基于 Apriori 分类事务库关联规则算法①

A Rule-Calculate of Classification Database Based on Apriori

曹月芹[1] 林 枫[2] 陈国浪[1] (1.温州职业技术学院 计算机系 浙江 温州 325035;
2.温州电力局基建处 浙江 温州 325027)

摘 要: Apriori 算法是通过定义的事务库来生成频繁项集，确定各数据之间的关联规则。在实际应用中生成事务库时常常会出现同一项目中重复类型的事务库，而同一项目的事务之间的支持度为零。因此，事务库的定义方法是直接影响生成关联规则的生成速度与效率，针对这一问题，本文提出并实现了一种基于 Apriori 分类事务库的关联规则算法。该方法改变了传统 Apriori 算法中所有事务统一定义的方法，对不同项目的事务进行分类定义，通过这种的实现，不但减少了计算机的大量运算，而且提高了关联规则的生成速度。

关键词: 关联规则 事务库 数据挖掘 算法 频繁项集 分类

数据挖掘现在越来越为更多的人们所关注，被认为是未来最有发展前景和广阔市场潜力的新兴学科之一。随着信息技术和数据库技术的不断发展，各行各业的人们掌握了大量的数据，在竞争日益激烈的现今社会里，如何迅速有效地获得隐藏在数据之后的有用的知识信息，成为众多企业决策者和管理者的当务之急。

1 引言

数据挖掘(Data Mining)[1-3]就是从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中，提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程。关联规则的挖掘是其中一个数据挖掘的重要方法，我们在对学校的实际事务数据库用传统的 Apriori 算法进行关联规则的挖掘时发现，由于每一项目集的每一事务都有唯一的事务代码，在进行频繁项目集生成时存在着大量的同一项目事务之间的支持度为零，在生成关联规则时，这样的项集通常是无效的，而这些无效的项集的也给计算机带来了一系列无效的计算。

针对以上问题，本文提出了基于 Apriori 的分类事务库关联规则算法。该算法不仅简化事务库的生成，而且大大减少了计算机的运算量，另外在挖掘的效率方面，也比传统的 Apriori 算法有所提高。

在下文中，笔者运用学校的“学生信息管理系统”中的学生信息,举例分析了传统与改进算法的生成事务库。然后笔者通过不同算法的运行速率的分析，从而验证了改进算法的优势所在。最后笔者对全文的研究工作进行了归纳小结，并提出了未来的研究方向。

2 基于Apriori分类事务库关联规则算法

2.1 算法的提出

基于 Apriori 分类事务库关联规则算法是对传统关联规则 Apriori 算法的改进。为了准确描述关联规则挖掘问题，需要给出关联规则挖掘问题的正式定义，需要用事务数据库来定义关联规则。事务库的定义正是生成关联规则的基础，由于事务库是存在于项目中，每个项目中有多个不同的事务，在传统的算法中，同一项目产生不同的多个事务库，而且在同一项目中多个事务产生的频繁集的支持度是零，在这种情况下，大大降低的关联规则生成的速率，同时也给计算机增加了大量不必要的计算量。

鉴于以上问题，笔者在借鉴相关文献的基础上，

① 基金项目:浙江省温州职业技术学院科研项目(WZY200806)
收稿时间:2008-09-19

© 中国科学院软件研究所 http://www.c-s-a.org.cn

提出了基于 Apriori 分类事务库关联规则挖掘算法，有效地弥补了上述缺陷，同时能高效挖掘出其中的关联规则。

2.2 基本概念

为了准确描述关联规则挖掘，需要给出关联规则挖掘问题的正式定义，下面用事务数据库来定义关联规则[4,5]。

定义 1. 记 D 为交易 T 的集合，D={t1,t2,…,tn}，这里交易 T 是项的集合，可以表述为:T={il, i2, ...ip)，并且 T∈D。T 中的元素 ij(J=1, 2, ...p)称为项。对应每一个交易有唯一的标识，如交易号，记作 TID。

定义 2. 设 I={i1,i2,...,im}是数据集中所有项的集合。I 中的任何子集称为项目集(itemset),若|X|=k，则称集合 X 为 k-项集。

设 tk、和 X 分别为 D 中的事务和项目集，如果 X⊆tk，称事务 tk 包含项目集 X。

据集 D 中包含项目集 X 的事务数称为项目集 X 的支持数，记为 δ x,。项目集 X 的支持度，记作:support(X)。

Support(X)=(δ x)/|D|*100%

其中|D|是数据集 D 中的事务数。若 support(X)不小于用户指定的最小支持度(记作:minsupport)，则称 X 为频繁项目集，否则称 X 为非频繁项目集。

定理 1(向下封闭定理).设 X, Y 是数据集 D 中的项目集

(1)若 X⊆Y，则 support (X)≥support (Y)

(2)若 X⊆Y，如果 X 是非频繁项目集，则 Y 也是非频繁项目集

(3)若 X⊆Y，如果 Y 是频繁项目集，则 X 也是频繁项目集。

定义 3. 一个关联规则是形如 X=>Y 的蕴涵式，这里 X,Y 都是项目集，且 X⊆I，Y⊆I，并且 X∩Y=ϕ。X, Y 分别称为关联规则 X=>Y 的前件和后件。设 I={i1, i2,…, im}是项目属性集。记 D 为交易事务(transaction)T 的集合，这里交易 T 是项目属性的集合，并且 T⊆I 。对应每一个交易有唯一的标识，如交易号,记作 TID。设 X 是一个 I 中项的集合,如果 X⊆T，那么称交易 T 包含 X。

2.3 算法描述

Apriori 算法是挖掘布尔关联规则频繁项集的有效算法。Apriori 使用一种称作逐层搜索的迭代方法，k-项集用于探索(k+1)-项集。

首先，找出满足用户设定的最小支持度阈值的频繁 1-项集的集合，记为 L1。L1 用于寻找频繁 2-项集的集合 L2，而 L2 用于寻找 L3，如此循环下去，直到不能找到频繁 k-项集。寻找每个 Lk 需要一次数据库扫描。寻找所有频繁项集是关联规则挖掘算法的核心部分，计算工作量最大。

其次，根据最小置信度 A 值从频繁项集中构造出有效的关联规则。

2.3.1 生成频繁项集[6,7]

Apriori 算法的核心就是寻找项集中的频繁 K-项集。产生频繁 K-项集的过程可以分为连接和剪枝两步[4,7]。

(1) 连接步:为寻找 L_k，通过 L_{k-1}，与自己连接产生候选 K-项集的集合。该候选项集的集合记作 C_k。设 l_1 和 l_2 是 l_{k-l} 中的项集。记号 $l_i[j]$表示 l_i 的第 J 项(例如，$l_1[k-2]$表示 l_1 的倒数第 3 项)。为方便考虑，假定事务或项集中的项按字典次序排序。执行连接 L_{k-1} ∞ L_{k-1}，其中，L_{k-1} 的元素是可连接的，如果它们的前(k-2)个项相同

(2) 剪枝步: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项是频繁的，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。为压缩 C_k，可以用 Apriori 性质。因此，如果一个候选 k-项集的(k-1)-项集不在 L_k.中，则该候选项也不可能是频繁的，从而可以由 C_k 中删除。这种子集测试可以使用所有频繁项集的散列快速完成。

2.3.2 生成关联规则

根据定理 2 (向下封闭定理)，频繁项集的所有子集均是频繁项集，关联规则可通过下列方法从频繁项集中产生[5]:

(1)对于每个频繁项集 l，产生 l 的所有非空子集。

(2) 对于 l 的每个非空子集 s，如果(support(l)support(s))>minconfidence(最小置supports)信度)，则输出规则 “s=>(l-s)”。

由于规则是在频繁项集中产生的，因此，所有的规则都自动满足最小支持度(minsupport)。

© 中国科学院软件研究所 http://www.c-s-a.org.cn

3　算法的应用

实验环境如下：

硬件配置：PC，CPU 为 Intel 的 P4　2.4G，256M 内存。

软件配置：windows 操作系统，FoxPro 数据库系统。

3.1 数据准备

在本算法中，笔者利用“学生信息管理系统”中学生的相关数据进行实际的应用。Apriori 算法可以把关联规则挖掘问题分解为两个子问题：①找出所有频繁项集，这些项集出现的频繁性至少和预定义的最小支持计数一样。②由频繁项集产生强关联规则，这些规则必须满足最小支持度和最小置信度。

在本案例中，采用布尔型关联规则，采集相关数据后进行清洗，后得到如下图 1 所示的数据集，此数据集中包括数值数据，布尔型数据多种类型。

性别	专业方向	收入	专业成绩	工作适应性
男	网络	1500	70	适应
男	网络	2300	84	适应
男	网络	2400	75	适应
女	网络	1900	90	适应
男	网络	2300	84	适应
男	网络	3200	75	适应
男	网络	2600	80	适应

图 1　采集学生数据

由于本案例采用的是布尔型关联规则，通过数据转换等相关处理得到如下图 2 所示的离散数据，这样就可以对采集的数据实行相关的挖掘。

性别	专业方向	收入	专业成绩	工作适应
男	网络	最低	一般	适应
男	网络	一般	一般	适应
男	网络	一般	一般	适应
女	网络	较低	较好	适应
男	网络	一般	一般	适应
女	网络	一般	一般	不适应
男	网络	较高	较差	不适应
女	网络	一般	较好	适应

图 2　离散处理后的学生数据

3.2 算法实现

3.2.1 Apriori 算法事务库生成

从上面整理好的数据中提取相关属性，用前面整理好的数据作为整个数据挖掘模块提供了数据挖掘对象。由于 Apriori 算法适用于事务库的数据挖掘，所以需要将关系表转换为相应的事务库，传统事务库是将所的有项目统一编排，改进分类事务库是将各项目分类生成，即同一项目为一类事务，如“专业方向”这个项目为一类编排，这样减少了同一项目之间生成项集的不必要运算，其事务库采用代码如表 1 所示，将关系表中的一条记录视为一个事务，记录中的字段值采用代码表 1 所示。

表 1　两种不同算法的代码表

项目	值	代码	
		Apriori 算法	分类事务库算
性别	男	R01	A01
	女	R02	A02
专业方向	网络	R03	B01
	软件设	R04	B02
	多媒体	R05	B03
收入	最高	R06	C01
	较高	R07	C02
	一般	R08	C03
	较低	R09	C04
	最低	R10	C05
专业成绩	较好	R11	D01
	一般	R12	D02
	较差	R13	D03
工作适应性	适应	R14	E01
	不适应	R15	E02

3.2.2 生成频繁项集

Apriori 算法的核心就是寻找项集中的频繁 K-项集。表 2 显示了两种不同算法的项集集合。

对上述生成的项集设定相关最小支持度与最小置信度生成一系列频繁项集。在算法实现的过程中，只要在执行语句中加上同类项目的判断语句，则对于那些同一项目的事务就不会产生相关的项集了，其规则生成结果如下。

3.2.3 规则生成

利用前面整理好的数据与生成的项集，用 Apriori 算法进行分析，在最小支持度设定为 5%，最小置信度为 15%的情况下，生成了一些频繁项集，通过整理，挖掘出很多条规则，这里从我校的实际情况出发，以

© 中国科学院软件研究所 http://www.c-s-a.org.cn

表 2　算法项集的比较

项集	Apriori 算法		分类事务库算法	
1 项集	(R01).(R02).(R03).(R04)....(R15)	15 个项集	(A01).(A02).(B01).(B02)...(E02)	15 个项集
2 项集	(R01,R02).(R01.R03)...(R01.R15) (R02,R03).(R02,R04)...(R02,R15)	14+13+...+1=105	(A01,B01).(A01,B02).(A01,B03) (A20,B01)....(A02,B03)	13*2+10*3+5*5+3*2=87
3 项集	...	...	...	

一位老师多年的教学经验主要选取了以下一些比较实用的规则进行分类显示如表 3 所示。

表 3　关联规则显示

	Apriori 算法	分类事务库算法
规则	R01 => R03 support: 19.3% confidence: 33.4% R01 => R04 support: 22.7 % confidence: 42.8% R01 ∧ R04 => R08 support:7.5 % confidence: 33.3 % R01 ∧ R08=> R04 support:7.5 % confidence: 40.7 %	A01 => B01 support: 19.3% confidence: 33.4% A01 => B02 support: 22.7 % confidence: 42.8% A01 ∧ B02 => C03 support:7.5 % confidence: 33.3 % A01 ∧ C03=> B02 support:7.5 % confidence: 40.7 % ....

上述规则转换成属性与值的关系相当于：性别=“男”=>专业方向=“网络”支持度为 19.3%，表示在数据库表中有 19.3 的男同学选网络方向的专业，置信度为 33.4%表示在男同学中有 33.4 的人选择网络方向。专业方向是学生在进行学习一年后，根据自己的兴趣自主选择专业，从某种程度上看，专业方向体现了学生对各专业的学习兴趣。通过上述规则可以看出，男同学对软件设计比较感兴趣，对多媒体是不感兴趣。女同学对多媒体专业最感兴趣。在这里女同学对软件设计的选择没有出现，说明女同学最不感兴趣的是软件设计。

3.3 算法比较

从以上算法的实现过程中可以看出：在 Apriori 分类事务库算法中，由于按照项目的种类进行分类，在生成事务库的过程中大提高了生成效率，在本例子中，2 项集的传统算法生成了 105 项集，而改进算法中生成了 87 项集，本案例中对计算机来说减少了 18% 的计算机，而对于一些实际的项目中，每个项目的事务是远远超过本案例的种类，这种情况下，会更加提高生成项集的效率。同样道理，对于后面的一些项集的生成也大大提高了很多。由于项集的生成少了，在生成频繁项集时，工作量也就大大减少，再从生成规则的结果比较来看，改进和方法对生成的结果没有任何影响。所以，在挖掘的效率方面，改进的分类事务库算法保持了原有 Apriori 算法规则结果，又提高了生成规则的效率。本改进算法是针对每个项目中有多个种类的情况而定，如果每个项目都只有一个种类的事务，则本算法与传统的算法就没有多大区别了。

4　结束语

4.1 小结

本文的主要内容是设计并实现了一种基于 Apriori 分类事务库的关联规则算法。本文的研究工作归纳如下：首先本文针对 Apriori 算法中的生成事务库的种类繁多，同一项目的事务之间出现支持度为零的情况下提出并实现了一种基于 Apriori 分类事务库关联规则算法。该算法针对每个项目设定一类，按照每一类进行事务码的生成，因此有效地弥补了上述原有的 Apriori 算法中的缺陷。其次，用传统算法与改进算法进行实际的应用对比。最后，两种算法进行了相关性的比较。

4.2 未来研究方向

在今后的研究中，我们的工作可以基于现有的工作从以下进行拓展：一是在对分类结果的研究中，发现产生的分类规则还是有比较大的误差，究其原因，

(下转第 143 页)

© 中国科学院软件研究所 http://www.c-s-a.org.cn

(上接第 56 页)

笔者认为在数据集中的属性字段选择上，还有许多影响学生成绩的因素没有考虑到，学生的信息调查的内容还不够全面，所使用的数据可能还不是最佳数据集，这一点有待将来进一步研究。二是在其它应用方面，高校教学管理中的诸多领域还有待进一步探讨。比如学生心理分析、教学质量评估等，这些方面的应用都是有待研究的新课题。

参考文献

1 陈文伟.数据仓库与数据挖掘教程.北京:清华大学出版社, 2006.

2 黄琳,秦化淑,郑应平,郑大钟.复杂控制系统理论:构想与前景.自动化学报, 1993,19(1):129－137.

3 王珊,等.数据仓库基础.北京:电子工业出版社, 1998.

4 Han JW, Kambe M.范明,孟小峰等译.数据挖掘概念与技术.北京:机械工业出版社, 2001.

5 刘南艳,杨君锐.多最小支持度下的关联规则及其挖掘方法研究.西安科技大学学报, 2005,(4):79－82.

6 吴安阳,赵卫东.基于多最小支持度的空间关联规则发现.计算机应用, 2005,(9):212－215.

7 肖基毅,邹腊梅,刘丰.频繁项集挖掘算法研究.情报杂志, 2005,(11):4－5.

© 中国科学院软件研究所 http://www.c-s-a.org.cn