 Hu 的方法，尤其当最小模糊支持度越小就越明显。这也显示出本文所提出的方法在使用 FP-growth 与运用表格 FGTTFS 于阶层架构下来产生高频模糊格及关联规则下，可有效的提升整体的执行效率。且较小的最小模糊支持度也不会严重影响到所提出方法的运行时间。

另外，在关联规则的产出数量上，Hong 等方法的特点是选取了每一个量化属性上最大模糊支持度的一维模糊格；而 Hu 的方法特点则是考虑所有的高频一维模糊格。至于本文所提出的方法，把上述两项特点结合，进而使所产生的高频模糊格数量能在每一量化属性上有着最大的模糊支持度并能把所有的高频一维模糊格作整体的挖掘。因此，能获得较优的效能。

表 11　10 000 笔交易的实验结果

Tab.11　Experimehtal results of 10 000 transactions

最小模糊支持度	Hong 等提出的方法		Hu 提出的方法		本文所提出的方法	
	运行时间/s	规则数	运行时间/s	规则数	运行时间/s	规则数
0.02	653	813	426	1 065	364	1 148
0.03	621	742	382	873	216	882
0.04	583	510	324	613	143	503
0.05	516	321	253	386	93	337
0.06	364	246	176	195	52	67
0.07	183	99	84	73	52	67
0.08	183	99	84	73	52	67
0.09	183	99	84	73	52	67
0.10	183	99	84	73	52	67

表 12 20 000 笔交易的实验结果

Tab.12 Experimental results of 20 000 transactions

最小模糊支持度	Hong 等提出的方法		Hu 提出的方法		本文所提出的方法	
	运行时间/s	规则数	运行时间/s	规则数	运行时间/s	规则数
0.02	1 324	833	811	1 015	666	1 198
0.03	1 289	722	784	852	394	857
0.04	1 021	527	602	596	220	511
0.05	986	303	486	359	163	302
0.06	703	211	331	201	55	67
0.07	337	99	150	73	55	67
0.08	337	99	150	73	55	67
0.09	337	99	150	73	55	67
0.10	337	99	150	73	55	67

4 结论

本文应用模糊分割法与 FP-Growth 算法的手段,提出一种在阶层概念里来找寻模糊关联规则的方法,该方法使用编码方式将阶层里的项逐一编码,然后使用模糊分割法来找出 1 维模糊格,再使用 FP-Growth 算法,在各阶层找寻可生成的模糊关联规则。最后,通过与 Hong 等及 Hu 的方法进行比较,验证本方法的有效性,由于 FP-Growth 之特性在于不用产生候选项集、且将数据库给压缩在 FP-tree 中,故本文的方法有着较佳的执行效率。

其次,在语意值的设定上,使用者可依据本身的偏好、过去的使用经验来主观设定语意值个数及其形状,如高斯分布或梯形隶属函数,如此能更符合使用者在主观上的认知。且 Pedrycz 也提到在模糊系统上使用三角形隶属函数是具有实用性及有效性[12],故本研究采用三角隶属函数作为模糊计算的隶属度函数。

本研究不足之处是实行时忽略了语意值个数与交易事务记录笔数会影响到 FGTTFS 表格的使用空间。当语意值个数越大时,FG 则需要越多的使用空间;当交易事务记录笔数越多时,TT 所设定的使用空间就越大。若能节省 FGTTFS 的使用空间,势必能让算法的效能更加提升。至于在未来的研究展望上,对于设定使用者所给定的参数上,遗传算法(Genetic algorithm)可考虑作为自动决定适合参数的工具,这是由于遗传算法具有自动搜寻的特性。至于最小模糊支持度以及在每个量化属性上所设定的个数 K,Hong 等[13]建议可透过遗传算法的方式来自动取得。而这些建议也是未来继续使用数据挖掘的方法来解决各类型问题的重要参考。

参考文献:

[1] Agrawal R, Srikant R. Fast algorithm for mining association rules in large database[C]//. Proceeding of 20th International Conference on Very Large Databases. Santiago:[s.n.],1994:478-499.

[2] Park J S, Chen M S, Yu P S. An effective hash based algorithm for mining association rules[C]//. Proceeding of the 1995 ACM SIGMOD International Conference on Management of Data. San Jose: ACM Press,1995:175-186.

[3] Brin S,Motwani R, Ullman J D, et al. Dynamic itemset counting and implication rules for market basket data[C]//. ACM SIGMOD International Conference on Management of Data. New York: ACM Press,1997:255-264.

[4] Han J, Pei J, Yin Y. Mining frequent patterns without candidate generation[C]//. Proceeding 2000 ACM SIGMOD International Conference Management of Data. Dallas: ACM Press,2000:1-12.

[5] Zadeh L A. Fuzzy sets[J].Information and Control,1965,8(3):338-353.

[6] Zadeh L A. The concept of a linguistic variable and its application to approximate reasoning-Ⅰ[J]. Information and

Science,1975,8(3):199-249.
[7] Zadeh L A. The concept of a linguistic variable and its application to approximate reasoning-Ⅱ[J].Information and Science, 1975,8(4):301-357.
[8] Zadeh L A. The concept of a linguistic variable and its application to approximate reasoning Ⅲ[J].Information and Science, 1976,9(1):43-80.
[9] Han J W, Fu Y J. Discovery of multiple-level association rules from large database [C]. In Proceedings of International Conference on Very Large Data Bases. Zurich, 1995: 420-431.
[10] Hong T P, Lin K Y, Chien B C. Mining fuzzy multiple-level association rules from quantitative data[J].Applied Intelligence,2003,18(1):79-90.
[11] Hu Y C. Mining association rules at a concept hierarchy using fuzzy partition [J]. Journal of Information Management, 2006,13(3):63-80.
[12] Pedrycz W. Triangular membership functions[J].Fuzzy Sets and Systems,1994,64(1):21-30.
[13] Hong T P, Chen C H, Lee Y C, et al. Genetic-fuzzy data mining with divide-and-conquer strategy[J].IEEE Transactions on Evolutionary Computation,2008,12(2):252-265.

(特约编辑:黄家瑜)

(上接第 585 页)

5　结语

“智能高速”是一个大的系统工程,包括全面高效的交通基础设施和载运工具运行状态感知体系、完备的数据中心体系和信息资源互通共享的开发应用体系、统筹协调的业务管理系统、快捷准确全面的信息服务体系、适应信息化智能化发展要求的技术支撑体系和可信可控的网络与信息安全保障体系等。本文针对“智能高速”中的应急指挥平台智能化提出了初步建设思路,为今后的具体建设提供了指导和参考意见。

参考文献:

[1] 游大磊,王倩.我国物联网发展现状及趋势分析[J].无线互联科技,2017(8):95-96.
[2] 陈晓静.高速公路移动通信应急指挥平台设计与实现[J].江苏科技信息, 2016(25):57-58.
[3] 高照.辽宁省高速公路指挥调度及应急管理平台的设计与实现[D].大连:大连海事大学,2016.
[4] 李宝. 安徽省 FY 市交通运输一体化管理研究[D]. 合肥:安徽大学,2016.
[5] 刘永,林鹰,蒋山,等.基于物联网的高速公路运行管理系统设计[J].微电子学与计算机,2015(1):165-168.
[6] 邓玉勇,李璨,刘洋.我国城市智慧交通体系发展研究[J].城市,2015(11):68-73.
[7] 林颢润.高速公路日常调度与应急指挥系统设计与实现[D].大连:大连理工大学,2014.
[8] 田旭旺.可视数字化综合指挥调度平台在高速公路中的应用[J].中国交通信息化,2014(11):89-90.

(责任编辑: 王圆圆)