关联规则挖掘研究①

Study on Associated Rules Algorithm

包 剑 （阜新 辽宁工程技术大学计算机系 123000）

摘要：关联规则挖掘是数据挖掘的知识模式中比较重要的一项任务，它的目的是发现数据集中所有的频繁模式。根据关联规则定义及属性，可发现关联规则。利用 Apriori 算法实现了关联规则的挖掘，关联规则可以产生清晰有用的结果；可以广泛应用于各个领域，既可以检验行业内长期形成的知识模式，也能够发现隐藏的新规律，是完成数据挖掘任务的一个重要手段。

关键词：关联规则 数据挖掘 Apriori

1 引言

关联规则是由 Agrawal 于 1993 年提出，它是数据中一种简单但很实用的规则。发现关联规则的算法属于无监督学习的方法。关联规则可用于发现交易数据库中不同商品之间的联系，以找出顾客购买行为模式，如购买了某一商品对购买其他商品的影响。现实例子很多，例如超级市场利用前端收款机收集存储了大量的售货数据，这些数据是一条条的购买事务记录，每条记录存储了事务处理时间，顾客购买的物品、物品的数量及金额等。这些数据中常常隐含形式如下的关联规则：在购买铁锤的顾客当中，有 70% 的人同时购买了铁钉。这些关联规则很有价值，商场管理人员可以根据这些关联规则更好地规划商场，如把铁锤和铁钉这样的商品摆放在一起，能够促进销售。

2 关联规则

2.1 关联规则的定义

关联规则就是描述在一个事务中物品之间同时出现的规律的知识模式。更确切的说，关联规则通过量化的数字描述物品甲的出现对物品乙的出现有多大的影响。

设 $I=\{i_1,i_2,\ldots,i_m\}$ 是一组物品集，其中的元素称为项。D 是一组事务集（称之为事务数据库）。D 中的每个事务 T 是一组物品，并且满足 $T\subseteq I$。设 X 是一个 I 中项的集合，如果 $X\subseteq T$，称事务 T 支持物品集 X。

关联规则是如下形式的一种蕴含：$X\Rightarrow Y$，其中 $X\subseteq I, Y\subseteq I$，且 $X\cap Y=\varnothing$。规则 $X\Rightarrow Y$ 在事务数据库 D 中的支撑度是事务集中包含 X 和 Y 的事务数与所有事务数之比，记为 support($X\Rightarrow Y$)，即 support($X\Rightarrow Y$) = $|\{T: X\cup Y\subseteq T, T\in D\}| / |D|$[1]。

规则 $X\Rightarrow Y$ 在事务数据库 D 中的可信度是事务集中包含 X 和 Y 的事务数与包含 X 的事务数之比，记为 confidence($X\Rightarrow Y$)，即

confidence($X\Rightarrow Y$) = $|\{T: X\cup Y\subseteq T, T\in D\}| / |T: X\subseteq T, T\in D\}|$

如果不考虑关联规则的支持度和可信度，那么在事务数据库中存在无穷多的关联规则。事实上人们一般只对满足一定的支持度和可信度的关联规则感兴趣。在文献中，一般称满足一定要求的（如较大的支持度和可信度）的规则为强规则。因此，为了发现出有意义的关联规则，需要给定两个阈值：最小支持度和最小可信度。前者即用户规定的关联规则必须满足的最小支持度，它表示了一组物品集在统计意义上的需满足的最低程度；后者即用户规定的关联规则必须满足的最小可信度，它反应了关联规则的最低可靠度。对于给定的一个事务集 D，挖掘关联规则问题就是产生支

① 基金项目：辽宁省教育厅高等学校科学研究项目（202182054）资助

©《计算机系统应用》编辑部 http://www.c-s-a.org.cn

持度和可信度分别大于用户给定的最小支持度和最小可信度的关联规则。

在实际情况下，一种更有用的关联规则是泛化关联规则。因为物品概念间存在一种层次关系，如夹克衫、滑雪衫属于外套类，外套、衬衣又属于衣服类。有了层次关系后，可以帮助发现一些更多的有意义的规则。例如"买外套→买鞋子"（此处，外套和鞋子是较高层次上的物品或概念，因而该规则是一种泛化的关联规则）。由于商店或超市中有成千上万种物品，平均来讲，每种物品（如滑雪衫）的支持度很低，因此有时难以发现有用规则：但如果考虑到较高层次的物品（如外套），则其支持度就较高，从而可能发现有用的规则[2]。

另外，关联规则发现的思路还可以用于序列模式发现。用户在购买物品时，除了具有上述关联规律，还有时间上或序列上的规律，因为，很多时候顾客会这次买这些东西，下次买同上次有关的一些东西，接着又买有关的某些东西。

2.2 关联规则的属性

一般用四个参数来描述一个关联规则的属性：

（1）可信度（Confidence）。设 W 中支持物品集 A 的事务中，有 c% 的事务同时也支持物品集 B，c% 称为关联规则 A⇒B 的可信度。简单地说，可信度就是指在出现了物品集 A 的事务 T 中，物品集 B 也同时出现的概率有多大。如上面所举的铁锤和铁钉的例子，该关联规则的可信度就回答了这样一个问题：如果一个顾客购买了铁锤，那么他也购买铁钉的可能性有多大呢？在上述例子中，购买铁锤的顾客中有 70% 的人购买了铁钉，所以可信度是 70% 。

（2）支持度（Support）。设 W 中有 s% 的事务同时支持物品集 A 和 B，s% 称为关联规则 A⇒B 的支持度。支持度描述了 A 和 B 这两个物品集的并集 C 在所有的事务中出现的概率有多大。如果某天共有 1000 个顾客到商场购买物品，其中有 100 个顾客同时购买了铁锤和铁钉，那么上述的关联规则的支持度就是 10% 。

（3）期望可信度（Expected confidence）。设 W 中有 e% 的事务支持物品集 B，e% 称为关联规则 A⇒B 的期望可信度度。期望可信度描述了在没有任何条件影响时，物品集 B 在所有事务中出现的概率有多大。如果某天共有 1000 个顾客到商场购买物品，其中有 200 个顾客购买了铁钉，则上述的关联规则的期望可信度就是 20% 。

（4）作用度（Lift）。作用度是可信度与期望可信度的比值。作用度描述物品集 A 的出现对物品集 B 的出现有多大的影响。因为物品集 B 在所有事务中出现的概率是期望可信度：而物品集 B 在有物品集 A 出现的事务中出现的概率是可信度，通过可信度对期望可信度的比值反映了在加入"物品集 A 出现"的这个条件后，物品集 B 的出现概率发生了多大的变化。在上例中作用度就是 70% /20% =3.5。

可信度是对关联规则的准确度的衡量，支持度是对关联规则重要性的衡量。支持度说明了这条规则在所有事务中有多大的代表性，显然支持度越大，关联规则越重要。有些关联规则可信度虽然很高，但支持度却很低，说明该关联规则实用的机会很小，因此也不重要。

期望可信度描述了在没有物品集 A 的作用下，物品集 B 本身的支持度：作用度描述了物品集 A 对物品集 B 的影响力的大小。作用度越大，说明物品集 B 受物品集 A 的影响越大。一般情况，有用的关联规则的作用度都应该大于 1，只有关联规则的可信度大于期望可信度，才说明 A 的出现对 B 的出现有促进作用，也说明了它们之间某种程度的相关性，如果作用度不大于 1，则此关联规则也就没有意义了。

2.3 关联规则的挖掘

在关联规则的四个属性中，支持度和可信度能够比较直接形容关联规则的性质。从关联规则定义可以看出，任意给出事务中的两个物品集，它们之间都存在关联规则，只不过属性值有所不同。如果不考虑关联规则的支持度和可信度，那么在事务数据库中可以发现无穷多的关联规则。事实上，人们一般只对满足一定的支持度和可信度的关联规则感兴趣。因此，为了发现有意义的关联规则，需要给定两个阈值：最小支持度和最小可信度，前者规定了关联规则必须满足的最小支持度：后者规定了关联规则必须满足的最小可信度。一般称满足一定要求的（如较大的支持度和可信度）的规则为强规则（Strong rules）[3]。

在关联规则的挖掘中要注意以下几点：首先充分理解数据。其次目标明确。第三，做好数据准备工作。能否做好数据准备又取决于前两点。数据准备将直接

©《计算机系统应用》编辑部 http://www.c-s-a.org.cn

影响到问题的复杂度及目标的实现。然后选取恰当的最小支持度和最小可信度。这依赖于用户对目标的估计，如果取值过小，那么会发现大量无用的规则，不但影响执行效率、浪费系统资源，而且可能把目标埋没；如果取值过大，则又有可能找不到规则，与知识失之交臂。最后，很好地理解关联规则。数据挖掘工具能够发现满足条件的关联规则，但它不能判定关联规则的实际意义。对关联规则的理解需要熟悉业务背景，丰富的业务经验对数据有足够的理解。在发现的关联规则中，可能有两个主观上认为没有多大关系的物品，它们的关联规则支持度和可信度却很高，需要根据业务知识、经验，从各个角度判断这是一个偶然现象或有其内在的合理性；反之，可能有主观上认为关系密切的物品，结果却显示它们之间相关性不强。只有很好的理解关联规则，才能去其糟粕，取其精华，充分发挥关联规则的价值。

关联规则发现要经过以下步骤：首先连接数据，作数据准备；第二，给定最小支持度和最小可信度，利用数据挖掘工具提供的算法发现关联规则；最后，可视化显示、理解、评估关联规则。

3 算法研究

(1) 找出所有具有超出最小支持度的支持度的项集(itemsets)，由 Apriori 算法实现。

这里讨论 Apriori 算法，因为根据此算法得到的大项集，在序列模式阶段是有实际用处的。

L_1 = {large 1 – itemsets};
for (k = 2; $L_{k-1} \neq \varphi$; k + +) do begin
  C_k = apriori – gen (L_{k-1}); // New candidates
  forall transactions t ∈ D do begin
    C_t = subset (C_k, t); // Candidates contained in t
    forall candidates c ∈ C_t do
      c. count + + ;
  end
  L_k = { c ∈ C_k | c. count ≥ minsup}
end
Answer = $\cup_k L_k$;

apriori – gen 函数以 L_{k-1}(所有大(k – 1) – 项集)作为输入参数，返回所有大 k – 项集的集合 L_k，以以下两步实现：

第一步，联合

insert into C_k
select p. $item_1$, p. $item_2$, …, p. $item_{k-1}$, q. $item_{k-1}$
from L_{k-1}p, L_{k-1}q
where p. $item_1$ = q. $item_1$, …, p. $item_{k-2}$ = q. $item_{k-2}$, p. $item_{k-1}$ < q. $item_{k-1}$;

第二步，剪枝(pruning)，如果存在 c 的(k – 1) – 子序列不包含于 L_{k-1} 之中，则删除所有项集 c ∈ C_k。

forall itemsets c ∈ C_k do
forall (k – 1) – subsets s of c do
if (s　　L_{k-1}) then
delete c from C_k;

(2) 利用大项集(Itemsets)产生所需的规则(rules)。算法的思想在于：如果说 ABCD 和 AB 是大项集，就可以通过计算可信度，也就是 conf = support(ABCD) / support(AB)，并通过 conf ≥ miniconf 来确定规则 AB→CD 是否确立(该规则由于 ABCD 是大项集故肯定具有最小支持度)[4]。

4 实例

若了解某网站被用户访问的情况，经常出现的数据组可表示为在某次站点访问的页面集。零售商可通过关联规则发现将相关的商品摆在一起进行组合销售，参阅表 1、2。

表 1 事实表

TID	Itemset
1	Milk
1	Butter
2	Milk
2	Honey
2	Butter
3	Milk
3	Bread
3	Butter
4	Milk
4	Bread
4	Honey

(下转第 62 页)

©《计算机系统应用》编辑部 http://www.c-s-a.org.cn

(上接第 58 页)

表 2　关联数据

support	Itemset
4	{ Milk }
3	{ Milk },{ Butter },{ Milk ,Butter }
2	{ Milk },{ Butter },{ Milk ,Butter } { Honey },{ Bread },{ Honey, Bread },{ Honey, Milk } { Honey,Butter },{ Bread,Milk },{ Butter,Butter }

一旦关联数据被推导出，即可用于生成关联规则。通过选定关联数据中的某一类为预测目标，给其它类赋值作为预测规则的条件。关联规则可以产生清晰有用的结果；支持间接数据挖掘；可以处理变长的数据；计算的消耗量可以预见。

5　结语

本文讨论了对于数据挖掘中关联规则的发现，以及一些基本的概念和算法，并进行了程序实现。关联规则可以广泛应用于各个领域，既可以检验行业内长期形成的知识模式，也能够发现隐藏的新规律。有效地发现、理解、运用关联规则，是完成数据挖掘任务的一个重要手段。

参考文献

1 Agrawal R, Imielinski T, Wami A S. Mining Association Rules Between Sets of Items in Large Databases. In : Proc. of the ACM SIGM OD Conference on Management of Data, Washington, D. C. ,1993 -05:207 - 216.

2 A grawal R, Srikant R. Fast algorithm for mining association rules. In: Proceedings of the 1994 International Conference on Very Large Data Bases. Santiago, Chile, 1994. 487 - 49.

3 王珊等，数据仓库技术与联机分析处理，北京 科学技术出版社，1998。

4 陈京民等编著，数据仓库与数据挖掘技术，北京 电子工业出版社，2002。

©《计算机系统应用》编辑部 http://www.c-s-a.org.cn