文章编号: 1674- 7046(2010) 06- 0063- 4

数据挖掘中关联规则衡量方法的改进

赵军民, 王军豪, 高 蔚
(河南城建学院, 河南 平顶山 467036)

摘 要: 对数据挖掘中的关联规则的有趣性进行了详细的阐述, 发现并指出传统的支持度- 信任度框架的不足之处, 提出了一种对关联规则进行综合衡量的新方法, 并对此方法进行了详尽的分析和阐述, 并在此基础上对强关联规则重新进行了定义。

关键词: 关联规则; 支持度; 信任度
中图分类号: TP301. 6 文献标识码: A

1 关联规则

假设 I 是一个项集, 对于一个特定的事务数据库 D, 其中每个事务 T 是 I 的非空子集, 即每一个交易都与一个唯一的标识符 TID(Transaction ID) 对应。对项集 I 和事务数据库 D, T 中所有满足用户指定的最小支持度 Minsup(Minsupport) 的项集(即大于或者等于 Minsup 的项集 I 的非空子集) 称为频繁项集(Frequent Itemsets) 。在频繁项集中挑选出所有不被其他元素包含的频繁项集称为最大频繁项集(Maximum Frequent Itemsets) 。事务数据库 D 在项集 I 上满足最小支持度以及最小信任度 Minconf(Minconfridence) 的关联规则被称为强关联规则, 否则就被称为弱关联规则[1]。

2 基于兴趣度约束的关联规则

在数据挖掘的过程中, 通常情况下可能会挖掘出上千条的规则, 但是并不是所有的规则都对用户有用。为了提高挖掘出的规则的质量, 使之对用户来说更新颖、更容易理解, 就需要一个更有效的挖掘算法。为了以用户的需求和兴趣为最大目标, 尽量减少扫描次数, 将无意义的和虚假的规则剔除掉, 这里引入了一种约束条件- 兴趣度约束。当前最常见、也是最常用的两个兴趣度度量是支持度和信任度, 但是仅靠这两个度量标准还存在一定的缺陷[2]:

⑴会产生大量的规则, 而其中的大部分是显而易见的或不相关的;

⑵用户没有充分利用领域知识和职业直觉。用户的职业直觉往往对知识发现的过程具有重大的价值;

⑶没有提供好的度量令人感兴趣程度的方法, 而从数据中发现令人感兴趣的规则是数据挖掘的一个重要目标。

在实际应用中, 挖掘的关联规则可能因为以下原因失去有趣性: ①挖掘的规则符合先验知识或期望值; ②挖掘的规则可能涉及非有趣属性或属性组合; ③规则冗余。本文主要讨论关联规则令人感兴趣程度的度量问题, 并给出一种新的综合度量方法。

3 关联规则的综合度量

要看一条规则是不是有趣的, 要从确定性、有用性、非预期性和简洁性几个方面进行综合的度量[3]。

3. 1 确定性

确定性是规则的有效性以及值得信赖程度的反映。在挖掘过程中, 对于像 A ⇒B 这样的关联规则,

收稿日期: 2010- 09- 28
第一作者简介: 赵军民(1978-) , 男, 河南平顶山人, 硕士, 河南城建学院讲师。

它的确定性度量就是信任度。信任度是对关联规则的准确度的衡量,表示一个商品的购买暗示着另一个商品的购买。

3.2 有用性

有用性是用来衡量一条规则是否具有有趣性的一个重要因素,它可以用支持度来进行度量。支持度表示用这条规则可以推出百分之几的目标,它也是对关联规则重要性的度量,支持度说明了这条规则在所有事务中占多大的代表性。

3.3 简洁性

简洁性是针对规则的形式而言的,一般指规则的总体简洁性,是用来衡量关联规则的最终可理解程度的指标,并用规则的属性个数或者规则中出现的操作符来进行定义的[4]。它表现在两个方面:一方面表现在规则的项数上,如果规则项数很多,会不利于对这条规则的理解。因此,规则的项数越少,规则的简洁性越好。另一方面表现在规则所包含的抽象层次上,规则包含的抽象层次越高,它对应的解释力越强。

3.4 非预期性

具有非预期性的规则是那些提供新的信息或者跟先验知识相矛盾的规则。非预期的规则出乎客户的意料,与用户的期望相矛盾。传统的关联规则挖掘算法中,用条件概率 $P(B \mid A) = P(A \cup B) / P(A)$ 来表示,也就是用信任度作为约束对规则的非预期性进行判断,但它只是给定的 A 和 B 的条件概率的估计值,而并没有度量 A 和 B 之间蕴涵的实际强度。我们将项集 I_1 和 I_2 之间的影响程度用提升度来进行度量。

定义 1:提升度(lift)是利用相关分析来描述规则内在价值的度量,它所描述的是项集 I_1 对 I_2 的影响力的大小。提升度越高,表示 I_1 的出现对 I_2 出现的可能性影响越大,它是对 I_1 和 I_2 之间蕴涵的实际强度的度量。提升度是一个比值:

$$lift(I_1 \Rightarrow I_2) = \frac{Conf(I_1 \Rightarrow I_2)}{Sup(I_2)} = \frac{Sup(I_1 \Rightarrow I_2)}{Sup(I_1)\,Sup(I_2)}$$

也可以用 $P(I_2 \mid I_1) / P(I_2)$ 来表示。它又分为两种情况:

(1)当 lift> 1 时,表示 I_1 和 I_2 是相关的,代表 I_1 的出现蕴涵了 I_2 的出现,此规则是非预期的、有效的或者有趣的。

(2)当 lift< 1 时,表示 I_1 和 I_2 是不相关的,规则不是非预期的,是无效的、无趣的,应将其删除。

提升度的值越大,表明两者之间的相关性就越强,规则越有效、越有趣,其利用的价值也就越大[5]。

根据上面规则的度量标准,需要对强关联度重新进行定义,定义如下所示:

定义 2:对于事务数据库 D,如果 $I_1 \Rightarrow I_2$ 能同时满足以下条件,则称之为强关联规则,否则就称之为弱关联规则,其中 Maxlen 为最大规划长度。

$Sup(I_1 \Rightarrow I_2) \geqslant Minsup$

$Conf(I_1 \Rightarrow I_2) \geqslant Minconf$

$Len(I_1 \Rightarrow I_2) \leqslant Maxlen$

$lift(I_1 \Rightarrow I_2) \geqslant 1$

那么,在事务数据库 D 中挖掘关联规则的问题也就变为:产生所有支持度、信任度分别大于最小支持度和最小信任度,规则长度小于最小规则长度,并且提升度的值大于 1 的关联规则,也就是找出所有的强关联规则。

4 基于综合度量的关联规则算法设计

以 Apriori 算法为基础,保留原有的最小支持度、最小信任度这两个衡量指标,并将提升度(lift)以及最大规则长度这两个衡量标准引入到算法中,使之挖掘出更有趣、有效的关联规则。

算法 1:产生长度不超过 Maxlen 的频繁项集

输入:事务数据库 D;最小支持度阈值 Minsup;最大规则长度的阈值 Maxlen

输出:频繁项集 L

L= Φ

L_1= {频繁 1- 项集}

for k= 2 to Maxlen then

C_k= Apriori- gen(L_{k-1}, Minsup)

for 所有事务 t∈D then

C_t= subset(C_k, t)

for 所有事务 C∈Ct then c. count = c. count+ 1

next

L_K= {c∈C_K| c. count/ | D| ≥Minsup }

next

L= L∪L_K

算法 2: 产生有趣的关联规则

输入: 频繁项集 L; 最小支持度阈值 Minsup; 最小信任度阈值 Minconf; 提升度阈值 lift; 最大规则长度的阈值 Maxlen

输出: 强关联规则 R

R= Φ

For k= 2 to Maxlen then

{频繁项集 L_k 的所有子集

for 每一个 s 都属于 S_K

If 信任度 ≥Minconf and 提升> 1 then 输出强关联规则 R

}

5 算法实现及性能分析

使用 Microsoft Visual Basic 实现 Apriori 算法, 并引入支持度、信任度、提升度及规则长度相结合的综合度量标准。此程序可以任意输入支持度、信任度、提升度和规则长度的阈值, 并统计运行过程所耗费的时间。

程序界面如图 1 所示。

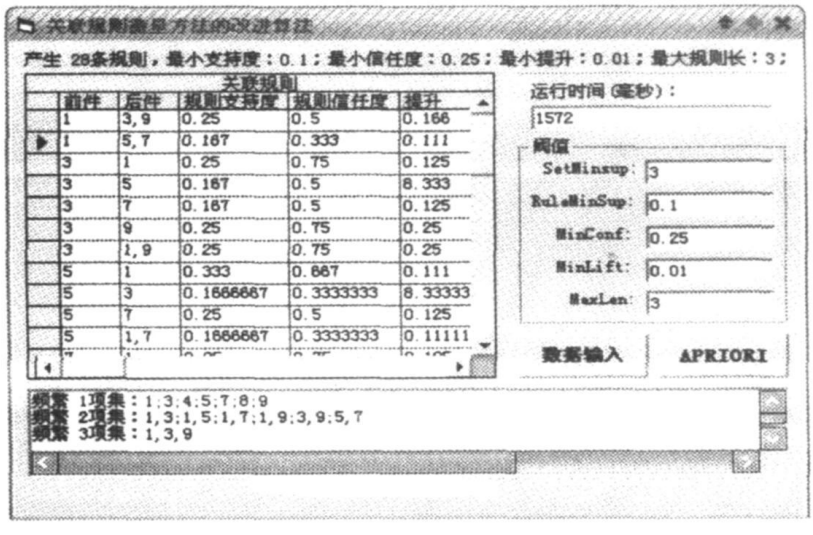

图 1　关联规则算法运行界面

由于 Apriori 算法在运行过程中需要多趟扫描数据库, 使得算法的运行非常耗时。可以分别利用事务数据量为 100、200、500、1 000、2 000、5 000、10 000 和 50 000 的数据集在相同环境下对程序进行测试,

结果如表 1 和图 2 所示。

表 1 算法运行时间

事务数据量	运行时间/ms	事务数据量	运行时间/ms
100	701	2 000	62 280
200	1 822	5 000	246 805
500	7 551	10 000	661 331
1 000	20 850	50 000	4 629 568

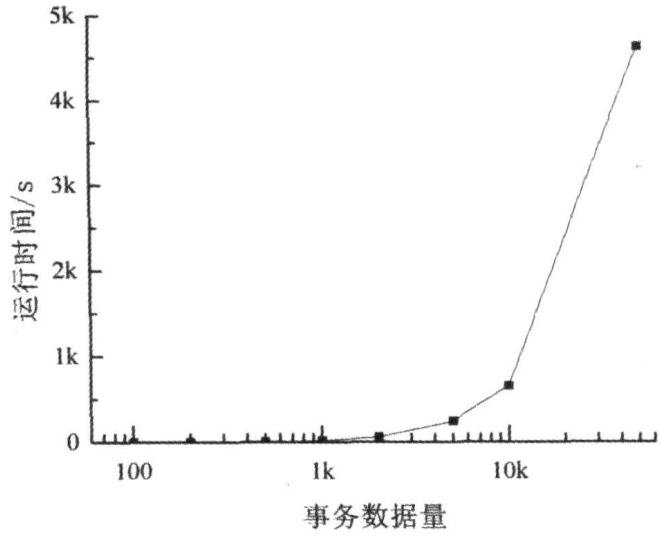

图 2 算法运行效率

表 1、图 2 利用综合衡量的方法处理经过整理的数据。这里设最小支持度为 0.1, 最小信任度为 0.25, 提升度(lift) 应大于或等于 1, 最大规则长度为 3。通过运行后得到了有效的关联规则, 与单纯的支持度– 信任度相比, 减少了无效的关联规则。表 2 是列出的部分关联规则。

表 2 关联规则

id	sup	conf	lift	len	rule	remark
1	0.28	0.63	2.23	2	105⇒961	洗面奶⇒面霜
2	2.21	0.46	2.21	2	4 616⇒732	牙膏⇒牛奶
3	0.30	0.52	1.87	2	3 946⇒4 616	洗发水⇒牙膏
4	0.27	0.45	1.67	3	245/423⇒1 830	熟肉/火腿⇒糕点
5	0.29	0.62	2.10	3	423/67⇒2 511	火腿/方便面⇒饼干
6	0.29	0.50	1.21	2	398⇒3 946	口香糖⇒洗发水
7	0.22	0.43	1.38	2	1 732⇒3 946	牛奶⇒洗发水
8	0.28	0.57	2.73	3	61/3 946⇒841	护发素/洗发水⇒沐浴露
9	0.23	0.37	1.61	2	500⇒32	卫生巾⇒牛奶
10	0.26	0.45	1.58	2	415⇒1 920	瓜子⇒薯片

6 小结

分析表 1 和图 2 可知, 随着数据量的变大, 算法所需时间呈指数上升趋势。因为需要多次扫描数据库, 所以 I/O 负载太大, 对每次 K 循环, 候选项集中的每个元素都必须通过扫描一次数据库来进行验证是否加入频繁项集。另外, 算法还有可能产生非常庞大的候选项集, 太大的候选项集对时间和主存空间都是一种很大的挑战。所以, 该算法虽然有一定改善, 仍需要进一步的改进。

(下转第 70 页)

参考文献

[1] 陈坚. 电力电子学- 电力电子变换和控制技术[M]. 北京: 高等教育出版社, 2002.

[2] 吴茂刚, 赵荣祥, 汤新舟. 空间矢量 PWM 逆变器死区效应分析与补偿方法[J]. 浙江大学学报, 2006, 40(3): 469- 473.

[3] 杨贵杰. 空间矢量脉宽调制方法的研究[J]. 中国电机工程学报, 2001, 21(5): 79- 83.

[4] 李波, 安群涛, 孙兵成. 空间矢量脉宽调制的仿真研究及其实现[J]. 电机与控制应用, 2006, 33(6): 40- 44.

[5] 陈建业. 电力电子电路的计算机仿真[M]. 北京: 清华大学出版社, 2003.

Investigation of Space Vector PWM based on PSPICE

LIU Na, ZHANG Xu-hui

(Henan University of Urban Construction, Pingdingshan 467036, China)

Abstract: Based on the principle of space vector pulse width modulation, simulation of it was analyzed and implemented using ABM in PSPICE and the wave of the theory was obtained. The foundational features of the harmonic distributions of SVPWM and the dominant factors affecting the distributions are obtained through the analysis on the harmonics of the waveforms, which provides us theoretical foundation to eliminate the harmonic pollution. Finally, experimental results have also been given to verify the efficacy of the method.

Key words: space vector pulse width modulation; PSPICE /ABM; harmonic analysis

---

(上接第 66 页)

参考文献

[1] 陈景民. 数据仓库与数据挖掘技术[M]. 北京: 电子工业出版社, 2002.

[2] B Padmanab han, A Tu zhilin. Unexpectedness as a measure of interestingness in knowledge discovery[J]. Decision Support System, 1999: 303- 318.

[3] Fayyad U, Piantesky- shapiro G. From Data Mining to Knowledge Discovery. Advances in Knowledge Discovery and Data Mining [J]. California: AAAI Press, 1996: 1- 363.

[4] Jiawei Han, Micheline Kamber. 数据挖掘概念与技术[M]. 范明, 孟小峰, 译. 北京: 机械工业出版社, 2001.

[5] 吴永梁, 陈烁. 基于改善度计算的有效关联规则[J]. 计算机工程, 2003, 29(13): 98- 100.

Improvement of association rules measure method in data mining

ZHAO Jun-min, WANG Jun-hao, Gao Wei

(Henan University of Urban Construction, Pingdingshan 467036, China)

Abstract: Data mining of association rules were discussed in detail. The author pointed out that the inadequacies of traditional support- trust framework, and proposed a comprehensive measure for the association rules with detailed analysis and elaboration, and the strong association rules are redefined on this basis.

Key words: association rule; support; trust