文章编号: 1007- 2985(2006)04- 0064- 04

基于 Rough 集的单维布尔关联规则的挖掘算法

卓月明[1], 覃遵跃[2], 胡　斌[2]
(1. 吉首大学物理科学与信息工程学院, 湖南 吉首　416000; 2. 吉首大学信息管理与工程学院, 湖南 张家界　427000)

摘　要: 利用 Rough 集理论中关于等价类的概念, 提出了单维布尔关联规则问题挖掘算法, 考虑到关联规则设定单一最小支持度阈值的局限性, 提出使用多个最小支持度的办法进行频繁项集的发现, 利用兴趣度对单维布尔关联规则进行评价.

关键词: 数据挖掘; 粗糙集; 关联规则; 分类
中图分类号:TP311.11　　文献标识码: A

1993 年 Agrawal 等设计了 Apriori 算法[1], 首先提出了挖掘顾客交易数据库中项集之间的关联规则问题. 近年来发展了很多挖掘算法[2-4]. 20 世纪 80 年代初, 波兰数学家 Pawlak Z 针对 Frege G 的边界线区域思想提出了粗糙集理论(Rough Set). 粗糙集理论根据一个系统的观察和测量所得的现实数据信息, 从分类的观点出发, 以集合近似、近似分类与不可分辨的概念为基础, 通过知识约简从中发现、推理知识和分辨系统的特点、过程、预测系统的结果等[5-6]. 笔者利用 Rough 集理论中关于等价类的概念, 针对单维布尔关联规则问题提出一种挖掘算法, 并利用兴趣度对规则进行评价.

1　挖掘关联规则的 DM_R 算法

关联规则的挖掘过程: (1) 找出所有频繁项集, (2) 由频繁项集产生强关联规则. 第 2 步最容易, 直接将频繁项集中的各项进行组合即可, 挖掘关联规则的总体性能由第 1 步决定. 下面以对表 1 所示的事务数据库进行单维布尔关联规则挖掘为例, 介绍 DM_R 算法. 交易集合 A = {牛奶(M), 面包(B), 鸡蛋(E), 盐(S), 米(R)}, 每个交易在某个事务 T 中或者出现, 或者不出现, 各项事务重复出现的次数也在表 1 中给出.

表 1　某商场的部分食品交易记录

事务集合 W	出现频度	交易记录中各项的 ID 列表	事务集合 W	出现频度	交易记录中各项的 ID 列表
$T1$	4	M, B, E, R	$T6$	8	M, B, E
$T2$	1	S, R	$T7$	2	B, E, S
$T3$	4	B	$T8$	14	M, B
$T4$	3	M, B, E, S	$T9$	6	M, E
$T5$	8	B, E			

1.1 频繁项集的生成

1.1.1 使用等价类概念挖掘候选项集　DM_R 算法借助不可分辨关系的概念, 将事务数据库按照交易集合划分等价类. 将事务数据库中的各项事务看作对象, 各项交易是属性, 对于某一交易而言, 如果它在某项事务中出现则其属性值为 1, 否则为 0, 即每项交易均被认为是一个二元属性, 这样就把事务数据库转化为一个信息系统, 于是即可按照属性或属性集合划分等价类. 对于交易集合 A_i, 按照事务是否包含将划分为

收稿日期: 2006-05-20
基金项目: 湖南省教育厅科学研究项目(05C141)
作者简介: 卓月明(1970-), 男, 湖南慈利人, 吉首大学物理科学与信息工程学院讲师, 硕士生, 高级程序员, 主要从事分布式计算和计算机网络研究.

$$W/A_i = \{\{W_i\}, \{W - W_i\}\},$$

其中: W_i 为包含交易集合 A_i 的事务集合. 交易集合 A_i 的出现频度为 card(W_i), 它是一个候选项集, 其支持频度 $s = \frac{\text{card}(W_i)}{\text{card}(W)}$, 若 $s = \min\_\sup$, 则交易集合 A_i 就是一个频繁项集.

对于候选项集的搜索采取分层递进方法, 即可利用 $k-$ 候选项集来产生 $(k+1)-$ 候选项集, 一个候选项集的任意子集也是一个候选项集. 若 A_i 和 A_j 分别是 $k-$ 候选项集, 则对于单项交易 $a \quad A_i$ 且 $a \quad A_j$, 可生成一个 $(k+1)-$ 候选项集 $\{a \quad A_j\}$, 它对事务数据库的划分结果为

$$W/\{a \quad A_j\} = \{\{W_i \quad W_j\}, \{W - W_i \quad W_j\}\},$$

其中: $W_i \quad W_j$ 为包含 $(k+1)-$ 候选项集 $\{a \quad A_j\}$ 的事务集合, 该项集是否为 $(k+1)-$ 频繁项集取决于其出现频度 card($W_i \quad W_j$) 是否小于 $\min\_\sup \quad \text{card}(W)$.

从 $k-$ 候选项集中可以直接产生频繁项集, 同时还可以生成 $(k+1)-$ 候选项集而无需搜索数据库, 因此 DM_R 算法只需在生成 $1-$ 候选项集时对数据库进行一次搜索, 从而大大减少计算时间.

1.1.2 使用多个最小支持度阈值挖掘频繁项集　进行关联规则挖掘时的一个核心要素是设定规则的最小支持度阈值. 多数现有的关联规则挖掘仅使用一个最小支持度阈值, 这意味着事务数据库中各种交易具有等同的出现频率, 而实际情况上有些交易出现得相当频繁, 而有些交易则很少出现. 在 DM_R 算法中, 笔者提出使用多个最小支持度阈值来反映数据表中各项交易出项频率的差别, 以期待较好地反映事务数据库的本质. 为数据表中的每项交易均指定一个最小支持度, 称之为最低交易频度(MIF), 而关联规则的最小支持度阈值取决于规则中包含的具体交易的最低交易频度.

定义 1　设关联规则 $R: a_1, a_2, \quad , a_k \quad a_{k+1}, \quad , a_r$, 其中 $a_j \quad D$, 且 $a_i \quad a_j, i, j \quad \{1, 2, \quad , r\}$. 令 MIF($a_i$) 表示交易 a_i 的 MIF 值, 则规则 R 的最小支持度阈值定义为

$$\min\_\sup(R) = \min(\text{MIF}(a_1), \text{MIF}(a_2), \quad , \text{MIF}(a_r)).$$

通过对各项交易设定不同的值, 用户可以灵活控制不同的关联规则的最小支持度阈值, 从而发现包含非频繁交易的具有较低支持度的关联规则, 以及具有较高支持度的包含频繁交易的关联规则, 同时又不会引入过多无意义规则.

下面对表 1 进行频繁项集的挖掘. 设 $\text{MIF}(M) = 0,5, \text{MIF}(B) = 0.6, \text{MIF}(E) = 0.4, \text{MIF}(S) = 0,1, \text{MIF}(R) = 0.05$.

(1) 数据表中出现的每个交易均为 $1-$ 候选项集 L_1 中的元素, 扫描整个数据表, 确定 L_1 中每个元素对数据表的划分结果. 计算结果如下:

$$W/\{M\} = \{\{T1, T4, T6, T8, T9\}, \{T2, T3, T5, T7\}\},$$
$$W/\{B\} = \{\{T1, T3, T4, T5, T6, T7, T8\}, \{T2, T9\}\},$$
$$W/\{E\} = \{\{T1, T4, T5, T6, T7, T9\}, \{T2, T3, T8\}\},$$
$$W/\{S\} = \{\{T2, T4, T7\}, \{T1, T3, T5, T6, T8, T9\}\},$$
$$W/\{R\} = \{\{T1, T2\}, \{T3, T4, T5, T6, T7, T8, T9\}\}.$$

(2) 为了发现 $2-$ 频繁项集, 首先确定 $2-$ 候选项集 C_2, C_2 中元素应为 C_1 中元素的两两组合, C_2 中元素对数据表的划分及交易集合的出现频度分别为:

$W/\{M, B\} = \{\{T1, T4, T6, T8\}, \{T2, T3, T5, T7, T9\}\}$, 出现频度为 21,
$W/\{M, S\} = \{\{T4\}, \{T1, T2, T3, T5, T6, T7, T8, T9\}\}$, 出现频度为 3,
$W/\{M, R\} = \{\{T1\}, \{T2, T3, T4, T5, T6, T7, T8, T9\}\}$, 出现频度为 4,
$W/\{B, E\} = \{\{T1, T4, T5, T6, T7\}, \{T2, T3, T8, T9\}\}$, 出现频度为 25,
$W/\{B, S\} = \{\{T4, T7\}, \{T1, T2, T3, T5, T6, T8, T9\}\}$, 出现频度为 5,
$W/\{B, R\} = \{\{T1\}, \{T2, T3, T4, T5, T6, T7, T8, T9\}\}$, 出现频度为 4,
$W/\{E, S\} = \{\{T4, T7\}, \{T1, T2, T3, T5, T6, T8, T9\}\}$, 出现频度为 5,
$W/\{E, R\} = \{\{T1\}, \{T2, T3, T4, T5, T6, T7, T8, T9\}\}$, 出现频度为 4,

$W/\{R,S\}=\{\{T2\},\{T1,T3,T4,T5,T6,T7,T8,T9\}\}$, 出现频度为1.

根据定义1可计算得到: $\min\_\sup(M,B)=\min(\mathrm{MIF}(M),\mathrm{MIF}(B))=0.5$, $\min\_\sup(M,E)=0.4$, $\min\_\sup(M,S)=0.1$, $\min\_\sup(M,R)=0.05$, $\min\_\sup(B,E)=0.4$, $\min\_\sup(B,S)=0.1$, $\min\_\sup(B,R)=0.05$, $\min\_\sup(E,S)=0.1$, $\min\_\sup(E,R)=0.05$, $\min\_\sup(S,R)=0.05$. 则对2-频繁项集L_2的搜寻结果如图1所示.

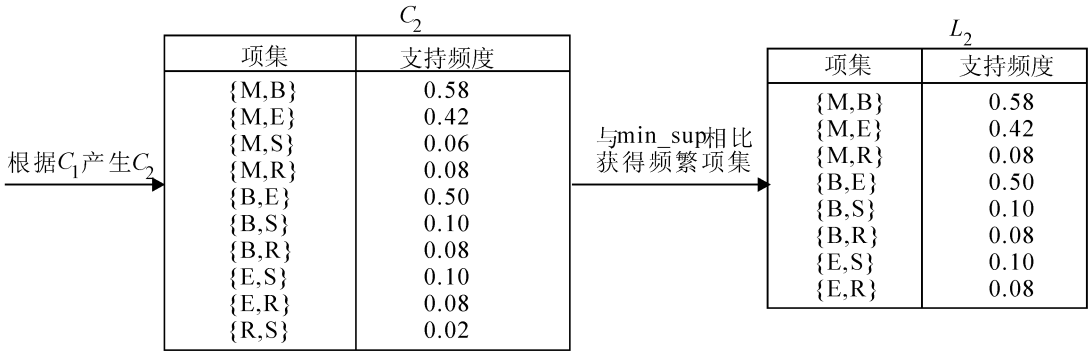

图1 2-候选项集C_2产生2-频繁项集L_2

(3) 发现3-频繁项集L_3的过程与发现L_2的过程类似, 根据定义1可计算得到: $\min\_\sup(M,B,E)=\min(\mathrm{MIF}(M),\mathrm{MIF}(B),\mathrm{MIF}(E))=0.4$, $\min\_\sup(M,B,S)=0.1$, $\min\_\sup(M,B,R)=0.05$, $\min\_\sup(M,E,S)=0.1$, $\min\_\sup(M,E,R)=0.05$, $\min\_\sup(M,S,R)=0.05$, $\min\_\sup(B,E,S)=0.1$, $\min\_\sup(B,E,R)=0.05$, $\min\_\sup(B,S,R)=0.05$, $\min\_\sup(E,S,R)=0.06$. 对L_3的搜索结果如图2所示.

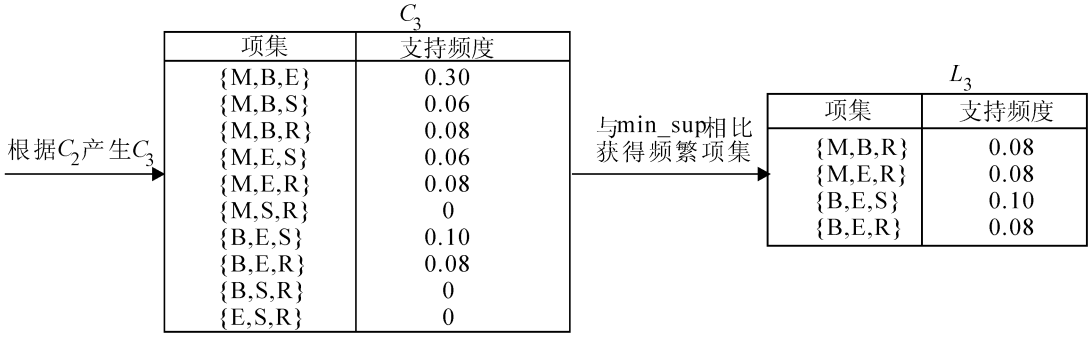

图2 3-候选项集C_3产生3-频繁项集L_3

(4) 发现4-频繁项集L_4的过程与发现L_2的过程类似. 其中由4-候选项集C_4产生4-频繁项集L_p的过程如图3所示.

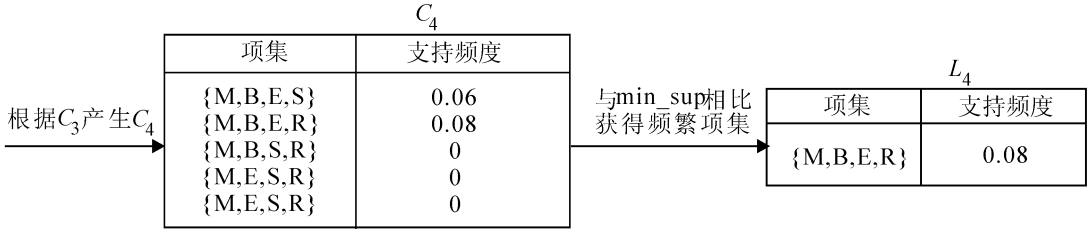

图3 4-候选项集C_4产生4-频繁项集L_4

当求5-候选集时, 5-候选项集C_5中只包含1个元素$\{M,B,E,R,S\}$: $W/\{M,B,E,R,S\}=\{\ ,\{T1,T2,T3,T4,T5,T6,T7,T8,T9\}\}$, 支持度为0, 而$\min_\sup(M,B,E,R,S)=\min(\mathrm{MIF}(M),\mathrm{MIF}(B),\mathrm{MIF}(E),\mathrm{MIF}(R),\mathrm{MIF}(S))=0.05$, 显然$L_5=\ $. 至此, 算法无法再发现新的频繁项集.

2.2 关联规则的生成

在从数据表中挖掘出所有的频繁项集后, 很容易就可以获得相应的关联规则, 也就是产生满足最小支持度阈值和最小信任度阈值的强关联规则. 由于规则是通过频繁项集直接产生的, 因此在生成强关联规则

的这一步骤所涉及的所有项集均满足对最小支持度阈值的要求. 设 min _ conf= 80%, 由表 1 生成强关联规则集合如表 2 所示.

表 2　由表 1 挖掘出的强关联规则

关联规则	置信度 %	关联规则	置信度 %	关联规则	置信度 %	关联规则	置信度 %
M B	82.9	*R M*	80.0	*B R M*	100	*E S B*	100.0
E B	80.6	*R B*	80.0	*B R E*	100	*E R B*	100.0
S B	83.3	*R E*	80.0	*R M B E*	80	*R E M*	100.0
S E	83.3	*S B E*	83.3	*M R B E*	100	*M B R E*	100.0
M R B	100	*R M B*	80.0	*B R M E*	100	*M E R B*	100.0
M R E	100	*R M E*	80.0	*E R M B*	100	*B E R M*	100.0
B S E	100	*R B E*	80.0				

3　结语

关联规则挖掘算法是为大规模数据集设计的一种相当高效的算法, 关于关联规则算法的研究往往强调的是计算效率, 而不是对算法规则的解释. 笔者提出的基于 Rough 集针对单维布尔关联规则挖掘问题的 DM _ R 算法, 按照某项交易是否出现将事务集合划分为不同等价类, 并通过对各项交易定义最低交易频度间接设定多个最小支持度阈值, 以更好地进行频繁项集的挖掘, 并在此基础上生成强关联规则集合.

参考文献:

[1] AGRAWAL R, SRIKANT R. Fast Algorithms for Mining Association Rules [A]. BOCCA J B, JARKE M, ZANIOLO C, ets. Proceedings of the 20th International Conference on VLDB [C]. SanFrancisco: Morgan Kaufmann Publishers, 1994.

[2] AGRAWAL R, SRIKANT R. Mining Sequential Patterns [A]. YU P S, CHEN A L P, eds. Proceedings of the 11th International Conference on Data Engineering [C]. Los Alamitos, CA: IEEE Computer Society, 1995.

[3] BAYARDO R J. Efficiently Mining Long Patterns from Databases [A]. HASS L M, TIWARY A, eds. Proceedings of the ACM SIGMOD InternationalConference on Management of Data [C]. New York: ACM Press, 1998.

[4] DAO- I LIN, ZVI M KEDEM. Pincer- Search: A New Algorithm for Discovering the Maximum Frequent Set (1997) Lecture Notes in Computer Science [A]. Proceedings of the 6th European Conference on Extending Database Technology Heidelberg: Springverlag, 1998.

[5] 吉根林, 杨　明, 宋余庆, 等. 最大频繁项目集的快速更新 [J]. 软件学报, 2005, 28(1): 128- 135.

[6] 曾黄麟. 智能计算 [M]. 重庆: 重庆大学出版社, 2004.

[7] 范　明, 孟小峰. 数据挖掘概念与技术 [M]. 北京: 机械工业出版社, 2005.

A Mining Algorithm of Single-Dimensional Boolean Association Rules Based on Rough Set

ZHUO Yue-ming[1], QING Zun-yue[2], HU Bin[2]

(1. College of Physics Science & Information Engineering, Jishou University. Jishou 416000, Hunan China; 2. College of Information Management & Engineering, Jishou University, Zhangjiajie 427000, Hunan China)

Abstract: Applying the concepts of equivalent class in the Rough Set Theory, the paper advances association rules to mine valuable knowledge that describe the interrelationship about data item, and put forward one mining algorithm. Considering the limit of localization of Association rules setting single minimum support threshold, it brings forward the discovery of numerous item class through many minimum supports and evaluating rule though interestingness.

Key words: data mining; rough set; association rule; classify

(责任编辑　陈炳权)