一种改进的Apriori算法在交通信息化中的应用

吴 昊，李军国
(吉林大学珠海学院 计算机科学与技术系，广东 珠海 519041)

摘 要：基于关联规则理论，在传统的单维单层布尔型Apriori算法的基础上提出一种改进的多维多数据类型Apriori算法，将算法用于分析复杂的交通事故数据库。理论分析和实验数据表明，算法是有效可行的，实验结论达到了交通管理部门的预期要求，可以用于辅助相关部门作出道路交通改进工作的决策。

关键词：关联规则；Apriori算法；数据挖掘
中图分类号：TP182　　**文献标识码：**A

An improved Apriori algorithm apply to traffic management

WU Hao，LI Jun Guo
(Computer Science and Technology program，Zhuhai College， Jilin University,Zhuhai 519041，China)

Abstract: Based on the association rules and the classical single-dimensional and single-layer boolean Apriori algorithm,the text puts forward an improved multi-dimensional and mutli-type Apriori algorithm to analyse the complicated database which records the data of traffic accidents. The theory and the experiment data indicates that the algorithm is effective and feasible.On the whole,the conclusions are coincident with the anticipated requests,and it can be used to assist the correlative department to make decision on improving the road traffic.

Key words: association rules；Apriori algorithm；data mining

在当今的网络信息社会，为及时处理与日剧增的数据信息，解决知识获取这一瓶颈问题，研究人员于20世纪80年代末提出了知识发现(KDD)的概念。知识发现是指从数据集中识别出有效的、新颖的、潜在有用的，以及最终可理解的模式的非平凡的过程，数据挖掘(DM)是知识发现过程的一个重要步骤，在KDD和DM的诸多方法中，关联规则数据挖掘算法尤为引人注目。

关联规则(Association Rules)的概念首先由AGRAWAL R等于1993年提出[1]，用来反映一个事物与其他事物之间的相互依赖性或相互关联性[2]。如果两个或者多个事物之间存在一定的关联关系，那么，其中一个事物就能够通过其他事物预测到。关联规则在决策支持系统、专家系统和智能信息系统等各个方面都能起到重要的作用。

Apriori算法是挖掘产生布尔关联规则所需频繁项集的基本算法，是根据有关频繁项集特性的先验知识(Prior Knowledge)而命名的。该算法利用了一个层次顺序搜索的循环方法来完成频繁项集的挖掘工作。这一循环方法就是利用k-项集来产生(k+1)-项集[3]。本文正是在此基础上进行了探讨，力求寻找一种更为实用的改进的Apriori算法。

本文从关联规则理论出发，提出一种改进的多维多数据类型Apriori算法。算法是通过改进候选集C_k的产生方式，即函数gen_candidate()中候选集的产生。理论分析和实验数据表明，文中提出的改进的Apriori算法是有效、可行的。

1 关联规则理论

1.1 关联规则基本概念

设I=$\{i_1,i_2,\cdots,i_n\}$是二进制文字的集合，其中的元素称为项(Item)。

规则$A\Rightarrow B$在交易数据库D中的支持度(Support)是交易集中包含A和B的交易数与所有交易数之比，记为support($A\Rightarrow B$)，即$P(A\cup B)$[2]

support($A\Rightarrow B$)=$P(A\cup B)$=|{T:$A\cup B\subseteq T$,$T\in D$}| / |D|

规则$A\Rightarrow B$在交易集中的置信度(Confidence)是指包含A和B的交易数与包含A的交易数之比，记为confidence($A\Rightarrow B$)，即$P(B|A)$[2]

$$confidence(A\Rightarrow B)=P(B|A)=support(A\cup B)/support(A)$$
$$=|\{T:A\cup B\subseteq T,T\in D\}|/|T:A\cup B\subseteq T,T\in D\}|$$

其中：$support(A\cup B)$为包含项集$A\cup B$的交易记录数目，$support(A)$为包含项集A的交易记录数目。

规则的支持度和置信度是两个规则兴趣度度量，它们分别反映发现规则的实用性和确定性。这两个阈值均在0～100%之间，而不是0～1之间。

最小支持度min_sup(Minimum Support Count)指用户规定的关联规则必须满足的最小支持度，表示一组物品集在统计意义上需满足的最低程度；最小置信度min_con(Minimum Confidence Count)指用户规定的关联规则必须满足的最小置信度，反映了关联规则的最低可靠度。一般将同时满足最小支持度阈值和最小置信度阈值的规则称为强规则。如果项集满足最小支持度，则称它为频繁项集(Frequent Itemset)，简称频集。频繁k-项集的集合通常记作L_k。

1.2 Apriori算法

经典的Apriori算法利用一个层次顺序搜索的循环方法来完成频繁项集的挖掘工作。这一循环方法就是利用k-项集来产生(k+1)-项集。挖掘或识别所有频繁项集是Apriori算法的核心，占整个计算量的大部分。

Apriori算法的性质：频繁项集中所有非空子集都必须也是频繁项集。

根据这一性质，如果项集I不满足最小支持度阈值min_sup，则I不是频繁的，即$P(I)$<min_sup。如果项A添加到I，则结果项集$I\cup A$不可能比I更频繁出现。因此，$I\cup A$也不是频繁的，即$P(I\cup A)$<min_sup[4]。Apriori算法利用这一重要性质来压缩搜索空间，生成较小的候选项目集，提高频繁项集逐层产生的效率。

用L_{k-1}找L_k的过程由连接和剪枝两步组成[5]：

(1)连接步骤

为找L_k，通过L_{k-1}与自己连接产生候选k-项集的集合C_k。设l_1和l_2是L_{k-1}中的项集。记$l_i[j]$表示l_i的第j项，例如：$l_1[k-2]$表示l的倒数第3项。假定项集中的项按字典次序排序，执行连接$L_{k-1}\Theta L_{k-1}$，其中L_{k-1}的元素是可连接的，如果它们前(k-2)个项相同。L_{k-1}的元素l_1和l_2是可连接的，如果$(l_1[1]=l_2[1])\wedge(l_1[2]=l_2[2])\wedge\ldots\wedge(l_1[k-2]=l_2[k-2])\wedge(l_1[k-1]<l_2[k-1])$，条件$(l_1[k-1]<l_2[k-1])$是简单地保证不产生重复。连接$l_1$和$l_2$产生的结果项集是$l_1[1]\ l_1[2]\cdots l_1[k-2]\ l_2[k-1]$。

(2)剪枝步骤

C_k是L_k的超集，即它的成员可以是也可以不是频繁的，但所有的频繁k-项集都包含在C_k中。扫描数据库，确定C_k中每个候选的计数，从而确定L_k，即根据定义，计数值不小于最小支持度计数的所有候选是频繁的，从而属于L_k。然而，C_k可能很大，这样所涉及的计算量就很大。如果一个候选k-项集的(k-1)-子集不在L_{k-1}中，则该候选也不可能是频繁的，从而可以由C_k中删除。可以利用哈希表来保存所有频繁项集以便能够快速完成这一子集测试操作。

2 改进的多维数据类型Apriori算法

2.1 算法思想

描述一个属性集的维间关联规则的分析算法主要依赖于迭代性质，事实上，候选集C_k是由两个(k-1)-频繁集作连接，然后进行剪枝，删除那些包含不属于L_{k-1}的子集生成的，这个剪枝过程提高了计算效率。然而，要浪费时间来对L_{k-1}作连接，然后还要检查其结果是否存在不合法的子集。如果数据表很稀疏并且k-频繁集不是很大，那么这个连接到剪枝过程是很有效的。但是如果属性集中的数据稠密，k-频繁集将变得巨大并且剪枝过程不能使C_k变小很多。这样，实现迭代性质将需要非常复杂的数据结构，那么仅仅考虑实现多维多数据类型Apriori算法将失去它的优点。下面是多维多数据类型Apriori算法的改进算法。

算法是通过改进候选集C_k的产生方式，即函数gen_candidate()中候选集的产生。在产生新的候选集时，在k-1项频繁集中，任取频繁子集l_1和l_2，判断l_1中每一个项i所在的维m是否在l_2中所有项所在维的集合M中存在，如果不存在，则将粒度i加入l_2中，产生k-项集C，然后判断C能否加入候选集C_k中。

然后在函数gen_infrquent，从候选集C_k中生成频繁项集L_k，在生成L_k时，首先通过连接操作$L_{k-1}\Theta L_{k-1}$来生成可能的k-itemsets，再利用Apriori性质对连接生成的可能的k-itemsets进行删除，只有满足此性质的将其加入到候选k-itemsets的集合C_k中。对于每一个候选k-itemsets $T\in C_k$，统计在k-维数据表中与之对应的计数值(支持度)，根据数据定义，存储在属性数据中的count值是从原始属性集数据中的一个聚集值，即为该属性数据所代表的项目集的频度，将它与最小支持度min_sup比较，如果大于，则将其加入到L_k中。

2.2 改进的多维多数据类型Apriori算法流程图

图1为改进的多维多数据类型Apriori算法数据流程。

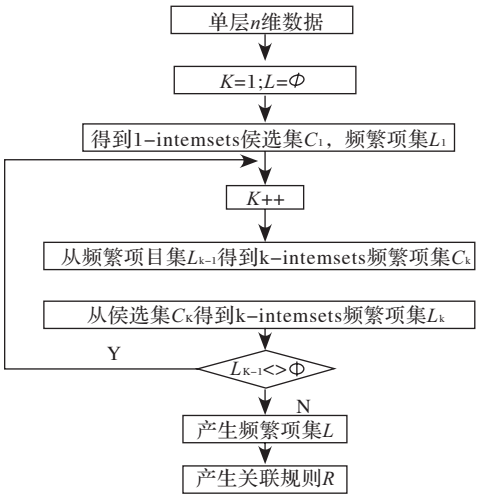

图1 改进的多维多数据类型Apriori算法数据流程

3 实验结果分析

使用标准C++语言实现了“改进的多维多数据类型Apriori算法”，将算法在某市市区1998年12月21日至2004年7月20日的道路交通事故数据进行测试。算法的硬件运行环境：DELL PC机Optiplex 270系列，奔腾Ⅳ 2.60GHz CPU，512M DDR内存。

3.1 数据集的预处理

由于关系表属性的数据类型比较丰富，可归结为数值型(区间型、离散型)与类别型(含布尔型)两大类。类别型属性的取值可映射为一组连续整数；离散数值型属性的取值可映射为一组保持属性值次序的连续整数；区间数值型属性的各个区间可映射为保持区间排列次序的连续整数。

实验通过概括属性值形成概念，一个概念是同一个属性的若干个互不交叠的取值区间的并集，这些区间可以是互不相邻的。这个定义对数值型与类别型都适用，对于类别型它实际上是类别值的并集。这样，每个属性都可以有概念层次，并且不同概念的内涵可以重叠。同时充分利用数据挖掘的基本方法，数据聚焦选择和当前分析相关的数据，包括维。如果某个属性包含大量不同值，同时在该属性上有概化操作符(即汇总函数)，则运用该操作符进行属性概化。

由于算法的输出结果就是相关规则。为了更明确表示决策规则，这里采用 “条件集合⇒结果”的方式。并且给出(sup,con)，即每条规则的支持度sup和置信度con，来说明规则的重要性和有效性。

3.2 算法复杂度分析

改进的多维多数据类型Apriori算法的第一步是找出所有符合最小支持度的频繁项集，在空间上的消耗主要考虑候选集的产生方式，即函数gen_candidate()中候选集C_k的产生。第二步是利用前面产生的频繁项集产生期望的关联规则，同时利用关联规则的重要性质“任意数据项集的支持度总是小于等于其子集的支持度”，减少了算法搜索数据项集的范围，提高了算法的效率。

3.3 算法的结果精确性

以分析2003年12月至2004年7月某市市区事故原因为例，首先得到的是各种事故原因发生的比例(事故数据个数为2 401)，例如不按规定让行(27.66%)、未保持安全距离(12.16%)、超速行驶(8.79%)、违章拐弯(7.58%)、其他机动车原因(5.46%)、逆向行驶(4.91%)、违章变更车道(4.33%)等，其他事故原因比例较小，可视为“噪音”，而且一般情况下也不为用户所感兴趣，在此不一一列出。

假设用户想了解导致“不按规定让行”这一结果，设置最小支持度阈值为5%，最小置信度阈值为40%。产生以下规则：

例如：条件：天气因素、道路因素

规则：

(1) 天气：晴；地形：平原；道路横断面：混合式；照明条件：白天；道路线形：平直；交通控制方式：无控制⇒不按规定让行(9.27%，36.31%)。

(2) 天气：晴；路面类型：沥青；道路横断面：混合式；照明条件：白天；交通控制方式：无控制⇒不按规定让行(11.84%，36.34%)。

(3) 天气：晴；路面情况：平坦；道路横断面：混合式；照明条件：白天；道路线形：平直；交通控制方式：无控制⇒不按规定让行(8.67%，36.33%)。

(4) 天气：晴；地形：平原；道路横断面：混合式；照明条件：白天；道路线形：平直；交通控制方式：无控制⇒不按规定让行(9.27%，36.31%)。

对1998年12月至2004年7月某市市区交通事故情况进行分析：机动车驾驶人违法驾车行为是导致交通事故发生的主要原因。平直道路事故频繁。晴天事故占绝大比例。从月统计周期分布来看，6~9月为事故多发时段，9月事故致人死亡较为突出。从24小时事故分布情况看，10~11、14~16时段事故多发，18~21时死亡率较高；其次为8~10、11~12、16~17时段。可见，中午、傍晚和下午时分是交通事故的多发时段。

传统的单维单层布尔型Apriori算法难于适应现实中海量、高维的数据库规模，因而实用高效的多维多数据类型Apriori算法是该理论研究的主要课题之一。本文基于关联规则理论，提出一种改进的多维多数据类型Apriori算法。理论分析和实验数据表明，文中提出的多维多数据类型Apriori算法是有效、可行的。实验结论达到交通管理部门的预期要求，可以用于辅助相关部门作出日后道路交通改进工作的决策。

当然，本文提出的改进算法还可以通过改进建立数据库中的数据模型，使模型包含更加详细的内容，从而促使算法对更加复杂的数据进行分析。进一步提高算法的效率是下一步需要深入探讨的问题，寻求真正行之有效的、可以作用于更复杂数据的算法，仍需不懈努力。

参考文献

[1] LUO Ke He，CAI Wang. An improved algorithm for mining association rules based on apriori algorithm[M]. Computer &Digital Engineering,2006.

[2] LIU Zheng Jiang,WU Zhao Lin.Data mining to human factors based on ship collision accident survey reports[J].Navigation of China.2004,59(2).

[3] OU Yu Ming，ZHANG Shi Chao,XU Zhang Yan,et al. Improved apriori algorithm for efficiency [J].Computer Engineering and Design，2004，25(5).

[4] MENG Gang. Research and improvement on Apriori algorithm of association rules [J].Journal of Changchun Institute of Technology(Natural Sciences Edition)，2007(04).

[5] 景永霞,王治和,杜跃. 一种新的Apriori改进算法[J].长春理工大学学报(自然科学版)，2007(02).

(收稿日期：2009-01-09)