文章编号: 1001-0920(2003)03-0277-04

关联规则衡量标准的研究

罗 可[1,2], 吴 杰[1]
(1. 湖南大学 电气与信息工程学院, 湖南 长沙 410082; 2. 长沙电力学院 计算机应用教研室, 湖南 长沙 410077)

摘 要: 关联规则采掘是数据采掘中重要的研究课题。针对当前关联规则采掘中可能产生许多无效关联规则的问题, 分析其原因, 提出在衡量标准中增加有效度, 并给出了有效度的定义。根据有效度的大小, 将关联规则分为正关联规则、无效关联规则、负关联规则, 提出了新衡量标准采掘关联规则的算法, 并用 Visual FoxPro 进行了试验。实验表明, 新方法能明显减少无效关联规则的数目。

关键词: 数据采掘; 关联规则; 有效度; 算法
中图分类号: TP311 文献标识码: A

Research on judgment criterion of association rules

LUO Ke[1,2], *WU Jie*[1]
(1. School of Electric and Information Engineering, Hu'nan University, Changsha 410082, China; 2. Division of Computer Application, Changsha University of Electric Power, Changsha 410077, China)

Abstract: Mining association rules are an important topic in the data mining research. The reasons for many invalid rules in mining association rules are analyzed. The validity is defined and added to the judgment criterion. According to the value of validity, association rules are classified into positive, invalid and negative association rules. An algorithm of new judgment criterion in mining association rules is presented and tested with Visual FoxPro. The test results show that the method can obviously reduce invalid association rules.

Key words: Data mining; Association rules; Validity; Algorithm

1 引 言

在事务(Transaction) 数据库中采掘关联规则是数据采掘领域中一个非常重要的研究课题, 它是由 Agrawal 等人首先提出的[1]。关联规则的采掘问题可形式化描述为: 设 $I = \{i_1, i_2, \cdots, i_m\}$ 是 m 个不同项目的集合, D 是针对 I 的事务的集合, 每一笔事务包含若干项目 $i_i, i_j, \cdots, i_k \in I$。关联规则表示为 $X \Rightarrow Y$, 其中 $X \subset I$, $Y \subset I$, 且 $X \cap Y = \varnothing$。X 称作规则的前提, Y 是结果。一般把一些项目的集合称作项目集(itemset), 在项目集中项目的数量称作项目集的长度。关联规则 $X \Rightarrow Y$ 成立的条件是: 1) 它具有支持度 s, s 是 D 中事务同时包含 X 和 Y 的百分比, 即概率 $P(XY)$; 2) 它具有置信度 c, c 是 D 中包含 X 的事务同时也包含 Y 的百分比, 即条件概率 $P(Y|X)$。

关联规则的采掘问题就是在事务数据库 D 中找出具有用户给定的最小支持度 minsup 和最小置信度 minconf 的关联规则, 因此, 该问题可分解成如下两个子问题:

1) 找出存在于事务数据库中的所有频繁项目集。若项目集 X 的支持度 support(X) 不小于用户给定的最小支持度 minsup, 则称 X 为频繁项目集; 否则, 为非频繁项目集。

收稿日期: 2002-02-27; 修回日期: 2002-04-19。
基金项目: 国家自然科学基金资助项目(10171030); 湖南省科技厅资助项目; 湖南省教育厅资助科研项目。
作者简介: 罗可(1961—), 男, 湖南长沙人, 教授, 博士生, 从事数据库技术、数据采掘的研究; 吴杰(1957—), 男, 湖南长沙人, 教授, 博士生导师, 从事自动化技术的研究。

2) 利用频繁项目集生成关联规则。对于每个频繁项目集 A, 若 $B \subset A, B \neq \varnothing$, 且 $\text{support}(A)/\text{support}(B) \geq \text{minconf}$, 则有关联规则 $B \Rightarrow (A - B)$。

子问题 2) 相对来说较为容易, 其生成算法可参见文献[2]。目前研究的重点集中在: 子问题 1) 的高效算法[3], 在分布式数据库中采掘关联规则, 发现有实际意义的关联规则等。本文讨论采掘有实际意义的布尔关联规则。

2 当前关联规则衡量标准的不足

目前衡量和生成关联规则的标准有 2 个, 即支持度(support) 和置信度(confidence), 按现有方法来生成关联规则, 可能会发现大量冗余的、虚假的关联规则[4]。先来看一个实例: 设想在一个底层事务数据库中有 N 条记录, 只讨论这 N 条记录购买咖啡和牛奶的情况, 如表 1 所示。

表 1 购买咖啡和牛奶的统计表

	购牛奶人数	不购牛奶人数
购咖啡人数	20	5
不购咖啡人数	70	5

现研究关联规则 咖啡 ⇒ 牛奶。支持度 $S = 20/100 = 0.20$, 置信度 $C = 20/25 = 0.8$。当把置信度和支持度阈值定位低于 0.8 和 0.2 时, 该规则将会作为目标规则之一被采掘出来。由此可得出结论, 将咖啡和牛奶放在一起会提高牛奶的销售量。然而, 事实并非这样。原始事务库中有 90% 的顾客会购买牛奶, 而从上述采掘出的关联规则可知, 买咖啡的顾客有 80% 的可能性购买牛奶。也就是说, 一个已知买了咖啡的顾客购买牛奶的可能性比一个我们不知道任何信息的顾客购买牛奶的可能性小。事实上, 不买咖啡会买牛奶的可能性更大, 其置信度 $C = 70/75 = 0.933$。

从上例可看出, 置信度和支持度很高的规则, 并非一定有实际利用价值。再看一个实例: 设有一组事务数据, 如表 2 所示。

我们仅讨论其中的若干长度为 2 的项目集。从直观上看, C 和 D 总是同时出现或不出现, 应该是有利用价值的关联规则, 计算得到 $C \Rightarrow D$ 的支持度和置信度分别是 0.5 和 1; 再来看 A 和 B, 不管 A 是否出现, B 总是出现, 应该不构成有利用价值的关联规则, 计算得到 $A \Rightarrow B$ 的支持度和置信度分别是 0.5 和 1。即 $C \Rightarrow D$ 与 $A \Rightarrow B$ 具有相同的支持度和置信度, 是等同的关联规则。显然, 上述结论是不正确的。

表 2 一组事务数据

TID	ITEMS
101	B, C, D, E, F
102	B, E, F, H, J
103	B, C, D, F, G, J
104	A, B, C, D, F, I
105	A, B, E, F, H, I
106	A, B, F, J
107	A, B, C, D, F, G, H, J
108	A, B, F, G, H, J
109	B, C, D, E, F, G, I
110	B, E, F, G

针对关联规则衡量标准中存在的问题, 文献[5] 给出了感兴趣的规则定义, [6] 又对此作了一些改进, [7] 则定义了负关联规则的兴趣度。然而, 上述文献中提到的方法仍未能很好地解决关联规则衡量标准中存在的问题。

3 改进关联规则衡量标准的方法

针对关联规则采掘中的上述问题, 其根本原因是关联规则 $X \Rightarrow Y$ 的置信度的定义只考虑了 X 出现时 Y 出现的可能性, 而未考虑 X 不出现时 Y 出现的可能性, 使得采掘出来的关联规则不能全面反映一些项目集对另一些项目集的影响程度, 从而导致采掘出来的某些关联规则是无效的。本文建议采用下列方法来解决。

3.1 引入新的关联规则衡量标准: 有效度

采掘关联规则时, 在原有关联规则衡量标准的基础上, 再引入有效度(Validity, 笔者自行定义)。若用 $P(X)$ 表示事务 X 发生的概率, 用 $P(\bar{X})$ 表示事务 X 不发生的概率, 则定义 $X \Rightarrow Y$ 的有效度为

$$\text{Validity} = P(XY) - P(\bar{X}Y)$$

上式的直观意义为: 有效度 = (在 D 数据库中 X 和 Y 同时出现的概率) - (在 D 数据库中 $\bar{X}$ 和 Y 同时出现的概率)。由于 $P(XY)$ 和 $P(\bar{X}Y)$ 的值区间均在 $[0, 1]$, 显然, 有效度的值区间在 $[-1, 1]$。

3.2 根据有效度将关联规则分为 3 类

以表 2 中的若干 2 项目集来讨论有效度。

1) 表 2 中每条记录中均有 B 和 F, $B \Rightarrow F$ 的有效度为

$$\text{Validity} = 10/10 - 0 = 1$$

2) 表 2 中每条记录均有 B, 但都没有 K, $K \Rightarrow B$ 的有效度为

$$\text{Validity} = 0 - 10/10 = -1$$

显然, B 和 K 不可能构成2项目频繁集, 生成关联规则时也不会计算此 2 项目集的有效度, 该例只是为了说明Validity = - 1时应满足的条件。也就是说, 关联规则有效度的值区间只能是(- 1, 1]。

3) 表2中有5条记录中包含 A, 另5条记录中不包含 A, 所有记录中均包含 B, $A \Rightarrow B$ 的有效度为

$$\text{Validity} = 5/10 - 5/10 = 0$$

4) 表 2 中有 5 条记录同时包含 C 和 D, 其余记录中不包含 C 或 D, $C \Rightarrow D$ 的有效度为

$$\text{Validity} = 5/10 - 0 = 0.5$$

5) 表 2 中, 5 条含有 C 的记录中有 2 条含有 J, 另 5 条未含 C 的记录中有 1 条含有 J, $C \Rightarrow J$ 的有效度为

$$\text{Validity} = 2/10 - 1/10 = 0.1$$

6) 表2 中, 5 条含有 C 的记录中有 1 条含有 H, 另5 条未含 C 的记录中有2 条含有 H, $C \Rightarrow H$ 的有效度为

$$\text{Validity} = 1/10 - 2/10 = -0.1$$

从上面分析和有效度的计算不难看出, 有效度可能的取值区间为(- 1, 1]。为此, 引入有效度的阈值 v_1 和 v_2, 他们均是大于 0 小于 1 的正数, 通常接近于 0。v_1 和 v_2 可以相等, 也可以不等。关联规则的有效度必然是下列 3 种情况之一: 1) $v_1 \leq \text{Validity} \leq 1$; 2) $-v_2 < \text{Validity} < v_1$; 3) $-1 < \text{Validity} \leq -v_2$。

根据上述有效度的取值范围, 将满足支持度和置信度要求的关联规则分为 3 类: 正关联规则、无效关联规则和负关联规则。

3.2.1　正关联规则

正关联规则满足下列条件:

1) 支持度不小于用户给定的支持度阈值;

2) 置信度不小于用户给定的置信度阈值;

3) 有效度不小于用户给定的正关联规则阈值 v_1。

由于正关联规则的有效度大于 0, 说明在 D 数据库中 X 和 Y 同时出现的概率大于 $\bar{X}$ 和 Y 同时出现的概率, 这类规则通常是有效的关联规则, 能够直接指导用户的决策过程。比如, 在医学应用中, 我们获得了正关联规则 $X \Rightarrow Y$, 其中 X 代表 "经常吸烟", Y 代表 "肺病患者", 则可直接得出 "经常吸烟的人更容易得肺病" 的结论。在以表 2 讨论的 6 种情况中, 1), 4), 5) 可能成为正关联规则。

在大多数情况下, 可只输出正关联规则。比如, 在事务数据库中采掘关联规则的目的就是要讨论顾客在购买某些商品的时候有多大倾向会购买另外一些商品, 其本质是讨论一些商品的销售对另外一些商品销售的影响程度。因此, 此类关联规则是用户最关心的。

3.2.2　无效关联规则

无效关联规则满足下列条件:

1) 支持度不小于用户给定的支持度阈值;

2) 置信度不小于用户给定的置信度阈值;

3) 有效度的值区间为($-v_2, v_1$)。

此类关联规则的有效度接近于 0, 说明在 D 数据库中 X 和 Y 同时出现的概率与 $\bar{X}$ 和 Y 同时出现的概率基本相等, 这类规则通常是无效的关联规则。在以表 2 讨论的 6 种情况中, 3) 是无效关联规则, 此外, 某些有效度接近于 0 的关联规则, 根据 v_1 和 v_2 的取值大小, 也可能成为此类关联规则。

3.2.3　负关联规则

负关联规则满足下列条件:

1) 支持度不小于用户给定的支持度阈值;

2) 置信度不小于用户给定的置信度阈值;

3) 有效度不大于用户给定的负关联规则阈值 $-v_2$。

由于负关联规则的有效度小于 0, 说明在 D 数据库中 X 和 Y 同时出现的概率小于 $\bar{X}$ 和 Y 同时出现的概率, 这类规则通常不能直接指导用户的决策过程, 但有时可通过反例来指导用户。比如, 在医学应用中, 我们获得了负关联规则 $X \Rightarrow Y$, 其中 X 代表 "经常参加体育锻炼", Y 代表 "高血压患者", 上述结果说明: "经常参加体育锻炼" 和 "高血压患者" 的百分率较大, 由于它是负关联规则, 则可从 X 的反例间接得出 "不经常参加体育锻炼的人更容易得高血压" 的结论。在以表 2 讨论的 6 种情况中, 6) 可能成为负关联规则, 然而 2) 不可能成为负关联规则, 因为它不可能满足关联规则的支持度的要求。

下面再来计算表 1 中咖啡 $\Rightarrow$ 牛奶的有效度, 不难发现它确实不是正关联规则。

$$\text{Validity} = 20/100 - 70/100 = -0.5$$

3.3　算法描述

在关联规则采掘中引入有效度后, 求频繁项目集的算法保持不变, 生成关联规则的算法将作适当修改。修改后的生成关联规则算法简述如下:

Step1: 输入置信度阈值和有效度阈值;

Step2: 由频繁项目集生成所有可能的候选关

表 3 支持度阈值为 0.2 时表 2 数据产生的频繁项目集的数目

1 项目频繁集	2 项目频繁集	3 项目频繁集	4 项目频繁集	5 项目频繁集	6 项目频繁集
10	38	61	45	15	2

表 4 支持度阈值为 0.3 时表 2 数据产生的频繁项目集的数目

1 项目频繁集	2 项目频繁集	3 项目频繁集	4 项目频繁集	5 项目频繁集
10	24	23	9	1

表 5 以表 2 数据生成的关联规则数目

支持度阈值	置信度阈值	有效度 v_1	有效度 v_2	关联规则总数	正关联规则数	无效关联规则数	负关联规则数	正关联规则占总数的百分比/%
0.2	0.6	0.01	0.01	842	272	74	496	32
0.2	0.6	0.11	0.11	842	93	574	175	11
0.2	0.9	0.01	0.01	472	74	56	342	16
0.2	0.9	0.11	0.11	472	21	276	175	4
0.3	0.6	0.01	0.01	262	182	35	45	69
0.3	0.6	0.11	0.11	262	50	161	45	19
0.3	0.9	0.01	0.01	118	38	35	45	32
0.3	0.9	0.11	0.11	118	20	53	45	17

联规则;

Step3: 扫描事务数据库 D, 分别计算各候选关联规则的置信度;

Step4: 删除其置信度低于置信度阈值的候选关联规则;

Step5: 扫描事务数据库 D, 分别计算各关联规则的有效度;

Step6: 将所有正关联规则、无效关联规则、负关联规则分别存入不同的数据表中, 根据用户需要, 输出所有正关联规则或全部关联规则, 也可分别输出 3 类关联规则的数目。

注 **1** 1) 在 Step2 中, 一项不同长度的频繁项目集生成的候选关联规则的数目是不同的。一项长度为 N 的频繁项目集生成的候选关联规则数目是 $C_N^1 + C_N^2 + \cdots + C_N^{N-1}$ 条。

2) 上述算法仅仅是本文提供的一种可行算法。用户可根据实际对该算法进行修改。

3.4 实验结果

用 Visual FoxPro 对表 2 中的数据进行了多次实验, 实验结果基本相同(运行环境为: Windows98, Visual FoxPro5.0)。实验方法如下:

Step1: 输入支持度阈值, 用 Apriori 算法[2] 求出所有长度为 1, 2, 3, … 的频繁项目集, 然后将不同长度的频繁项目集分别存入不同的数据表中。

Step2: 输入置信度阈值, 通过长度为 2, 3, 4, … 的频繁项目集生成候选的关联规则, 分别计算各候选关联规则的置信度, 删除其置信度低于置信度阈值的候选关联规则。

Step3: 输入有效度阈值, 分别计算各关联规则的有效度, 分别输出正关联规则、无效关联规则、负关联规则以及 3 类规则的数目。

表 3 ~ 表 5 是若干组实验数据。

在大多数情况下, 用户关心的是正关联规则, 而正关联规则的数目一般只占关联规则数目的一小部分, 当有效度的阈值较大时更是如此。实验表明, 用该方法能明显减少无意义的关联规则的数目。

4 结 语

本文针对当前关联规则采掘中产生许多无效关联规则的问题, 分析了产生的根源, 提出了改进方法, 即在关联规则衡量标准中增加有效度。有效度的值区间为(-1, 1], 根据关联规则有效度值的大小, 将关联规则分为正关联规则、无效关联规则和负关联规则。正关联规则通常是有效的关联规则; 无效关联规则通常是无意义的关联规则; 负关联规则通常是无效的关联规则, 但其反例在某些场合可能是有效的关联规则。一般来说, 只有正关联规则才是有效的关联规则, 它通常只占关联规则总数的一小部分。

用本文方法来改进关联规则采掘, 先前求频繁项目集的方法可继续使用, 而生成关联规则的方法将会作适当修改。对本文提出的在关联规则采掘中引入有效度的改进算法进行了实验, 实验表明该方法能明显减少无效的关联规则。

(下转第 284 页)

Control of Dynamic Systems [M]. Mass: Addison-Wesley, 1986.

[2] 王广雄, 王新生. 鲁棒镇定问题的 H_∞ 优化设计[J]. 自动化学报, 2002, 28(4): 601-605.
(Wang Guangxiong, Wang Xinsheng. H_∞ Optimal design for robust stabilization [J]. *Acta Automatica Sinica*, 2002, 28(4): 601-605.)

[3] Boyd S, Ghaoui L El, Feron E, et al. *Linear Matrix Inequalities in System and Control Theory* [M]. Philadelphia: SIAM, 1994.

[4] Wang Shaopeng, Chow J H. Low-order controller design for SISO systems using coprime factors and LMI[J]. *IEEE Trans Automatic Control*, 2000, 45(6): 1166-1169.

[5] Gahinet P, Nemirovski A, Laub A J, et al. *LMI Control Toolbox* [M]. Mass: The Math Works Inc, 1995.

[6] Gahinet P, Apkarian P. A linear matrix inequality approach to H_∞ control [J]. *Int J Robust and Nonlinear Control*, 1994, 4(1): 421-448.

(上接第 276 页)

参考文献(References):

[1] Vapnik V. *The Nature of Statistical Learning Theory* [M]. New York: Springer, 1998.

[2] Vapnic. 统计学习理论的本质[M]. 张学工, 译. 北京: 清华大学出版社, 2000.

[3] Platt J C. Fast training of SVMs using sequential minimal optimization[A]. *Advances in Kernel Methods Support Vector Learning* [C]. MIT Press, 1998. 185-208.

[4] 马笑潇. 智能故障诊断中的机器学习新理论及其应用[D]. 重庆: 重庆大学, 2002.

[5] CHIH-WEI HSU, CHIH-JEN LIN. A comparison of methods for multi-classification support vector machines[EB/OL]. http://www.csie.ntu.edu.tw/~cjlin/bsvm/.

[6] Weston J, Watkins C. Multi-class support vector machines[A]. *Proc of ESANN'99*[C]. Brussels D Facto Press, 1999.

[7] 耿尊敏, 宋孔杰, 李兆前, 等. 关于柴油机振声特点及动态诊断方法的研究与讨论[J]. 内燃机学报, 1995, 13(2): 140-147.
(Geng Z M, Song K J, Li Z Q, et al. The research and discussion on the characteristic of vibration of diesel engine and dynamic diagnosis methods [J]. *Trans CSICE*, 1995, 13(2): 140-147.)

(上接第 280 页)

参考文献(References):

[1] Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large databases[A]. *Proc of ACM SIGMOD Conf on Management of Data*[C]. Washington, 1993. 207-216.

[2] Agrawal R, Srikant R. Fast algorithms for mining association rules[A]. *Proc of the 20th Int Conf on Very Large Databases*[C]. Santiago, 1994. 487-499.

[3] Agrawal R, Mannila H, Srikant R, et al. Fast discovery of association rules[A]. *Advances in Knowledge Discovery and Data Mining*[C]. AAAI/MIT Press, 1996. 307-328.

[4] Brin S, Motwani R, Silverstein C. Beyond market basket: Generalizing association rules to correlations[A]. *Proc 1997 ACM-SIGMOD Int Conf Management of Data*[C]. Tucson, 1997. 265-276.

[5] Srikant R, Agrawal R. Mining generalized association rules[A]. *Proc of the 21st Int Conf on Very Large Data Bases*[C]. Zurich, 1995. 407-419.

[6] Srikant R, Agrawal R. Mining quantitative association rules[A]. *Proc of the ACM SIGMOD*[C]. Montreal, 1996. 1-12.

[7] Savasere A, Omiecinski E, Navathe S. Mining for strong negative association in a large database of customer transactions[A]. *Proc of the Int Conf on Data Engineering*[C]. Orlando, 1998. 494-502.