关联规则在汽车销售中的应用

刘斐
（同济大学软件学院 上海 200000）

摘 要：该文主要介绍了关联规则挖掘的FP-tree算法。并基于对FP-tree算法的研究，在Microsoft 2010中用C#语言实现FP-tree算法，并将应用程序应用到某汽车销售企业的汽车销售数据进行关联规则挖掘。根据程序所得到的结果，由lift值判断，证明了所得规则的有效性。
关键词：数据挖掘 关联规则 FP-tree算法 汽车销售 lift
中图分类号：TP311.13 文献标识码：A 文章编号：1674-098X（2013）01（a）-0006-03

Application of Association Rules toAutomobile Sales

Liu Fei
(School of Software Engineering of Tongji Univ.Shanghai)

Abstract：This paper describes the mining association rules, and the FP-tree algorithm.Based on FP-tree algorithm, the algorithm is implemented to the application by using tool Microsoft 2010 with C # language.And then Use thisapplication to mine an automobile sales database in order to obtain useful association rules.According to the results obtained by the program, it proves that the resulting rules are valid by using lift.
Key Words：data mining association rules FP-tree algorithm automobile sales lift

数据挖掘（Data Mining，DM）就是从大量的、不完全的、有噪声的、模糊的数据中，提取隐含在其中的、人们事先不知道的但又是潜在的可理解、可接受、可应用的有用信息和知识的过程，并最终利用其来进行重要的商业决策[1-2]。该文重点研究关联规则中的FP-Tree算法。根据对这种算法的研究，并运用其对某公司的汽车销售数据进行挖掘，运用关联规则分析销售数据，找出影响汽车销量的因素，为汽车销售提供决策支持。

1 关联规则算法

1.1 关联规则概述

关联规则是美国IBM Almaden Research Center的RakeshAgrawal等人于1993年首先提出来的知识发现（KDD：Knowledge Discovery in Databases）研究的一个重要课题[3]。由于关联规则挖掘形式简洁、思路清楚、易于理解，并可以有效的捕捉数据间的重要关系，因此从大型数据库中挖掘关联规则的问题已经成为近年来数据挖掘研究领域的一个热点。

1.2 关联规则算法

在目前的许多算法中，以1994年RakeshAgrawal等人提出了的Apriori算法最有影响力[4]，其他大多数算法也是以Apriori算法为核心的。Apriori算法是使用一种称作逐层搜索的迭代方法。首先，产生1-频繁项集，记作L_1。然后用L_1找出2-频繁项集L_2，直到不能找到更多的频繁项集为止。在k次循环中，过程先产生k-候选项集的集合C_k。然后通过扫描数据库生成支持度，并测试产生k-频繁项集L_k。找每一个L_k就需要扫描一次事务数据库。Apriori算法虽然简单明了，但是却存在难以克服的性能瓶颈。Apriori算法在执行的过程中需要很大的I/O负载，并且可能产生庞大的候选集。

针对Apriori算法的性能瓶颈问题-需要产生大量候选项集和需要重复地扫描数据库，2000年Jiawei Han等人提出了基于FP-tree生成频繁项集的FP-growth算法。该算法只进行2次数据库扫描且它不使用侯选集，直接压缩数据库成一个频繁模式树，最后通过这棵树生成关联规则。研究表明它比Apriori算法大约快一个数量级[5]。

FP-growth算法是一种不产生候选模式而采用频繁模式增长的方法挖掘频繁模式的算法。算法只需要扫描2次数据库：第一次扫描数据库，得到1维频繁项集；第二次扫描数据库，利用1维频繁项集过滤数据库中的非频繁项，同时生成FP-tree。由于FP-TREE蕴涵了所有的频繁项集，其后的频繁项集的挖掘只需要在FP-TREE上进行。FP-TREE挖掘由两个阶段组成：第一阶段建立FP-tree，即将数据库中的事务构造成一棵FP-tree；第二阶段为挖掘FP-tree，即针对FP-tree挖掘频繁模式和关联规则。由于FP-growth算法的优点，该文即基于FP-growth算法进行实现。

下面给出FP-growth算法：

输入：FP-tree，α；

输出：所有频繁模式集；

(1) begin

(2) FP-growth (tree，α)

(3) if (tree has single Path P) then

(4) forall$\beta \in P$ //对路径P中的结点的任一组合记为β

(5) 生成$\alpha \cup \beta$频繁项集，使其支持度等于β中所有节点的最小支持度

(6) else

(7) foralla_i //对Tree头上的每个节点记为a_i

(8) begin

(9) $\beta = a_i \cup \alpha$

(10) β.sup port=a_i .sup port;

(11) end

(12) 构造β的条件模式基和β的条件FP树$Tree_\beta$

(13) if $Tree_\beta \neq \varnothing$ then

(14) FP-growth ($Tree_\beta$，β)

(15) end

由FP-growth算法得到频繁项集之后，就可以提取其关联规则了。从已知频繁项集产生关联规则为两步：

(1) 对于每个频繁项集l，产生l的所有非空子集；

(2) 对于每个l的非空子集s，若$\frac{\text{sup port}(l)}{\text{sup port}(s)} \geq \min_conf$，则产生关联规则“$s \Rightarrow (l-s)$”，其中$\min_conf$是最小可信度阈值。

1.3 关联规则的有效性

在关联规则挖掘中，lift[6]是用来衡量关联规则（$R: X \Rightarrow Y$

）或目标模型是否有效的量。lift表示的是关联规则当中Y在X条件下的效应对与Y在整体平均水平下的高低。如果lift大于1，那么目标模型的效应就比一般平均水平高，即关联规则Y在X条件下表现的更好；lift小于1，则表明没有平均水平高，即关联规则Y在X条件下没有一般情况下好；lift等于1，则表示Y的表现与X不相关。lift的定义为：

$$lift(X \Rightarrow Y)=\frac{P(Y \mid X)}{P(Y)}=\frac{confidence(X \Rightarrow Y)}{support(Y)}$$

例1　表1是关于咖啡与茶的列链表，表中分别表示喝茶和喝咖啡的人数。

表1　咖啡与茶的列链表

	喝咖啡	不喝咖啡	
喝茶	50	20	70
不喝茶	20	10	30
	80	20	100

对于关联规则（喝茶⇒喝咖啡），由表中数据可得confidence（喝茶⇒喝咖啡）=50/70=0.714，即说明了在喝茶的人当中有71.4%的人和咖啡，这是一个很高的比率，如果最小置信度是70%的话，这个关联规则就是强关联规则。但是考虑到{喝咖啡}的支持度=80/100=0.8，说明人群中有80%的人喝咖啡。这个比率要比喝茶的人中喝咖啡的比率大，即说明了喝茶并不真的促进喝咖啡，而是相反。这个例子说明了，关联规则的置信度并不能完全反映关联规则的有效性，它需要更进一步判断才可以得出正确的结论。

我们应用lift来判断例1中关联规则的有效性，根据lift的计算公式：

$$lift(\text{喝茶} \Rightarrow \text{喝咖啡})=\frac{confidence(\text{喝茶} \Rightarrow \text{喝咖啡})}{support(\text{喝咖啡})}=\frac{0.714}{0.8}=0.893<1$$

可以知道，喝咖啡与喝茶负相关，即喝茶并不对喝咖啡有积极影响。

2 实验

2.1 确定挖掘对象

该文的挖掘对象是某公司的汽车销售数据，挖掘的目的是找出隐藏在汽车销售数据中顾客的性别、住址和购买车型中所蕴含的不为人知的知识和信息。

2.2 数据准备和预处理

首先，进行数据选择。对于该文，笔者需要从某公司的销售数据的数据库中提取出所要用到的销售数据，这些销售数据中必须包含本挖掘问题所关心的属性，即是购买者的性别、住址以及购买车型这三个必要属性。表2就是笔者在数据选择过程后的到的一部分数据。其中第一列sex是性别属性，包括男、女和单位。第二列brand是购买车型，有朗逸382、途观232和新POLOC22等。第三列address是顾客的地址信息。第五列y和第六列x分别是顾客地址的经纬度坐标。ID列是销售编号。

然后，进行数据的预处理。在该文中，经过数据选择后，需要对某些数据进行预处理，例如sex列，在数据表中，有很多交易这个属性为空值。所以需要在数据表中去掉sex属性为空的交易，负责会影响到sex列属性的支持度计算。第二列brand也需要清洗，因为在数据项中发现有同一车型但名称有差异的车型。如“新POLO　C22”和“新POLOC22”，它们虽然是同一车型，但记录的名称中却相差一个空格。所以，需要将相同车型的名称进行不一致清洗。同时需要将各品牌车的名称统一进行清洗，将细分车型改变成品牌名称。同样需要将第三列进行类似的清洗，将详细地址改变成每个区县的名称。由于前三列已经满足了挖掘所需的必要数据特征，为了提高挖掘效率，减少特征维数，第四列至第七列应予以丢弃，如表3所示。这样数据的预处理就已完成。

表2　数据选择后得到的部分数据表

sex	brand	address	region	y	x	ID
男	朗逸382	河西区大沽南路	高德	39.08405	117.247536	1
女	朗逸682	河东区大桥道	高德	39.109637	117.249706	2
女	途观232	静海县大邱庄镇	高德	38.833352	117.063213	3
男	新POLO B22	宝坻区大口屯镇	高德	39.585519	117.236515	4
男	朗逸382	河北区调纬路	高德	39.166588	117.207428	5
男	途观322	河西区绍兴道	高德	39.10727	117.212879	6
男	志俊BT2	武清区东蒲洼街	高德	39.422319	117.011098	7
女	新POLO C22	和平区山西路	高德	39.122396	117.19615	8
女	朗逸602	南开区迎风道	高德	39.084931	117.146804	9
男	POLO 842	河东区成林道	高德	39.128989	117.259822	10
男	朗逸382	河西区大沽南路	高德	39.08405	117.247536	11
女	朗逸682	河东区大桥道	高德	39.109637	117.249706	12
单位	志俊BT2	河东区卫国道	高德	39.142383	117.24348	25

表3　数据预处理后得到的部分数据表

sex	brand	address
男	朗逸	河西区
女	朗逸	河东区
女	途观	静海县
男	新POLO	宝坻区
男	朗逸	河北区
男	途观	河西区
男	志俊	武清区
女	新POLO	和平区
女	朗逸	南开区
男	POLO	河东区
男	朗逸	河西区
女	朗逸	河东区
单位	志俊	河东区

2.3 用FP-tree算法挖掘

根据FP-tree算法，该文使用Microsoft 2010 C#语言进行编程。所编得的应用程序界面如图1。本程序可以选择数据源文件，数据源文件须是文该文件，如*.dat或*.txt文件。

文件中的每行为一个交易数据，交易数据中的各属性用空格隔开。

在完成数据文件选择后，同样可以设置产生频繁项集的最小支持度和最小置信度。对于本次实验对象，该文选用的最小支持度为10%，最小置信度为60%。

运行本程序，得到最小支持度为10%，最小置信度为60%的两条关联规则为：朗逸⇒男，其对应置信度为67.73%；新POLO⇒女，其对应置信度为67.24%。运行结果如图1所示。第一个规则说明了买朗逸的车主中，有67.73%的是男车主，即选择买朗逸的男性居多。第二个规则说明了买新POLO的车主中，有67.24%的是女车主，即选择买新POLO的女性居多。由图1的运行结果知，（朗逸⇒男）的lift值为1.2052，（新POLO⇒女）的lift值为2.2493，均大于1，说明有效。可以根据其来作为汽车销售中相关决策的依据。

图1 FP-tree算法程序运行结果

3 结语

该文通过对FP-tree算法的实现，并将其应用到汽车销售数据中，得出了有效的关联规则。说明了关联规则可以有效的挖掘客户类型与购买车型之间的关系，为汽车销售的决策提供了有效的依据。目前对于关联规则挖掘的研究主要集中在如何提高发现频繁项集的效率，但对如何提高挖掘规则的有效性和可用性的研究则较少，所以，在海量的数据挖掘中很容易出现冗余项和无效规则。因此，在今后的研究当中，本人将会更关注如何提高关联规则挖掘的有效性和可用性方面的研究，以给出更准确有用的信息提供给决策者，达到科学决策的目的。

参考文献

[1] 朱明.数据挖掘 [M].2版.合肥：中国科学技术大学出版社，2008.
[2] Simoudis, Evangelos.IEEE Expert：Intelligent Systems and Their Applications.Reality Check for Data Mining, 1996:26-33.
[3] Agrawal, R.Imieli ski, T.Swami, A."Mining association rules between sets of items in large databases".Proceedings of the 1993 ACM SIGMOD international conference on Management of data - SIGMOD '1993:207-216.
[4] Agrawal, Rakesh；and Srikant, Ramakrishnan；Fast algorithms for mining association rules in large databases, in Bocca, Jorge B.；Jarke, Matthias；andZaniolo, Carlo；editors, Proceedings of the 20th International Conference on Very Large Data Bases (VLDB), Santiago, Chile, 1994:487-499.
[5] Jiawei Han, Jian Pei, YiwenYin.Mining Frequent Patterns without Candidate Generation. In：Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data.Dallas, 2000：l-12.
[6] Coppock, David S.(2002-06-21)."Data Modeling and Management：Why Lift?".2012-12-19.