文章编号: 1000-5862(2012)01-0095-04

基于分辨矩阵的含负属性项关联规则挖掘

王培吉[1], 章树玲[1], 赵玉琳[2]

(1.内蒙古科技大学数理与生物工程学院, 内蒙古 包头 014010; 2. 内蒙古第一机械集团公司二分公司, 内蒙古 包头 014032)

摘要: Apriori 算法存在候选集、频繁集产生效率低, 丢失有趣强关联规则等问题, 提出一种基于分辨矩阵可以采掘含负属性项强关联规则的改进算法, 最后给出一个实际例子实现该算法.

关键词: 数据挖掘; 分辨矩阵; 兴趣度; 关联规则
中图分类号: TP 301.6　　文献标志码: A

0 引言

Apriori 算法是由 R . Agrawal 和 R. Srikant[1-2]提出的最有影响的挖掘关联规则频繁项集的算法, Apriori 算法使用逐层搜索的迭代方法[3], k-项集用于寻找(k+1)-项集, 找每个 L_K 需先产生候选项集, 再通过扫描数据库来计算候选项集的支持计数, 消除非频繁项集, 当数据库较大时, 会产生庞大的候选项集, 且因多次扫描数据库将导致不可低估的输入输出开销[4].

另外, 基于 Apriori 算法进行关联规则挖掘, 可能生成无用关联规则, 如关联规则"买台式电脑 ⇒ 买笔记本"的支持度 37. 5%、置信度 75%分别大于最小支持度、最小置信度, 这样可得规则: "买台式电脑 ⇒ 买笔记本", 但同时得到: "88%的人肯定会买笔记本", 所以生成的此规则是无用关联规则, 反而含负属性项关联规则: "买笔记本 ⇒ 不买台式电脑"(其支持度和置信度分别为 50%、57%)更为合理、有用, 而传统的算法对挖掘含负属性项的关联规则是无能为力的.

本文提出一种改进的关联规则挖掘方法—基于 0-1 矩阵的含负属性项的关联规则挖掘模式, 使用这种方法只对数据库扫描一次, 无需候选集, 即可得到频繁集, 同时可过滤无趣关联规则, 产生含负属性项的有趣关联规则, 使获得的关联规则更有效、合理.

1 基本概念

1.1 事务数据库中的项集及关联规则

设项的集合 $I=\{i_1, i_2,\cdots, I_m\}$, 其子集称为项集. 项集中所包含的元素(项)的个数称为项集长度, 项集长度等于 K 的项集叫做 K-项集. 所有含有某个项集的事务数, 叫做该项集的支持计数.

设事务数据库 $D=\{T_1, T_2,\cdots, T_n\}$, 其中每个事务 $T_i \subset I$, (i=1, 2,⋯, n)每个事务有唯一标识符 TID, A 是项集, 事务 T_i 包含项集 A 当且仅当 $A \subset T_i$. 当 $A \subset I$, $B \subset I$ 并且 $A \cap B=\varnothing$ 时, 蕴涵式 $A \Rightarrow B$ 叫做事务数据库 D 中的关联规则[3].

1.2 支持度、置信度及强关联规则

给定事务数据库 D 中的关联规则 $A \Rightarrow B$, D 中事务同时包含 A、B 的百分比 S 称为关联规则 $A \Rightarrow B$ 在事务数据库 D 中的支持度(support); 包含项集 A 的事务中同时包含项集 B 的百分比 C 称为关联规则 $A \Rightarrow B$ 在事务数据库 D 中的置信度(confidence).

当关联规则 $A \Rightarrow B$ 在事务数据库 D 中的支持度、置信度分别大于等于各自的阈值时, 认为关联规则 $A \Rightarrow B$ 是有趣关联规则, 此两值叫做关联规则 $A \Rightarrow B$ 在事务数据库 D 中成立的最小支持度和最小置信度.

设关联规则 $A \Rightarrow B$ 在事务数据库 D 中成立的最小支持度和最小置信度分别为 $sup_{\min}$和 $conf_{\min,}$ 事务数据库 D 中事务总数为$|D|$.

当项集的支持计数大于或等于 $sup_{\min} \times |D|$时,

收稿日期: 2011-09-14
基金项目: 国家自然科学基金(81060238)资助项目.
作者简介: 王培吉(1968-), 男, 内蒙古包头人, 副教授, 硕士, 主要从事数据库、数据挖掘方面的研究.

此项集称事务数据库 D 中的频繁项集, 当项集的支持计数可能大于或等于 $sup_{\min}\times|D|$时, 此项集称事务数据库 D 中的侯选项集. 同时满足 $sup_{\min}$ 和 $conf_{\min}$ 的关联规则称为强关联规则.

如表 1 所示, $sup_{\min}$=25%, $conf_{\min}$=50%, 则 support(买台式电脑 ⇒ 买打印机)=2/8=25%, confidence(买台式电脑 ⇒ 买打印机)=2/4=50%, 满足 $sup_{\min}$ 与 $conf_{\min}$, 这样"买台式电脑 ⇒ 买打印机"是强关联规则.

表 1 事务数据库

TID	项 集
T_1	台式电脑、打印机、
T_2	台式电脑、打印机、笔记本
T_3	台式电脑、笔记本
T_4	台式电脑、笔记本
T_5	笔记本
T_6	笔记本
T_7	笔记本
T_8	笔记本

1.3 Apriori 算法

Apriori 算法在 1994 年由 R.Agrawal 和 R.Srikant 提出, 它使用逐层搜索的迭代方法,首先找出频繁 1-项集 L_1 , L_1 用于找频繁 2-项集 L_2 ,而 L_2 用于找 L_3 , 如此下去, 直到不能找到频繁 K-项集, 找每个 L_K 需要对数据库进行扫描. 它通过 2 个步骤来完成.

(1) 连接步, l_1、$l_2\in L_{K-1}$(), $l_i[j]$表示 l_i 的第 j 项, 如果 l_1、l_2 的前面的(k–2)项均一样, 则称 L_{K-1} 中的项 l_1、l_2 可以连接. l_1、l_2 连接生成的结果是 K-项集 $l_1[1]$、$l_1[2]$、…、$l_1[K-1]$、$l_2[K-1]$, 这样由 L_{K-1} 连接生成 C_K.

(2) 剪枝步, 因为频繁项集的非空子集必是频繁项集. 如果一个侯选 K-项集的(K–1)-子项集不在 L_{K-1} 中, 则该侯选项集也不可能是频繁项集, 从而可以对 C_K 进行剪枝, C_K 经剪枝后, 再对 C_K 中每个侯选项集进行支持计数.

2 改进的关联规则挖掘模式

在实际应用中, 对含有大量事务的事务数据库 D, 找每一个频繁 K-项集都需要先产生候选项集, 并扫描数据库, 这样明显影响算法复杂度, 降低了挖掘的计算效率. 另外, 传统的关联规则挖掘方法常生成无趣的强关联规则, 而未能生成含负属性项的有趣关联规则, 因此, 传统的关联规则挖掘模式在一定程度上限制了更有效的应用. 一些学者对这些问题进行了研究[5-10].

针对上述问题, 本文提出一种基于分辨矩阵(0-1 矩阵)的含负属性项的关联规则挖掘方法, 使用这种方法可以做到只对数据库扫描 1 次, 无需生成候选集, 即可得到频繁项集, 并将无趣的关联规则转化并生成含负属性项的有趣关联规则, 克服了上述缺点.

2.1 分辨矩阵及分辨向量

定义 **1** 设 $S=\{U,I,F\}$是一信息系统, 所有对象的非空有限集合 $U=\{X_1,X_2,\cdots, X_P\}$, 属性集 $I=\{I_1, I_2,\cdots, I_m\}$, F: $U\times I\to V$, V 是 I 的值域.

映射 φ_i: $V_i\to\{0, 1\}$: 当 $x\in \boldsymbol{v}_i^1$, $\varphi_i(x)=1$; 当 $x\in \boldsymbol{v}_i^2$, $\varphi_i(x)=0$ $\boldsymbol{v}_i^1\cup \boldsymbol{v}_i^2=V_i$ $i=1,2,\cdots,m$. 定义映射φ: $V\to\{0, 1\}$, $\forall\ x\in V$ $(x\in V_i)$ 有 $\varphi(x)=\varphi_i(x)$, 这样从 V 通过映射 φ 得到一个矩阵 $\boldsymbol{D}_{p\times m}$, 把 $\boldsymbol{D}$ 称为 I 的分辨矩阵. 第 j 列 $\boldsymbol{D}_j$ 称为项 I_j 的分辨向量:

$$\text{support-count}(I_j)=\sum_{i=1}^{p} d_{ij}.$$

定义 **2** R_{ij} 表示 2 项集$\{I_i, I_j\}$, 其分辨向量定义为 $\boldsymbol{D}_{ij}=\boldsymbol{D}_i\wedge\boldsymbol{D}_j$, 其中当 $d_{ki}=0$ 或 $d_{kj}=0$,$d_{kij}=0$; 当 $d_{ki}=1$ 且 $d_{kj}=1$, $d_{kij}=1$; K-项集 $R_{12\cdots K}=\{R_{12\cdots(K-1)}, I_K\}$, $R_{12\cdots K}$ 的分辨向量定义为 $\boldsymbol{D}_{12\cdots K}=(\boldsymbol{D}_1\wedge\boldsymbol{D}_2\wedge\cdots\wedge\boldsymbol{D}_{k-1})\wedge\boldsymbol{D}_k$

2.2 下标集及下近似集

定义**3** 由项集的集合中每个项集的下标构成的集合称项集下标集.

定义 **4** $L'(w)=\{y|y\in$ 非频繁项集下标集 L'且 $y\subset w\}$称下标 w 的下近似集 $L'(w)$.

2.3 利用分辨矩阵挖掘频繁项集 $\boldsymbol{L}$ 的算法

算法的具体步骤为: (1) 输入信息系统(数据库)S,并生成分辨矩阵 $\boldsymbol{D}$; (2) 输入最小支持度 $sup_{\min}$; (3) 利用 $\boldsymbol{D}$ 及 $sup_{\min}$ 产生 L_1, 令 K=2; (4) 由频繁项集 L_{k-1} 中的项集连接成 K-项集的集合 P; (5) 计算 P 的项集下标集 W; (6) 对每一 w_i, $\in W$ 计算其下近似集 $L'(w_i)$, 若 $w_i\in W$, 满足 $L'(w_i)\neq\varnothing$, 则 $W=W-\{w_i\}$, $L'=L'\cup\{w_i\}$; (7) 对剪枝后每一 w_i, $\in W$, 生成对应的分辨量 D_{wi}, 计算支持计数, 若支持计数$<sup_{\min}\times|D|$, 则 $W=W-\{w_i\}$, $L'=L'\cup\{w_i\}$; (8) 当 $W\neq\varnothing$, 则按 W 中的下标, 找到 P 中对应的项集, 形成频繁项集 L_k, 令 $K=K+1$ 转至(4), 否则输出信息系统(数据库)S 中满足最小支持度 $sup_{\min}$ 的频繁项集 L.

2.4 含负属性项关联规则

定义 **5** 事务数据库 $\boldsymbol{D}$, 对项集 $\boldsymbol{I}=\{i_1, i_2,\cdots, i_n\}$(正文字集)中的元素求反得负文字, 项集 A 包含

正文字，项集 B 包含正文字、负文字，则规则 $A \Rightarrow B$ 称为含负属性项的关联规则.

定义 6　事务数据库 $\boldsymbol{D}$，关联规则 $A \Rightarrow B$ 的期望置信度(expectedconfidence)定义为 D 中事务包含 B 的百分比.

期望置信度度量无条件时事务包含项集 B 的确定性:

$A \Rightarrow B$ 的兴趣度 $i=\dfrac{\text{confidence}(A \Rightarrow B)}{\text{expectedconfidence}(A \Rightarrow B)}$.

兴趣度度量事务包含项集 A 与事务包含项集 B 的相关程度.

表 1 中，因为“买台式电脑 $\Rightarrow$ 买笔记本”是强关联规则，其期望置信度 EC=7/8，兴趣度 i=0.86<1，所以，强关联规则“买台式电脑 $\Rightarrow$ 买笔记本” 缺乏意义，被过滤掉；当最小兴趣度 i_0=1.1 时，强关联规则“买笔记本 $\Rightarrow$ 不买台式电脑” (其支持度和置信度分别为 50%、57%)的期望置信度为 50%，兴趣度 i=1.14>i_0，所以，强关联规则“买笔记本 $\Rightarrow$ 不买台式电脑”是有意义的.

2.5　从 $\boldsymbol{L}$ 中挖掘含负属性项关联规则的算法

对于 $l_k \in L_k$, $k \geqslant 2$;

{H_1={l_k 生成只有一个项的规则后件};

genrules(l_k, H_1); }

输出规则集 $R=\bigcup_k R_K$.

genrules(l_k: 频繁 k 项集, H_m: 规则后件的 m-项集)

{若(m<=k–2)

{H_{m+1}=Apriori-gen(H_m);

对于每一个 $h_{m+1} \in H_{m+1}$

{$conf$=sup(l_k)/sup (l_k–h_{m+1});

ec=sup(h_{m+1});

int= $conf$/Ec;

若 int>$int_{\min}$

若($conf \geqslant conf_{\min}$),

则 R_K=R_K ∪ {l_k-h_{m+1} $\Rightarrow$ h_{m+1} with sup、$conf$、int}

若 int<1,干则对每一个反向项集 h_{m+1}' of h_{m+1}

{S=sup(l_k–h_{m+1} ∪ h_{m+1}');

C=sup(l_k–h_{m+1} ∪ h_{m+1}')/ sup(l_k–h_{m+1});

E=sup(h_{m+1}') ;

I=C/E

若($S \geqslant sup_{\min}$ and $C \geqslant conf_{\min}$ and $I \geqslant int_{\min}$),

则 R_K=R_K ∪ { l_k–h_{m+1} $\Rightarrow$ h_{m+1}' with S, C, I}}}

genrules(l_k, H_{m+1}); }}.

3　实例

3.1　数据及预处理

选用内蒙古科技大学 2003—2005 三届金融数学专业 158 名学生 28 门课程成绩表.

数据清理：(1)补考成绩的计算问题. 为了保证数据挖掘结果的准确与合理，对于不及格学生的成绩应以他们的原成绩为准而不按他们的补考成绩计算. (2)没有正常参加考试的成绩计算. 所有成绩均不计入挖掘数据中，以消除噪音、维护数据完整性.

数据转换: (1)成绩离散化. 将成绩数据转换为 2 个等级数据. (2)课程编码. 将成绩表中课程以 1、2, …, 28 编码.

3.2　挖掘频繁项集

最小支持度 $sup_{\min}$=0.25，挖掘得到的频繁项集如图 1 所示.

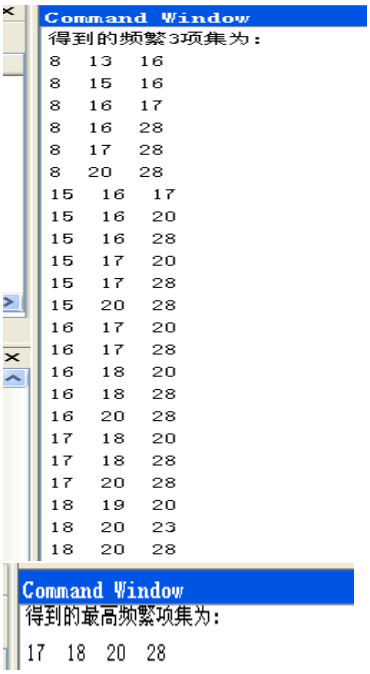

图 1　频繁项集

3.3　挖掘关联规则

当 $conf_{\min}$=0.75, $int_{\min}$=1.5 时得到的关联规则见图 2.

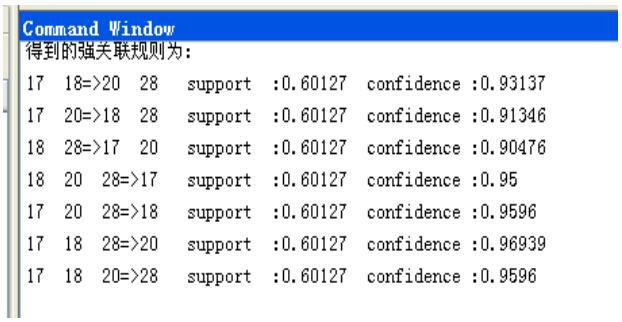

图 2　关联规则

由规则“保险学、金融工程学 75 分以上 $\Rightarrow$ 计量经济学、数理金融学 75 分以上”, 可得保险学、金融工程学、计量经济学、数理金融学同时优良的支持度为 60.127%, 说明它们之间具有较强影响力, 且这 3 门课程设置顺序合理, 同时计量经济学、数理金融学任课教师的教学有的放矢: 由保险学、金融工程学的成绩可对学生进行分类, 已知学生甲的成绩优良, 可预测学生甲的计量经济学、数理金融学成绩可能优良, 在教学中可对学生甲提出更高要求, 已知学生乙的成绩较差, 可预测学生乙的计量经济学、数理金融学成绩可能较差, 在教学中可给学生乙较多的帮助, 有利于教师分类教学.

4 参考文献

[1] Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large databases [C]. Washington: ACM Press, 1993: 207-216.

[2] Agrawal R, Srikant R. Fast algolithms for mining association rules in large databases [R]. WSA: IBM Almaden Research Center, 1994.

[3] Han Jiawei, Micheline Kamber. 数据挖掘概念与技术 [M]. 北京: 机械工业出版社, 2002: 152-158.

[4] Tan Pangning, Steinbach M, Kumar V, et al. Introduction to data mining [M]. Beijing: Posts & Telecom Press, 2006: 209-221

[5] Han Jiawei, Pei Jian, Yin Yiwen. Mining frequent patterns without candidate generation [C]. New York: ACM Press, 2000: 1-12.

[6] 朱玉全, 孙志挥. 快速更新频繁项集 [J]. 计算机研究与发展, 2003, 40(1): 94-99

[7] 颜跃进, 李舟军, 陈火旺. 基于 FP-Tree 有效挖掘最大频繁项集 [J]. 软件学报, 2005, 16 (2) : 215-222

[8] 吉根林, 杨明. 最大频繁项目集的快速更新 [J]. 计算机学报, 2005, 28(1): 129 -135.

[9] 柯丽, 王明文, 何世柱, 等. 基于频率共现熵的跨语言网页自动分类研究 [J]. 江西师范大学学报: 自然科学版, 2011, 35(3): 240-245.

[10] 栗晓聪, 滕少华. 频繁项集挖掘的 Apriori 改进算法研究 [J]. 江西师范大学学报: 自然科学版, 2011, 35(5): 498-502.

Mining Association Rules Including Negative Items Based on Recognizable Matrix

WANG Pei-ji[1], ZHANG Shu-ling[1], ZHAO Yu-lin[2]

(1. School of Mathematics, Physics and Biological Engineering, Inner Mongolia University of Science and Technology, Baotou Inner Mongolia 014010, China; 2. Branch 2, Inner Mongolia First Machinery Group Corporation, Baotou Inner Mongolia 014032, China)

Abstract: Recognizable matrix, association rules including negative items and submit an improved Apriori algorithm, aim at low efficiency of mining candidate itemsets, frequent itemsets and losing useful rules is presented. The implementation of algorithm by introducing an example is described.

Key words: data mining; recognizable matrix; interestingness; association rules

(责任编辑: 冉小晓)