一种改进的关联规则挖掘方法研究

徐 勇,周森鑫
(安徽财经大学 信息工程学院,安徽 蚌埠 233041)

摘 要:关联模式挖掘研究是数据挖掘研究领域的重要分支之一,旨在发现模式之间存在的关联或相关关系。然而,传统的基于支持度-可信度框架的挖掘方法存在着一些不足:一是会产生过多的模式(包括频繁项集和规则);二是挖掘出来的规则有些是用户不感兴趣的、无用的,甚至是错误的。所以在挖掘过程中能有效地对无用模式进行剪枝是必要的。利用相关关系对模式进行评价是一种有效的剪枝方法。实验结果分析表明,在传统挖掘方法的基础上引入相关关系度量可以有效地对非相关模式进行剪枝,从而减小频繁项集和规则的规模。

关键词:数据挖掘;关联规则;相关关系;兴趣度
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1005-3751(2006)03-0077-03

Research on an Improved Approach for Association Rules Mining

XU Yong, ZHOU Sen-xin
(School of Information Engineering, Anhui University of Finance & Economics, Bengbu 233041, China)

Abstract: Association rules mining is an important branch of research on data mining, its purpose is to find the association or correlation among items. However the common approach based on support-confidence framework has some shortcomings: Firstly, there are a great number of redundant association patterns (including frequent items and rules), then it is difficult for user to find interesting association rules in them; Secondly, some of these rules are uninteresting. Therefore it is necessary to prune the useless rules effectively, where it is a valid approach to evaluate the rules via correlation. The experimental result shows that introducing correlation measuring based on common approach to association rules mining could prune the unwanted rules, reduce the scale of frequent items and rules.

Key words: data mining; association rule; correlation; interestingness

0 引 言

在过去的几十年内,随着微电子技术、信息技术的发展,一方面计算机硬件稳定、飞速的进步导致了计算机系统在数据采集、存储方面的功能日益强大;另一方面,数据库系统的研究开发经过从层次、网状数据库系统到关系数据库技术的广泛被接受,数据库和信息产业亦取得了很大的发展,大量被存储在数据库系统中的数据被广泛地应用于事务处理、信息检索、分析等领域。然而,尽管这些重要的商业数据被较为完整地存储起来了,但现有的数据库系统却没有为用户提供必要的进一步处理这些海量数据的功能,从而阻碍了将这些宝贵的数据真正融合到现有的数据库系统应用中。正是在这种"数据的丰富带来了对强有力的数据分析工具的需求"[1]的背景下,数据库中知识发现技术(KDD,Knowledge Discovery in Database)应运而生,并得到迅速发展,成为数据库研究、开发和应用最活跃的分支之一。

关联分析[2](association analysis)是数据挖掘的一项重要功能,典型应用是购物篮或事务数据分析。关联分析旨在发现隐藏在大规模数据集中属性之间的关联模式,反映属性之间频繁关联或同时发生的模式。这些模式可以用关联规则(association rules)的形式表示,如[1]:

computer⇒financial_management_software[support=2%,confidence=60%]

R. Agrawal 等在文献[2,3]中提出关联规则挖掘问题并介绍了一个有效的基于频集思想的关联规则挖掘算法——Apriori 算法。随后,国内外的研究人员在此基础上又提出了许多其他的有效的关联规则挖掘算法,这些算法绝大多数也都是基于传统频集思想的。然而实践表明,根据传统的基于频集思想的关联规则挖掘算法挖掘出来的规则不一定都是用户感兴趣的。

1 改进的关联规则挖掘方法

例1 现假设对某超市交易活动中涉及购买钢笔和圆珠笔的事务进行考察,一段时间内所形成的交易数据库中共包含 10000 条交易记录。其中购买钢笔的事务为

收稿日期:2005-06-15
基金项目:安徽省高等学校自然科学研究项目(2005KJ312ZC);安徽财经大学教研课题(ACJY200544)
作者简介:徐 勇(1978—),男,安徽泾县人,讲师,硕士,主要从事数据挖掘和数据库技术教学与研究;周森鑫,副教授,研究方向为数据库技术与网格计算。

9000 条，购买圆珠笔的事务为 2500 条，既购买了钢笔又购买了圆珠笔的事务为 2000 条。见表 1。

表 1 购买钢笔和圆珠笔的 2×2 相依表

	pen	$\overline{\text{pen}}$	
ballpen	2000	500	2500
$\overline{\text{ballpen}}$	7000	500	7500
	9000	1000	10000

设定最小支持度和可信度阈值分别为 support = 5%，confidence = 60%，针对此交易数据集执行传统的关联规则挖掘过程，可获得规则：

圆珠笔⇒钢笔[support = 20%，confidence = 80%] (1)

然而一般来说，圆珠笔和钢笔是负相关(negative correlated)的，也即增加其中一种商品的购买将会减少对另外一种商品购买的可能性。文献[1,4,5]等也提到过类似的例子并作了一定的分析。对原始交易数据进行分析发现有 90%的交易中包含了购买钢笔的行为，正是由于购买钢笔事件的概率很高——说明该购买行为在实际生活中普遍发生，它的发生受其它购买行为影响的程度很小——所以导致了规则(1)具有比较高的支持度和可信度，进而使得这个并不正确的规则被挖掘出来。

通过大量实验，发现基于支持度－可信度框架的关联规则挖掘过程应用到实际数据分析时，可能会得到支持度和可信度均满足用户要求的规则，但其实际利用价值却不如理论预期的那么高，甚至是错误的。

1.1 关联规则的形式化定义

定义 1[2]：设 $I = \{i_1, i_2, \cdots, i_m\}$ 是一个项的集合，称之为项集；设 TD 是一个交易数据库，T 表示其中的一个交易，且有 $T \subseteq I$；设 $X \subseteq I$，当且仅当 $X \subseteq T$ 时称事务 T 包含或支持项集 X。

定义 2[2]：关联规则是形如 $A \Rightarrow B$ 的蕴涵式，其中 $A \subset I, B \subset I$，且 $A \cap B = \emptyset$；交易数据库 TD 中的关联规则具有支持度 s 和可信度 c；支持度是指 TD 中同时包含 A 和 B 的事务的百分比，可信度指 TD 中包含 A 的事务同时也包含 B 的百分比。对于一个给定的交易事务集 TD，关联规则挖掘的任务是挖掘所有满足一定约束的强规则。

1.2 相关的工作

R. Agrawal 等在文献[2]中针对大规模顾客交易数据库中的购物篮数据提出关联规则挖掘时，指出关联规则是满足最小支持度和最小可信度阈值的蕴涵式 A⇒B；在文献[3,6]中介绍了基于支持度－可信度框架的挖掘算法。其后的研究很多都是致力于提高挖掘算法的效率。关联规则挖掘是从大规模数据库中提取客观存在的、隐含的、用户感兴趣的模式，因此挖掘出来的关联规则就必须具有一定的代表性、普遍性；反之对于偶然出现的模式，是没有必要提供给用户的。R. Agrawal 介绍的支持度约束是描述模式显著性的一种较为恰当的形式，利用支持度约束可以除去出现频率较小的项集，保留大于用户指定最小支持度阈值的项集，称其为频繁项集。并且基于支持度约束的频繁项集产生过程是向下封闭的(Download Closure)[1]，所以这也保证了应用支持度约束于实际算法的可行性。

强关联规则的兴趣度问题首先由 Piatetsky－Shapiro 在文献[7]中研究，文献[8～11]等对此作了进一步研究。其中文献[7]中提出的定义关联规则兴趣度的 3 个原则得到了大多数人的认同，文献[12]讨论了影响关联规则兴趣度的其它一些因素，文献[4]对多种兴趣度度量方法进行了比较分析，认为大部分情况下(medium support)各种度量方法的性能相似。

综上所述，已有的度量规则兴趣度的方法有很多优点，但是也存在着以下不足：

(1) 使用支持度－可信度框架的关联规则挖掘在许多场合下是有用的，但如在上文中所述，基于支持度－可信度框架的关联规则挖掘方法导出的规则并不都是有趣的，有些是多余的，有些甚至具有很强的误导性。

(2) 关联规则的相关度度量方法[1]、PS 方法[7]把规则 A⇒B、B⇒A 的兴趣度同等看待，显然这是不恰当的。

其它文献，如文献[4,12,13]等对相关的问题作了进一步的研究，如对影响关联规则兴趣度度量的因素进行了详细的分析，提出了一些改进的度量方法等。但是这些文献的研究成果侧重对该问题的理论分析，其算法实现难度较大。

1.3 关联规则兴趣度的 ARI 度量方法

为了克服上文中提到的一些不足，引入一个新的关联规则兴趣度度量方法，并将其引入到关联规则挖掘过程中。在交易数据库中有理由假设一个项集的出现比其不出现更令人感兴趣，由此对于规则 A⇒B 给出如下的兴趣度度量表达式：

$$\text{ARI} = \left(\frac{P(AB)}{P(A)P(B)}\right)^{\frac{P(B/A)}{P(B/\overline{A})}} \tag{2}$$

在公式(2)中，$P(AB)$ 的统计意义是事件 A, B 同时发生的可能性，同时它也表示在交易数据库中规则的支持度；$P(A)P(B)$ 表示事件 A, B 相互独立的可能性；$P(B/A)$ 表示规则的可信度，也即 B 事件相对于 A 事件的重要程度。容易理解，除式 $\left(\frac{P(AB)}{P(A)P(B)}\right)$ 的值是一个非负数，且有：

$$\text{当 ARI 的值为}\begin{cases}[0,1) & A,B \text{ 是负相关的}\\ 1 & A,B \text{ 是相互独立的}\\ (1,+\infty) & A,B \text{ 是正相关的}\end{cases}$$

$P(B/A)$、$P(B/\overline{A})$ 分别反映了一项集相对于另一项集的影响程度，通过它们的比值修正了由于规则前件、后件具有高支持度时带来的负面影响。

2 实验过程及其结果分析

笔者在一个实际超市交易数据集上将引入的兴趣度度量标准 ARI 的挖掘过程与 Apriori 算法进行了对比实验，具体实验情况如下：

数据集中交易总数为 298 条，交易商品的种类总数为

88 个,交易的平均长度(平均包含的商品种类数)为 6 个。实验用微机为联想台式机:P42.6G,512M 内存;Windows 2000 操作系统环境。实验结果如图 1、图 2 所示。

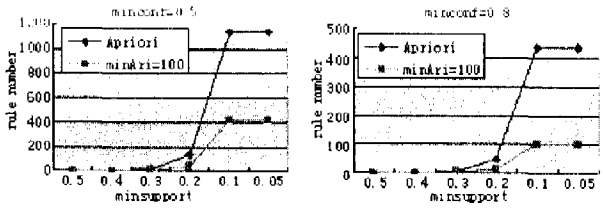

图 1　应用不同支持度的规则数

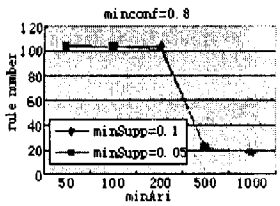

图 2　应用不同 ARI 的规则数

上述实验结果显示:

(1) 在不考虑可信度约束时,ARI 度量标准的剪枝效果较好,剪枝率(使用 ARI 度量标准进行规则挖掘时,去除的规则数占总规则数的比例)保持在 50%左右;

(2) 在支持度阈值较高时(>0.2)剪枝效率不明显;

(3) 当可信度阈值较高和支持度阈值保持较低水平时,剪枝效率较高,剪枝率保持在大于 60%;

(4) 当保持较高可信度(≥0.8)、较低支持度(≤0.1)时,设定较高的 ARI 阈值(≥300)比设定较低的阈值的剪枝效率更高;

(5) 在任何可信度阈值情况和保持支持度较低时(≤0.1),在以某一值为分界点的较低兴趣度阈值或较高兴趣度阈值范围内,各种兴趣度阈值情况下的剪枝效果差别不明显(该分界点根据不同的领域而变化)。

3　结论与下一步的工作

通过对实验结果的分析,发现对购物篮数据进行关联规则挖掘分析时,通过设计合理的数据结构与算法,在支持度-可信度框架的基础上,引入关联规则兴趣度度量方法 ARI,可以在保持时间性能不降低的前提下明显减少规则的数量,达到提高挖掘效率的目的。

下一步的工作准备在 ARI 度量基础上引入卡方检验,在频繁项集产生的过程中寻找更加简洁的闭合频繁相关项集,以减小产生规则的项集空间,提高挖掘过程的效率;结合领域知识寻找对可信度阈值、支持度阈值、ARI 阈值进行规范计算的策略与途径。

参考文献:

[1] Han jiawei, Kamber M. Data Mining - concepts and techniques[M]. San Francisco, CA: High Education Press, Morgan Kaufman Publishers, 2001.

[2] Agrawal R, Imielinski T, Swami A. Mining Association rules between sets of items in large databases[A]. In Proc 1993 ACM - SIGMOD Int Conf Management of Data(SIGMOD '93)[C]. Washington D.C.:[s.n.],1993.207-216.

[3] Agrawal R, Srikant R. Fast algorithms for mining association rules in large database[R]. Technical Report FJ9839, San Jose, CA: IBM Almaden Research Center, 1994.

[4] Tan P, Kumar V. Interestingness Measures for Association Patterns: A Perspective[R]. Technical Report # TR00-036, Department of Computer Science, University of Minnesota, 2000.

[5] 周　欣,沙朝锋,朱扬勇,等. 兴趣度-关联规则的又一个阈值[J]. 计算机研究与发展, 2000,37(5):627-633.

[6] Agrawal R, Srikant R. Fast algorithms for mining association rules[A]. In Proc. of 20th Int Conf Very Large Databases (VLDB'94)[C]. CA:[s.n.],1994.487-499.

[7] Piatetsky-Shapiro G. Discovery, Analysis, and Presentation of Strong Rules[A]. In Piatetsky-Shapiro G. Knowledge Discovery in Databases[C]. California: AAAI/MIT Press, 1991.229-248.

[8] Brin S, Motwani R, Silverstein C. Beyond market basekets: Generalizing association rules to correlations[A]. In: Proc of 1997 ACM SIGMOD Int'l Conf on management of Data[C]. Tucson, Arizona, UAS: ACM Press, 1997.265-276.

[9] Aggarwal C C, Yu P S. A new framework for itemset generation[A]. In Proc 1998 ACM Symp Principles of Database Systems(PODS'98)[C]. NY:[s.n.],1999.18-24.

[10] Ahmed K M, El-Makky N M, Taha Y. A note on Beyond market basket: Generalizing association rules to correlations [Z]. In: Fayyad U. ACM SIGMOD Explorations Newsletter, 2000,1(2):46-48.

[11] Chen M S, Han J, Yu P S. Data mining: An overview from a database perspective[J]. IEEE Trans Knowledge and Data Engineering, 1996,8:866-883.

[12] Freitas A A. On rule interestingness measures[J]. Knowledge-Based Systems, 1999,12(5-6):309-315.

[13] 杨建林,邓三鸿,苏新宁. 关联规则兴趣度的度量[J]. 情报学报, 2003,22(4):419-424.

(上接第 76 页)

参考文献:

[1] 李　敏,李仁发,杨大山,等. 基于虚拟原型技术的虚拟网络实验室[J]. 计算机工程与应用,2002(7):151-153.

[2] 徐雷鸣,庞　博,赵　耀. NS 与网络模拟[M]. 北京:人民邮电出版社,2003.56-125.

[3] 黄筱俊,郑善贤. 基于 NS 的移动网络仿真研究[J]. 微机发展,2004,14(5):25-27.

[4] 吴仕浩,林庆华,胥布工. 网络仿真器 NS-2 及其一个应用实例[J]. 计算机仿真,2004(7):96-98.

[5] Stevens W R. TCP/IP 详解卷 1:协议[M]. 北京:机械工业出版社,2000.61-81.