基于间接关联规则的数据挖掘算法研究

薄 宏,任玉杰,曹惠茹
(中山大学 南方学院,广东 广州 510970)

摘 要:简单数据集可以通过关联规则得到在数据间的相互关系;相当多的情况下,由于不能从关联规则得到隐藏在数据间的相互关系,需要按间接关联规则分析出数据项集在交易集合中出现的频度,挖掘隐藏在数据间的相互关系。文中通过使用概念分层和基于近邻的方法,探讨利用FP树产生的频繁项集,对候选关联检验其是否满足项对支持度条件,并利用这个频繁项集挖掘事务的间接关联,找到挖掘事务的间接关联的内在规律,构造出不依赖中介条件的间接关联挖掘算法。

关键词:间接关联;FP树;数据挖掘
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1673-629X(2012)11-0120-03

Research of Data Mining Algorithms Based on Indirect Association Rules

BO Hong, REN Yu-jie, CAO Hui-ru
(Nanfang College of Sun Yat-sen University, Guangzhou 510970, China)

Abstract: Simple data set can get its correlation among data by the association rules. While in many other cases, the correlation hidden in the data cannot be found in this way. So, indirect association rules are needed to analyze database frequentness in transaction set and reveal the correlation hidden in the data. The research uses the concept hierarchy and affinity methods to discuss the frequent item sets occurred in the FP-tree, and examine the candidate association test to verify whether it is eligible for the support condition. Then use the frequent item sets to find the indirect association and its inherent rules in order to build indirect association mining algorithm free of intermediary conditions.

Key words: indirect association; FP-tree; data mining

0 引 言

数据挖掘作为一门交叉学科,日益受到人们的重视。数据挖掘(Data Mining)是在大量数据中,发现数据之间的联系和规则,从而提取有用的信息。这种概念把数据挖掘的对象定义为大量数据,其中,数据指的是一个有关事实的集合,如移动客户的通话清单,是用来描述事务有关方面的信息和进一步发现知识的原材料。关联分析(association analysis)、关联规则的数据挖掘就是假设数据之间是有内在的必然联系的,通过挖掘算法以发现变量间的某种规律性,从而得到数据库中存在的重要的、可被发现的信息。

在关联规则挖掘的算法研究中,R. Agrawal(1994)提出布尔关联规则挖掘频繁项集的算法[1]。频繁项集的算法采用逐层迭代的方法,产生候选项集,并使用候选项集去递归频繁项集。Apriori算法在处理大型事务数据时,由于重复的扫描数据库以及复杂的候选项集生成策略,降低了效率[2]。

Han Jiawei(2000)提出不产生候选项集的基于FP-tree算法结构,在对源数据库扫描过程中,将数据信息对应到FP-tree结构里,避开了候选项集发生的数据交换的额外开销,从而将数据库频繁模式的挖掘问题转化成挖掘FP-tree的问题。R. Agrawal(2001)利用FP-tree生成频繁项目集的投影算法对候选项集进行改进[3],Mohammed J(2004)提出"挖掘非冗余关联规则"来发现大量数据之间的联系[4]。

国内学者陆楠(2003)给出FP-growth结构对挖掘关联规则的影响[5],易彤(2004)引入支持度函数用于关联规则维护[6],程苗(2007)通过处理数据挖掘中的异常数据来建立关联分析[7],陆觉民(2009)对具有代表性的与分析任务相关的数据进行样本抽取,将对数据库的操作转化为对在样本中关联的操作[8],谢志强(2009)提出敏感性关联规则[9],张笑达(2010)采用去除大量冗余的非频繁项集来重构关联规则[10],汪成

收稿日期:2012-02-23;**修回日期**:2012-05-27
作者简介:薄 宏(1960-),女,高级工程师,主要研究方向为计算机建模、数据库技术。

亮(2010)通过构建条件模式基,以组合方式挖掘频繁项集[11]。从数据的相关关系分析中,可找出数据项间相关的紧密程度。刘宏(2008)提出基于灰色关联的挖掘算法[12],从数据库中抽取极少的数据为统计样本,以灰色关联分析找出事务中项目间的关系,并且在给定条件下的数据挖掘中表现出较高的效率。采用灰色关联的前提是数据有高度的一致性,即不同时间周期的数据要有足够小的差异。

1　关联规则挖掘

在一般情况下,数据之间存在某种关联规则,对事务数据库定义关联规则如下:

设 $I=\{i_1,i_2,\cdots,i_m\}$ 是项的集合,其中 $i_k(k=1,2,\cdots,m)$ 定义为消费者购物篮中的物品。

数据 D 是事务数据库的集合;事务数据库是由一系列 T 事务组成,即事务 T 是项的集合,即 $T\subseteq I$。

设项集 X,若存在项集 $X\subseteq I, X\subseteq T$,事务 T 包含项集 X。设定 X 为事件前提,Y 为事件的结果。

关联规则是如下形式的逻辑蕴涵:$X\rightarrow Y$,其中 $X\subseteq I, Y\subseteq I$,且 $X\cap Y=\varnothing$。当规则的支持度大于事先设定的最小支持度,且置信度大于事先设定的最小置信度时,则存在必然的关联。

设 $S(X)$ 为支持度,则项集 X 的 $S(X)$ 定义:$S(X)=|\{X\subseteq I, T\in D\}|/|D|$,其中 $|\{X\subseteq I, T\in D\}|$ 表示数据库中包含有项集 X 的事务数,$|D|$ 表示数据库 D 中总共有的事务数。

由于数据之间的关系的模糊性,对于关联规则的置信度在期望可信度下限以内时,则表明了 X 的出现对 Y 的出现有了促进关系,即表明了 X 与 Y 之间存在某种程度的相关性。这种相关性通过支持度测量以确定并非偶然出现的关联,发现较高支持度的关联规则。

给定数据库 D 的关联规则 $X\rightarrow Y$ 的支持度,可以用包含 X 和 Y 的事务数占所有事务数的比值大小来确定,且当 X 与 Y 是不相交的项集时,支持度 $S(X\rightarrow Y)$ 的度量定义为:

$$S(X\rightarrow Y)=|\{T|X\cup Y\subseteq T, T\in D\}|/|D|$$

在数据库 D 中支持度较高的关联规则,即强规则 $X\rightarrow Y$ 所对应的项目集 $X\cup Y$ 的关联规则,当项目集 $X\cup Y$ 是频繁项目集时,计算 $X\rightarrow Y$ 的置信度的问题转化为求解 $X\cup Y$ 的支持度的问题。

关联规则 $X\rightarrow Y$ 在数据库 D 中的置信度是指包含 X 和 Y 的事务数与包含 X 事务数的比值,记为 $C(X\rightarrow Y)$,即

$$C(X\rightarrow Y)=|\{T|X\cup Y\subseteq T, T\in D\}|/|\{T|X\subseteq T, T\in D\}|$$

为便于分析,将关联规则挖掘划分为两个子问题:首先是找出所有的频繁项集,其目标是发现所有支持度满足给定的最小支持度阈值的项目集。其次,在频繁项目集的基础上,找到强关联规则;即从得到的频繁项目集中,得到高置信度的规则,通常应不小于用户给定的最小置信度,这些高置信度的规则称为强规则。

2　基于间接关联规则的挖掘

2.1　间接关联规则

相当多的情况下,由于不能从关联规则得到隐藏在事务数据库中的相互关系,采用一组用户数据项集或用户交易数据集合,按间接关联(indirect association)规则分析出数据项集在交易集合中出现的频度关系。Tan(2000)等给出挖掘序列与非序列的间接关联的思想。

假设商品 (a,b) 很少被顾客同时购买,如果 a 与 b 是不相关的商品,例如无糖面包和运动鞋,则对这个购物篮的支持度期望较低。如果 a 与 b 是相关的商品,例如牛奶和无糖面包,则对这个购物篮的支持度期望较高。在这里,运动鞋与无糖面包是不同的产品类别,通常预期支持度低是正常的现象。然而,观察消费者购买行为时,有一个值得注意的现象,一些需要减体重的人,购物出现了运动鞋与无糖面包的频繁事件。因此,预期在减体重的人和运动鞋,减体重的人和无糖面包,两者之间有着某种关联。

为了说明期望支持度,以两个不同属性类 P、Q,使用概念分层来挖掘非频繁模式(见图 1)。

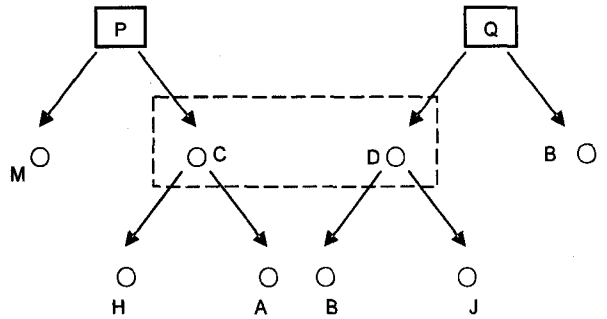

图 1　分层挖掘的非频繁模式

假定项集 $\{C,D\}$ 是频繁的,用 $S()$ 表示模式的实际支持度,$\varepsilon()$ 表示期望的支持度。则 (A,B) 的期望的支持度为:

$$\varepsilon(A,B)=s(C,D)\times\frac{s(A)}{s(C)}\times\frac{s(B)}{s(D)}$$

事实上,间接关联是建立在 (a,b) 具有相同的隐含中介基础上的关联。这种隐含中介可以称为中介集(mediator),设 Y 为中介集,则:

$y\in A$ 且 $y\in B$

确定是否存在间接关联,需要设定间接关联模式的支持度,对于研究目标所给定支持度记为 t_p,与 Y 中

介集的支持度记为 t_q，则有：

$s(\{a,b\}) < t_p$

中介支持度条件为：

$s(\{a\} \cup Y) \geqslant t_q$ 并且 $s(\{b\} \cup Y) \geqslant t_q$

设定中介支持度的目的是为确定 Y 为 a,b 的近邻。

2.2 基于 FP-growth 的频繁项集的挖掘算法

在包括 Apriori 算法在内的多种算法中，一般基于强关联规则的算法。在间接关联规则情况下，为生成所有频集，需要使用递推算法，这样会产生大量的候选集。更为难以处理的是，在大量的候选集中掩盖了弱相关的数据之间的关系。

Han Jiawei(2000)提出了不依赖候选项集的频繁项集挖掘算法 FP-growth，由于减少了数据交换和频繁匹配的开销，使得间接关联规则变得简捷。

基于 FP-growth 的频繁项集算法的基本思想是采取分治策略，FP 增长方法发现频繁项集而不产生候选，即将一个问题转换成若干个较小的子问题，得到某个后缀结尾的全部频繁项集。

假设商品 (a,b) 两者之间有着某种关联，则 a、b 之间应共享某些共同项。进一步，Y 为 a、b 的近邻的情况下将含有更多的共同项；更好的情况是所有的事务都具有共同项，FP 树只包含一条结点路径，这时 FP 树将具有最大的效率。

已知频繁项集 F_k，间接关联的算法是对每项 k 项集找到候选的间接关联 (a,b,Y)，对候选关联检验其是否满足项对支持度条件。P. N Tan 给出中介依赖条件下的挖掘算法。由于依赖条件的兴趣因子等，通常情况下并非构成必要的约束条件，而且候选间接关联是通过合并频繁项集得到的，去除中介依赖条件后的挖掘算法更为简捷，降低了搜索开销。间接关联规则的挖掘算法为：

for $k = 2$ to k_{max} do

$C_k = \{(a,b,Y) \mid \{a\} \cup Y \in F_k, \{b\} \cup Y \in F_k, a \neq b\}$

for $(a,b,Y) \in C_k$ do

if $s(\{a,b\}) < t_p$ then

$I_k = I_k \cup \{(a,b,Y)\}$

end if

end for

end for

Result = $\cup I_k$

结合 FP 增长算法，通过支持度计数去除中介依赖条件，使事务数据集得到压缩，从而高效地产生频繁项集，并利用这个频繁项集实现挖掘事务的间接关联。

关联规则从一组给定的数据项以及交易集合中，分析出数据项集在交易集合中出现的频度关系。由于不能从关联规则得到隐藏在数据间的相互关系，需要从一组给定的数据项以及交易集合中得到。

3 结束语

在相当多的情况下，间接关联规则可以得到隐藏在数据间的相互关系，可以通过分析数据项集在事务数据集合中出现的频度关系来确定。基于 FP 树的间接关联规则算法的核心是由 FP 树提供频繁项集的数据库压缩到一棵频繁模式树上，使用事务数据集的压缩，有效地产生频繁项集，并利用这个频繁项集挖掘事务的间接关联。事实上，间接关联是建立在具有相同的隐含中介基础上的关联。这种隐含中介可以称为中介集(mediator)，通过中介集，可以确定间接关联。由于候选间接关联是通过合并频繁项集得到的，文中最后给出不依赖中介条件下改进的间接关联挖掘算法。

参考文献：

[1] Agrawal R, Srikant R. Fast algorithms for mining association rules in large database[C]//Proceedings of the 20th International Conference on Very Large Data Bases. San Francisco: Morgan Kaufmanm, 1994: 478-499.

[2] 王 威，陈 梅. 基于位集合的 Apriori 算法的改进[J]. 计算机技术与发展，2011，21(12)：70-72.

[3] Agarwal R, Aggarwal C C, Prasad V V V. A tree projection algorithm for generation of frequent itemsets[J]. Journal of Parallel and Distributed Computing, 2001, 61(3): 352-368.

[4] Zaki M J. Mining Non-redundant Association Rules[J]. Data Mining and Knowledge Discovery, 2004, 9(3): 39-46.

[5] 陆 楠，王 喆，周春光. 基于 FP-tree 频集模式的 FP-Growth 算法对关联规则挖掘的影响[J]. 吉林大学学报，2003，41(2)：180-185.

[6] 易 彤，徐宝文，吴方君. 一种基于 FP 树的挖掘关联规则的增量更新算法[J]. 计算机学报，2004，27(5)：703-706.

[7] 程 苗. 关联分析在数据挖掘中的应用[J]. 激光杂志，2007(3)：65-65.

[8] 陆觉民，马国栋，郑 宇. 基于数据挖掘技术的图书馆流通数据的关联分析[J]. 现代情报，2009(9)：108-110.

[9] 谢志强，朱孟杰，杨 静. 基于 FP-Tree 的敏感性关联规则隐藏的研究[J]. 哈尔滨工程大学学报，2009，30(10)：1134-1141.

[10] 张笑达，徐立臻. 一种改进的基于矩阵的频繁项集挖掘算法[J]. 计算机技术与发展，2010，20(4)：93-96.

[11] 汪成亮，罗昌银. 一种基于组合方式改进的频繁项集挖掘算法[J]. 计算机系统应用，2010，19(1)：68-71.

[12] 刘 宏，吴 江，耿国华，等. 基于灰色关联分析的高感兴趣度数据挖掘算法研究[J]. 计算机工程与设计，2008(8)：4242-4244.