文章编号:1001 - 9081(2009)01 - 0189 - 04

基于激励的关联规则的挖掘

刘旭辉[1], 邵世煌[2], 余光柱[2,3]
(1. 长江大学 机械工程学院,湖北 荆州 434000; 2. 东华大学 信息科学与技术学院,上海 201600;
3. 湖北警官学院 计算机系,武汉 430034)
(Guang216@126.com)

摘 要:基于支持度的关联规则挖掘算法无法找到那些非频繁但效用很高的项集,基于效用的关联规则会漏掉那些效用不高但发生比较频繁、支持度和效用值的积(激励)很大的项集。提出了基于激励的关联规则挖掘问题及一种自下而上的挖掘算法 HM-miner。激励综合了支持度与效用的优点,能同时度量项集的统计重要性和语义重要性。HM-miner 利用激励的上界特性进行减枝,能有效挖掘高激励项集。

关键词:关联规则;基于激励;支持度;效用;兴趣度
中图分类号: TP182 **文献标志码**:A

Motivation-based association rule mining

LIU Xu-hui[1], SHAO Shi-huang[2], YU Guang-zhu[2,3]
(1. *College of Mechanical Engineering, Yangtze University, Jingzhou Hubei* 434000, *China*;
2. *College of Information Science and Technology, Donghua University, Shanghai* 201600, *China*;
3. *Department of Computer Science, Hubei University of Police, Wuhan Hubei* 430034, *China*)

Abstract: The existing algorithms for support-based Association Rule Mining (ARM) cannot find the itemsets that are not frequent but have high utility values, while Utility-Based Association Rule Mining (UBARM) cannot find the itemsets whose utility values are not high but the product of the support and utility of the same itemset (defined as motivation) is very large. This paper proposed motivation-based association rule and a down-top algorithm called HM-miner to find all high motivation itemsets efficiently. By integrating the advantages of support and utility, the new measure, i.e., motivation can measure both the statistical and semantic significance of an itemset. HM-miner adopted a new pruning strategy, which was based on the motivation upper bound property, to cut down the search space.

Key words: association rule; motivation-based; support; utility; interestingness

0 引言

传统的关联规则[1-3]假设用户只对频繁发生的项集感兴趣,用支持度度量项集的重要性,无法找到那些非频繁但效用很高、更能引起用户兴趣的项集。基于效用的关联规则[4-6]假设用户只对能带来高效用的项集感兴趣,用效用来度量项集的重要性,致力于发现所有高效用项集而忽略了那些效用不高(项集的效用略小于用户定义的效用阈值)但发生比较频繁的项集。这些项集的效用不是很高,但项集的支持度与效用的积很大,很可能引起用户的兴趣:由该项集生成的规则可能蕴涵着一个具有很大吸引力的决策方案。为此,提出了基于激励的关联规则及一种自下而上的挖掘算法。

期望理论[7]认为,激励是评价、选择的过程,人们采取某项行动的动力(或激励力)取决于其对行动结果的价值评价(效价)和预期实现目标可能性的估计(期望值)。换言之,激励力的大小取决于效价与期望值的乘积:

$$\text{激励力} = \text{效价} \times \text{期望值} \tag{1}$$

根据这一理论,只有那些支持度(期望值)与效用(效价)同时处于较高水平的项集,才可能产生强大的激励力。基于激励的关联规则用激励(支持度与效用的积)来度量项集的重要性,主要任务是发现所有高激励项集。

1 背景知识

1.1 概念与定义

设 $I = \{i_1, i_2, \cdots, i_m\}$ 为项目集合, $T = \{t_1, t_2, \cdots, t_n\}$ 为事务数据库。每一事务 $t_q (t_q \in T)$ 都有一个唯一的编号 Tid,所包含的项目构成 I 的一个子集,记为 $t_q \subseteq I$。如果对于 I 中的一个子集 S,有 $S \subseteq t_q$,就说 t_q 包含 S。

定义 1 项集 S 的事务集,记为 T_s,是所有包含 S 的事务的集合,即:

$$T_s = \{t_q \mid S \subseteq t_q, t_q \in T\} \tag{2}$$

定义 2 项目 i_p 的事务效用,记为 $u(i_p, t_q)$,指事务 t_q 发生时项目 i_p 带给用户的效用(效益)。本文假设所指效用为经济效用。在事务数据库中,项目的事务效用是该项目的单位利润与销售量之积。

定义 3 项目 i_p 在项集 S 中的效用值,记为 $u(i_p, S)$,是 i_p 在事务 $t_q (t_q \in T_s)$ 中的事务效用的和。即:

$$u(i_p, S) = \sum_{t_q \in T_s} u(i_p, t_q) \tag{3}$$

定义 4 项集 S 的效用,记为 $u(S)$,指 S 中所有项目 i_p 在项集 S 中的效用值的和,即:

收稿日期:2008 - 07 - 10;**修回日期**:2008 - 09 - 14。 **基金项目**:国家教育部博士点基金资助项目(20060255006)。

作者简介:刘旭辉(1966 -),男,湖北荆州人,副教授,博士研究生,主要研究方向:数据挖掘; 邵世煌(1938 -),男,上海人,教授,主要研究方向:智能控制、数据挖掘; 余光柱(1969 -),男,湖北荆州人,副教授,博士,主要研究方向:数据挖掘。

$$u(S) = \sum_{i_p \in S} u(i_p, S) \tag{4}$$

定义5 如果 $u(S) \geqslant minutil$ 成立，S 是高效用项集，否则，S 是低效用项集。$minutil$ 是用户指定的阈值。基于效用的关联规则挖掘，就是要发现所有高效用项集。

定义6 项集 S 的激励，记为 $m(S)$，指项集 S 的支持度 $s(S)$ 与效用值的积，即：

$$m(S) = s(S) \times u(S) \tag{5}$$

定义7 一条基于激励的关联规则(motivation-based association rule)，就是一个形如“$X \rightarrow Y$”的蕴涵式，同时满足下列条件：

1) 关联规则“$X \rightarrow Y$”的支持度 $s(X \cup Y)$ 大于等于用户定义的阈值 $minsup$，即：

$$s(X \cup Y) = \frac{\text{Support}(X \cup Y)}{n} \geqslant minsup \tag{6}$$

2) 关联规则“$X \rightarrow Y$”的置信度 $conf$ 大于等于用户定义的阈值 $minconf$，即：

$$conf = \frac{\text{Support}(X \cup Y)}{\text{Support}(X)} \geqslant minconf \tag{7}$$

3) 项集 $(X \cup Y)$ 的效用值 $u(X \cup Y)$ 大于等于用户定义的阈值 $minutil$，即：

$$u(X \cup Y) \geqslant minutil \tag{8}$$

4) 项集 $(X \cup Y)$ 的激励大于等于用户定义的阈值 $minmotivation$，即：

$$m(X \cup Y) \geqslant minmotivation \tag{9}$$

式(6)中，Support$(X \cup Y)$ 为数据库中包含 $X \cup Y$ 的交易数，n 为数据库中的交易总数。我们把满足条件1、3和4的项集称为高激励项集。

1.2 相关研究

支持度与效用都是人们兴趣的一种度量。支持度反映的是项集的统计特性，属于一种客观度量，其不足是不能反映项集的语义特性；相反，效用反映的是项集的语义特性，属于一种主观度量[8]。决定人们兴趣的，往往是主观因素与客观因素的综合。因此，单纯依靠支持度或效用来度量项集重要性的关联规则挖掘模型难以全面表达人们的兴趣。

文献[3]提出了一个面向目标的基于效用的关联规则挖掘模型(OOA模型)。OOA模型同时用支持度和效用度量项集的重要性，能发现数据集中高效用频繁集。但是，OOA模型及他提出的OOApriori算法与我们的目标有下列不同：1)OOA关联规则不要求项集的激励大于等于某一阈值；2)OOA模型中支持度阈值必须设置得比较大，否则会产生太多的频繁集。因此，OOA模型仍然会丢失一些支持度不高，但激励很大的模式。在基于激励的关联规则挖掘中，支持度阈值 $minsup$ 和效用阈值 $minutil$ 往往较小，主要通过激励阈值 $minmotivation$ 来除去不太重要的规则。设置 $minsup$ 和 $minutil$ 的目的是为了过滤掉一些非常频繁但效用太低或虽然效用很高但属于偶然的无用模式。

文献[9]提出用“一般效用(general utility)”度量项集的重要性。按定义，项集 S 的一般效用 $gu(S)$ 等于它的支持度与效用的加权和，即 $gu(S) = \lambda s(S) + (1-\lambda)u(S)$。“一般效用”同时反映了项集的语义特性与统计特性，但权值 λ 的确定带有较大的随意性，概念的意义也不如激励直观。激励基于概率论与管理学，更好理解。

文献[4]提出了一种基于效用的关联规则挖掘算法UMining，为我们的研究奠定了基础。

2 一种自底向上的挖掘算法

2.1 激励约束的特性

文献[4]表明，效用约束既不是单调的(monotone)、非单调的(anti-monotone)、可转换的(convertible)，也不是简洁的(succinct)。根据激励的定义可知，激励约束既不是单调的、非单调的、可转换的，也不是简洁的。

定义8 设 S^k 是一个 k-项集(S 中含有 k 个不同的项目)。S^k 的所有长度为 $(k-1)$ 的子集构成一个集合 $L^{k-1} = \{S^{k-1} \mid S^{k-1} \subset S^k\}$。显然，$|L^{k-1}| = k$。

定义9 包含 i_p 的所有 S^{k-1} 组成一个新的集合 $L_{ip}^{k-1} = \{S^{k-1} \mid i_p \in S^{k-1}, S^{k-1} \in L^{k-1}\}$。显然，$|L_{i_p}^{k-1}| = k-1$。

例如：一个4-项集 $S^4 = ABCD$，根据定义，有 $L^3 = \{ACD, ABD, ABC, BCD\}$，$L_A^3 = \{ACD, ABD, ABC\}$。

定理1 激励的上界特性。设 $m(S^k)$ 是 k-项集 S^k 的激励，那么下式成立：

$$m(S^k) \leqslant \frac{\sum_{S^{k-1} \in L^{k-1}} m(S^{k-1})}{k-1} \tag{10}$$

证明 设 $u(S^k)$ 是 k-项集 S^k 的效用值，根据文献[4]的定理2(效用的上界特性)，有：

$$u(S^k) \leqslant \frac{\sum_{S^{k-1} \in L^{k-1}} u(S^{k-1})}{k-1} \tag{11}$$

把式(11)两边同乘以 S^k 的支持度 $s(S^k)$，得

$$s(S^k) \times u(S^k) \leqslant s(S^k) \times \frac{\sum_{S^{k-1} \in L^{k-1}} u(S^{k-1})}{k-1} \tag{12}$$

由于 $s(S^k) \leqslant s(S^{k-1})$，上式可写为：

$$m(S^k) \leqslant s(S^k) \times \frac{\sum_{S^{k-1} \in L^{k-1}} u(S^{k-1})}{k-1} \leqslant \frac{\sum_{S^{k-1} \in L^{k-1}} u(S^{k-1}) s(S^{k-1})}{k-1} = \frac{\sum_{S^{k-1} \in L^{k-1}} m(S^{k-1})}{k-1} \tag{13}$$

证毕。

定义10 S^k 的候选项集，记为 C^{k-1}，是 S^k 的 $(k-1)$-子集的集合，即 $C^{k-1} \subseteq L^{k-1}$。

定义11 S^k 的激励上界，记为 $bm(S^k)$，定义为：

$$bm(S^k) = \frac{\sum_{S^{k-1} \in C^{k-1}} m(S^{k-1})}{|C^{k-1}| - 1} \tag{14}$$

式(12)中的 C^{k-1} 即 S^k 的候选项集，式中 $|C^{k-1}|$ 是 C^{k-1} 的基数。

定理2 减枝策略。如果每一个 $k-1$-项集 $S^{k-1} \in (L^{k-1} - C^{k-1})$ 是一个低激励项集，且 $b_m(S^k) < minmotivation$，那么，S^k 是低激励的，即：$m(S^k) < minmotivation$。

证明 既然每一个 $S^{k-1} \in (L^{k-1} - C^{k-1})$ 是低激励项集，那么有

$$\sum_{S^{k-1} \in (L^{k-1} - C^{k-1})} m(S^{k-1}) < \sum_{S^{k-1} \in (L^{k-1} - C^{k-1})} minmotivation \tag{15}$$

根据定理1即式(10)，可得：

$$m(S^k) \leqslant \frac{\sum_{S^{k-1}\in L^{k-1}} m(S^{k-1})}{k-1} = \frac{\sum_{S^{k-1}\in (C^{k-1}\cup (L^{k-1}-C^{k-1}))} m(S^{k-1})}{k-1} =$$

$$\frac{\sum_{S^{k-1}\in C^{k-1}} m(S^{k-1}) + \sum_{S^{k-1}\in (L^{k-1}-C^{k-1})} m(S^{k-1})}{k-1} \leqslant$$

$$\frac{\sum_{S^{k-1}\in C^{k-1}} m(S^{k-1}) + \sum_{S^{k-1}\in (L^{k-1}-C^{k-1})} minmotivation}{k-1} \leqslant \quad //\text{根据式(15)}$$

$$\frac{\sum_{S^{k-1}\in C^{k-1}} m(S^{k-1}) + |L^{k-1} - C^{k-1}| \, minmotivation}{k-1} \leqslant$$

$$\frac{(|C^{k-1}|-1) b_m(S^k) + |L^{k-1} - C^{k-1}| \, minmotivation}{k-1} = \quad //\text{根据式(14)}$$

$$\frac{(|C^{k-1}|-1) b(S^k) + (k-|C^{k-1}|) minmotivation}{k-1} \leqslant$$

$$\frac{(|C^{k-1}|-1) minmotive + (k-|C^{k-1}|) minmotivation}{k-1} \leqslant \quad //b_m(S^k) < minmotivation$$

$$\frac{(k-1) minmotivation}{k-1}$$

因此,$m(S^k) < minmotivation$。　　证毕。

2.2 算法

基于上述的减枝策略,我们提出了一种新的类似于 UMining 的算法,称为 HM-Miner (High motivation itemset miner)。HM-Miner 算法采用了自底而上的搜索策略,反复从 $k-1$- 项集生成 k- 项集,并计算候选集的支持度、效用与激励。程序描述如下:

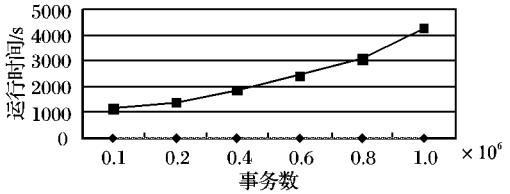
```
输入:数据库 T,阈值 minsup, minutil, minmotivation
输出:高激励项集集合 HM
1)   {
2)     I = scan(T);
3)     C1 = I;
4)     k = 1;
5)     Ck = CalculateAndStore(Ck,T)
6)     H = Discover(Ck, minsup,minutil,minmotivation);
7)     While (| Ck | > 0)
8)     {
9)       k = k + 1;
10)      Ck = Generate(Ck-1,I);
11)      Ck = Prune(Ck, Ck-1, minsup,minutil,
              minmotivation);
12)      Ck = CalculateAndStore(Ck,T);
13)      HM = HM ∪ Discover(Ck, minsup,minutil,
              minmotivation);
14)    }
15)    Return HM;
16)  }
```

函数 scan 扫描数据库 T 并发现所有的项目的集合 I; CalculateAndStore 函数计算 C_k 中每一个 k- 项集的支持度、效用和激励,存储在相应的数据结构中; Discover 用于找出 C_k 中满足条件的高激励项集; Generate 函数负责从 C_{k-1} 中的 $k-1$- 项集生成可能的候选 k- 项集; Prune 函数为 C_k 中的每一个 k- 项集计算激励上界,结合它的支持度与效用值,决定是否删除它。如果 $c \in C_k$ 的激励上界小于阈值 $minmotivation$,将它从 C_k 中删除。只有留在 C_k 中的候选集才需要准确计算激励。与 UMining 不同, HM-Miner 同时利用频繁集向下封闭特性、效用的上界特性和激励的上界特性剪枝。

3 试验与分析

实验在浪潮 XEON 服务器上进行。CPU 主频 2.4 GHz,内存 2 GB,运行 Windows 2003,程序用 Delphi 7 编写。实验用的数据集为 T20.I6.D100Ok,项目数为 1 k,由 IBM 的数据发生器生成。数据集中只有 0 和 1,分别代表某项目是否出现在事务中,没有效用值。因此,实验中用 Delphi 随机函数“RandG”产生随机值(高斯分布)来模拟事务中各项目的单位效用,用事务编号的模 100(Tid mod 100)来表达销售数量。这样,某一事务中某一项目的效用就等于项目的销售数量乘以该项目的单位效用。显然,各项目的单位效用是随机的,这决定了每次挖掘的结果不同。

图 1 显示了事务的变化对算法性能的影响。HM-miner 需要多次扫描数据库,当事务增长时,扫描时间变长,候选集可能增加, HM-miner 运行时间变长。



$minsup=minutil=0.25\%$　　$minmotivation=25\times10^{-6}$

图 1　事务数变化对算法性能的影响

图 2 中, $minsup$ 与 $minutil$ 设为 0.2%,然后 $minmotivation$ 从 2×10^{-6} 变化到 40×10^{-6}。当 $2\times10^{-6} \leqslant minmotivation \leqslant 4\times10^{-6}$ 时,$minmotivation \leqslant minsup \times minutil$ 成立,满足支持度约束(支持度不小于 $minsup$) 和效用约束(效用不小于 $minutil$) 的项集一定满足激励约束(激励不小于 $minmotivation$), 减枝效果与 $minmotivation$ 无关。当 $minmotivation \geqslant 4\times10^{-6}$ 时,算法的减枝效率变为更加依靠激励约束。$minmotivation$ 越大,激励约束的减枝效果越明显。

4 结语

本文分析了基于支持度的关联规则与基于效用的关联规则的不足,提出了基于激励的关联规则挖掘问题。作为一种

新的兴趣度量方法,激励综合了支持度与效用的优点,较好地反映了项集的语义特性与统计特性,能更好地服务于决策。本文还分析了激励约束的性质,证明了激励上界特性的存在,并在HM-miner算法中利用此特性进行减枝。实验证明了算法的正确性与有效性。

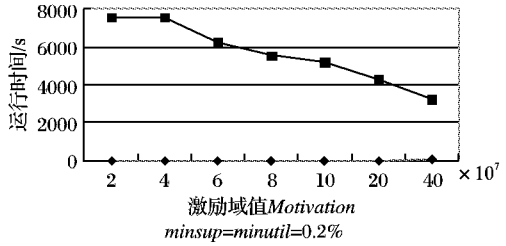


图2 激励的变化对算法性能的影响

参考文献:

[1] AGRAWAL R, SRIKANT R. Fast algorithms for mining association rules [C]// Proceedings of VLDB 1994. Santiago, Chile: VLDB Endowment, 1994: 487 - 499.

[2] LU S F, HU H P, LI F. Mining weighted association rules [J]. Intelligent Data Analysis, 2001, 5(3): 211 - 225.

[3] SHEN Y D, ZHANG Z, YANG Q. Objective - oriented utility - based association mining [C]// Proceedings of the 2002 IEEE International Conference on Data Mining. Maebashi, Japan: IEEE Computer Society, 2002: 426 - 433.

[4] YAO H, HAMILTON H J. Mining itemset utilities from transaction databases [J]. Data & Knowledge Engineering, 2006, 59(3): 603 - 626.

[5] LIU Y, LIAO W K, CHOUDHARY A. A fast high utility itemsets mining algorithm [C]// Proceedings of the 1st International Workshop on Utiliy-based Data Mining. New York: ACM Press, 2005: 90 - 99.

[6] 余光柱,李克清,易先军,等. 一种基于划分的高效用长项集的挖掘算法[J]. 计算机工程与应用, 2007, 43(29): 11 - 13.

[7] VROOM V H. Work and motivation [M]. Hoboken, NJ: John Wiley, 1964.

[8] GENG L, HOWARD J, HAMILTON H J. Interestingness measures for data mining: A survey [J]. ACM Computing Surveys (CSUR), 2006, 38 (3): 61 - 93.

[9] WANG J, LIU Y, ZHOU L, *et al.* Pushing frequency constraint to utility mining model [C] // ICCS: Proceedings of International Conference on Computational Science. Berlin: Springer-Verlag, 2007: 685 - 692.

(上接第180页)

3.3.3 实际检测结果

在真实的环境中经过多次实验,发现在把单端口出现频度阈值设置 k = 600, CCFPM 算法支持数阈值设置 sup = 500 的情况下,算法表现出了比较良好的检测效果,检测率达到80%以上。列举部分检测结果如表3所示(注:"检测时间"是指数据格式的转化以及挖掘的时间之和)。表3中各源IP地址在5 min的时间间隔内,发出的流数都在7 000个以上。

表3 实际检测结果

日 期	源 IP	端口模式	检测时间/s
2007—11—23	202.195.163.138	80	28
2007—12—24	10.2.89.52	80	22
2008—03—02	10.2.73.121	1433	41
2008—03—08	10.2.43.153	80,1755	28
2008—04—20	10.2.100.248	1433	39
2008—06—13	10.2.65.236	139,445	31

4 结语

本文针对骨干网IP流数据环境,通过实时地监控用户的出口流活跃度和访问目的IP地址的增幅来发现可疑源主机,然后将蠕虫攻击的频繁模式特征与频繁模式挖掘算法的特性相结合,提出了一种基于候选组合频繁模式挖掘的网络蠕虫检测算法。实验证明该算法能很好挖掘蠕虫的攻击模式,及时发现新型的蠕虫病毒。

参考文献:

[1] CNCERT/CC. 网络蠕虫灾害对应急处理的挑战[EB/OL]. [2008 - 05 - 01]. http://www.cert.org.cn.

[2] ZOU C C, GONG W, TOWSLEY D, *et al.* The monitoring and early detection of Internet worms [C]// IEEE/ACM Transactions on Networking. [S.l.]: IEEE Press, 2005, 13(5): 961 - 974.

[3] YAMADA Y, KATOH T, BISTA B B, *et al.* A new approach to early detection of an unknown worm [C]// AINAW '07: 21st International Conference on Advanced Information Networking and Applications Workshops. [S.l.]: IEEE Press, 2007, 1: 194 - 198.

[4] DUBENDORFER T, PLATTNER B. Host behaviour based early detection of worm outbreaks in Internet backbones[C]// WETICE '05: Proceeding of the 14th IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprise. [S.l.]: IEEE Press, 2005: 166 - 171.

[5] ZOU C C, GONG W, TOWSLEY D. Code red worm propagation modeling and analysis[C]// CCS '02: Proceeding of 9th ACM Conference on Computer and Communications Security. [S.l.]: ACM Press, 2002: 281 - 287.

[6] 顾荣杰, 晏蒲柳, 邹涛, 等. 基于频繁模式挖掘的Internet骨干网攻击发现方法研究[J]. 计算机科学, 2006, 33(9): 76 - 80.

[7] 王方伟, 张运凯, 王长广, 等. 网络蠕虫的扫描策略分析[J]. 计算机科学, 2007, 34(8): 105 - 108.

[8] CLAFFY K C. Internet traffic characterization. Dissertation for the degree Doctor of Philosophy. University of California[D]. San Diego: University of California, 1994.

[9] HUNG J C, LIN K C, CHANG A Y. A behavior based anti-worm system[C]// AINA '03: Proceeding of the 17th International Conference of Advanced Information Networking and Applications. Washington, DC: IEEE Press, 2003: 812 - 815.

[10] LIU BIN, LIN CHUANG, RUAN DONG - HUA, *et al*. Netflow based flow analysis and monitor [C]// International Conference on Communication Technology. Washington, DC: IEEE Press, 2006: 1 - 4.

[11] PAO TANG-LONG, WANG PO-WEI. NetFlow based intrusion detection system [C]// Proceeding of the 2004 IEEE International Conference on Networking, Sensing Control. Washington, DC: IEEE Press, 2004, 2: 731 - 736.

[12] KANTARDZIC M. 数据挖掘——概念、模型、方法和算法[M]. 闪四清, 陈茵, 程雁, 等译. 北京: 清华大学出版社, 2003: 144 - 153.

[13] Cisco. NetFlow overview[EB/OL]. [2008 - 05 - 01]. http://www.cisco.com.