第 36 卷 第 23 期 计 算 机 工 程 2010 年 12 月
Vol. 36 No. 23 Computer Engineering December 2010
·软件技术与数据库· 文章编号:1000—3428(2010)23—0079—03 文献标识码:A 中图分类号:TP311

动态加权关联规则算法的分析与实现

傅国强， 郭向勇
(深圳职业技术学院信息中心，广东 深圳 518055)

摘 要:加权关联规则算法存在 2 个不足:(1)不满足向下封闭性要求,即频繁集的子集未必是频繁集;(2)加权关联规则不能很好地处理不同项目的不同重要性,真正地体现不同项目重要性的不同。针对上述问题,提出一种动态加权关联规则算法,算法根据项目的重要性和最大频繁项目集数量确定项目不同阶段的不同权重,充分体现不同项目的重要性是不同的,从而使算法的向下封闭性得到证明。实验结果表明,该算法具有较高的准确性及效率。

关键词:动态加权; 关联规则; 向下封闭性

Analysis and Implementation of Dynamic Weighted Association Rule Algorithm

FU Guo-qiang， GUO Xiang-yong
(Information Center, Shenzhen Polytechnic, Shenzhen 518055, China)

【Abstract】 Weighted association rule algorithm for less than two: one does not meet the requirement of closed down, that is a subset of frequent sets may not be frequent sets; another weighted association rules can not handle the different importance of different items, and truly embody the the importance of different items different. In this paper, the dynamic weighted association rules algorithm, algorithm based on the importance of the project and the largest number of frequent itemsets at different stages of the project to determine the different weights, fully reflect the importance of different items are different, and algorithms are proven closed down. Experimental results show that the algorithm improves the high accuracy and efficiency.

【Key words】 dynamic weighted; association rule; downward closure

1 概述

关联规则挖掘是一个重要的实用的数据挖掘的算法,是用于在交易记录中发现各项目之间的关系。关联规则挖掘中最著名的算法是 Apriori 算法,该算法假设数据中各项目具有相同的重要性,即在该算法框架下,数据库中的各个项目以平等一致的方式处理。然而,在实际中不同的项目往往有着不同的重要性,这几乎是现实世界数据库的内在特征。加权关联规则被引入可以反映现实世界中各个项目的不同重要性。

近年来,对这一问题国内外已有相关文献。但可归为 2 类:(1)不满足向下封闭性要求;(2)不能很好地处理不同项目的不同重要性。文献[1]提出基于项目 K-支持期望概念的加权关联规则挖掘算法。算法能够满足封闭性,但支持度的计算增加一个项目集外的权重,使权重的意义大打折扣。另外还有许多加权关联规则算法都因为满足了权重的区别,却不能满足算法向下封闭性[2]的要求。

为了反映各个项目的不同重要性,在数据中发现加权关联规则,引入项目权值概念,从而扩展了现有的问题模型,提出一种新的加权关联规则算法,项目的权重是基于指数递阶的动态加权权重。

本文给出一种指数加权关联规则挖掘模型及算法,完全满足向下封闭性,并给出证明。算法模型充分体现项目之间的不同的重要性。仿真实验表明在大规模数据集挖掘中,算法的效率和所获得的规则的质量都得到明显改善。

2 关联规则算法分析

文献[3]提出了有效挖掘布尔关联规则的著名 Apriori 算法,此后又在关联规则中引入权重的概念,出现了许多加权关联规则算法,如文献[4]提出了一种模糊关联规则算法,实现了加权关联规则的挖掘。现有的加权关联规则主要实现方法是以 Apriori 为基础,引入项目的权重。

设 $I=\{i_1,i_2,\cdots,i_m\}$为项的全集,每个项都有一个权值与之对应。它们的权值分别是$\{r_1,r_2,\cdots,r_m\}(r_i\in[0,1])$,最小加权支持度为 wminsup,最小可信度为 wminconf。

定义 1 设项目集 $X=\{i_1,i_2,\cdots,i_k\}$,则项目集 X 的加权支持度[5] wsup 为:

$$\text{wsup}(X)=\max\{r_1,r_2,\cdots,r_k\}\times\sup(X)$$

其中,$\sup(X)$为 X 的传统支持度计数,$\max\{r_1,r_2,\cdots,r_k\}$称为 X 的权值,如果 $\text{wsup}(X)\geqslant\text{wminsup}$,则 X 是加权频繁集。

定义 2 设项目集 $X=\{i_1,i_2,\cdots,i_k\}$,则项目集 X 的加权支持度 wsup 为:

$$\text{wsup}(X)=\sum\{r_1,r_2,\cdots,r_k\}\times\sup(X)$$

基金项目:深圳市科技基金资助重点项目(06KJh041)
作者简介:傅国强(1966—),男,高级工程师,主研方向:数据挖掘,软件工程;郭向勇,研究员
收稿日期:2010-02-25 **E-mail**:fredfu@tom.com

其中，sup(X)为 X 的传统支持度计数；$\Sigma\{r_1,r_2,\cdots,r_k\}$称为 X 的权值，如果 wsup(X)≥wminsup，则 X 是加权频繁集。但以上 2 个方法确定项目集权重都不能保证关联规则向下封闭性成立。

wsup(X)=($\Sigma r_i+\Sigma r_j$)×sup(X)(所有不属于 X 中的项目的 k 个最大的权重之和)。由于增加后一个式子明显改变的加权的意义，甚至完全否认项目集本身的权重。

3 基于动态加权关联规则模型

定义 3 设项目集 $X=\{a_1,a_2,\cdots,a_k\}$，相应的权重 $R=\{r_1,r_2,\cdots,r_k\}$，则项目集 X 的加权支持度 wsup 为：

$$\text{wsup}(X)=\sum_{i=1}^{k}r_i\times\frac{1}{\max\{r_1,r_2,\cdots,r_k\}}\times 2^{-k}\times\text{sup}(X)$$

其中，sup(X)为 X 的传统支持度计数；$\sum_{i=1}^{k}r_i$ 为权重之和；$\max\{r_1,r_2,\cdots,r_k\}$称为 X 的最大权值，如果 wsup(X)≥wminsup，则 X 是加权频繁的。

定理 1 对于项目集 X 为 k 项加权频繁集，Y 为 X 的 k-1 项子集，即 Y∈X，则 Y 为 k-1 项频繁集。

证明：

X 动态为加权支持度为：

$$\sum_{i=1}^{k}r_i\times\frac{1}{\max\{r_1,r_2,\cdots,r_k\}}\times 2^{-k}\times\text{sup}(X)$$

为简化起见，设 $r_1\leqslant r_2\leqslant,\cdots,\leqslant r_k$，$\text{wminconf}(X)=\frac{\sum_{i=1}^{k}r_i}{r_k\times 2^k}\times\text{sup}(X)$，Y 为 k-1 项集，可能有 2 种情况：(1)Y 中包括 $r_1,r_2,\cdots,r_{k-1}$即少最后一项；(2)Y 中包括 $r_1,r_2,\cdots,r_{j-1},r_{j+1},\cdots,r_k$，X 中少中间项 r_j。

下面分别对 2 种情况证明：

$$(1)\text{wminconf}(Y)=\sum_{i=1}^{k-1}r_i\times\frac{1}{\max\{r_1,r_2,\cdots,r_{k-1}\}}\times 2^{-k+1}\times\text{sup}(Y)=\frac{\sum_{i=1}^{k-1}r_i}{r_{k-1}\times 2^{k-1}}\times\text{sup}(Y)$$

由支持度的定义不难看出，sup(X)≤sup(Y)，所以，

$$\text{wminconf}(X)\text{-wminconf}(Y)=\frac{\sum_{i=1}^{k}r_i}{r_k\times 2^k}\times\text{sup}(X)-\frac{\sum_{i=1}^{k-1}r_i}{r_{k-1}\times 2^{k-1}}\times\text{sup}(Y)\leqslant\frac{\sum_{i=1}^{k}r_i}{r_k\times 2^k}\times\text{sup}(Y)-\frac{\sum_{i=1}^{k-1}r_i}{r_{k-1}\times 2^{k-1}}\times\text{sup}(Y)=\frac{r_{k-1}\sum_{i=1}^{k}r_i-2r_k\sum_{i=1}^{k-1}r_i}{r_kr_{k-1}\times 2^k}\times\text{sup}(Y)=\frac{r_{k-1}r_k+r_{k-1}\sum_{i=1}^{k-1}r_i-2r_k\sum_{i=1}^{k-1}r_i}{r_kr_{k-1}\times 2^k}\times\text{sup}(Y)=\frac{r_k(r_{k-1}-\sum_{i=1}^{k-1}r_i)+(r_{k-1}-r_k)\sum_{i=1}^{k-1}r_i}{r_kr_{k-1}\times 2^k}\times\text{sup}(Y)$$

因为 $r_j\leqslant r_k$，所以上式≤0，wminconf(X)≤wminconf(Y)。

$$(2)\text{wminconf}(X)\text{-wminconf}(Y)=\frac{\sum_{i=1}^{k}r_i}{r_k\times 2^k}\times\text{sup}(X)-\frac{\sum_{i=1}^{k}r_i-r_j}{r_k\times 2^{k-1}}\times\text{sup}(Y)\leqslant\frac{\sum_{i=1}^{k}r_i}{r_k\times 2^k}\times\text{sup}(Y)-\frac{\sum_{i=1}^{k}r_i-r_j}{r_k\times 2^{k-1}}\times\text{sup}(Y)=\frac{-\sum_{i=1}^{k}r_i+2r_j}{r_k\times 2^k}\times sup(Y)\leqslant\frac{-(r_j+r_k)+r_j}{r_k\times 2^{k-1}}\times\text{sup}(Y)$$

因为 $r_j\leqslant r_k$，所以上式≤0。

综上所述，wminconf≤wminconf(X)≤wminconf(Y)，所以，子集 Y 为 K-1 项频繁集。

定理 2 对于项目集 X 为 k 项加权频繁集，Z 为 X 的子集，即 Z∈X，则 Z 为 k-1 项频繁集。

证明：由定理 1 可得：

wminconf(X)≤wminconf(Y(k-1))

wminconf(Y(k-1))≤wminconf(Y(k-2))

依此类推，对任意维子集 Z，wminconf(X)≤wminconf(Z)，所以 Z 为加权频繁集。

性质 对于项目集 X 有：wsup(X)≤sup(X)。

证明：因为 0≤hi≤1，所以有：wsup(X)=max{$h_1,h_2,\cdots,h_k$}×sup(X))≤sup(X)。

命题 如果项目集 X 是加权频繁的，Y 为 X 子集，则 sup(Y)≥wminsup。

证明：因为 X 是加权频繁的，由性质可知，sup(X)≥wminsup。而由上面性质易得，sup(Y)≥sup(X)。所以有 sup(Y)≥wminsup。

推论 如果项目集 Y 为非频繁集，即：sup(Y)<wminsup，则 Y 的任意超集都不是加权频繁集。

3.1 指数加权关联规则频繁集发现算法

本文的关联规则挖掘算法的向下封闭性在上一节已经得到了。但算法在计算支持度已将值缩小了为原支持度的 2^{1-k} 倍，明显缩小，为了发现频繁集需要将最少支持度的值同样缩小，同样根据可能的最大频繁集维数动态确定，如 K 维则缩小为 2^{1-k} 倍。挖掘频繁项集的具体算法描述如下：

算法 学习模式发现

输入 交易库 D，最小支持度 min_sup，最大可能频繁集维数 m，相应权重 $R=\{r_1,r_2,\cdots,r_k\}$

输出 频繁项集 L

```
(1)  L1=find_frequent_1-itemsets(D); // 挖掘频繁 1-项集，
(2)  min_sup=min_sup * exp((1-k) * ln(2))
(3)  for (k=2;Lk-1≠Φ;k++)
(4)  Ck=apriori_gen(Lk-1,min_sup); // 调用 apriori_gen 方法生
     //成候选频繁 k-项集
(5)  for each transaction t∈D { // 扫描事务数据库 D
(6)    Ct=subset(Ck,t);
(7)    for each candidate c∈Ct
(8)    c. count++; // 统计候选频繁 k-项集的计数
(9)  CR=Get_role(c, R);
(10) m=|CR|
(11) For(i=1; i<=m; i++) {
(12) c. support=c. support+CRi * c. count * exp((1-m) * ln(2));
(13)   }
(14) Lk={c∈Ck|c. count≥min_sup} // 满足最小支持度的 k-项
     //集即为频繁 k-项集
(15) }
(16) return L=∪k Lk; // 合并频繁 k-项集(k>0)
```

其中，第(1)步挖掘到的 1 项频繁项集，置信度大于给定最小置信度 Min_sup 的关联规则称为频繁关联规则(Frequent Association Rule)。

第(12)步计算支持度：

$$\sum_{i=1}^{k} r_i \times \frac{1}{\max\{r_1, r_2, \cdots, r_k\}} \times 2^{-k} \times \sup(X)$$

第(13)步，首先需要从频繁项集入手，挖掘出全部的关联规则(或者称候选关联规则)，然后根据 min_sup 来得到频繁关联规则。

```
Procedure
apriori_gen(Lk-1:frequent(k-1)-itemsets;min_sup:minimum,
      support threshold)
  for each itemset l1∈Lk-1
for each itemset l2∈Lk-1
if (l1[1]=l2[1])∧(l1[2]=l2[2])∧
(l1[k-2]=l2[k-2])∧(l1[k-1]=l2[k-1]) then {
  c=l1 join l2;   file://join step;generate candidates
  if has_infrequent_subset(c,Lk-1) then
    delete c   file://prune step:remove unfruitful candidate
  else add c to Ck;
  }
return Ck;
procedure has_infrequent_subset(c:candidate k-itemset;
     Lk-1:frequent(k-1)-itemset);
for each (k-1)-subset s of c
if s Lk-1 then
return TRUE;
return false;
```

3.2 加权关联规则的挖掘

上节设计的基于加权关联规则的学习模式发现算法对交易资源库使用日志数据库中挖掘出频繁集 Lk。现在要从频繁集 Lk 挖掘出强关联规则，也就是满足最小支持度和最小信任度的强关联度规则，可用条件概率计算关联规则的置信度：

$$\text{confidence}(A \rightarrow B) = P(B|A) = \frac{\text{sup port\_count}(A \cup B)}{\text{sup port\_count}(A)} = \frac{\text{C. count}}{\text{A. count}} \quad (1)$$

其中，C=A∪B，C. count 和 A. count 为包含项集 C 和 A 的各交易记录的加权数的和。

在所有形如“学生-教学资源”的频繁项集，根据式(1)得到的频繁项集导出置信度不小于最小置信度的强规则，不同类型的学生对应不同的教学资源。

挖掘频繁关联规则的算法描述如下：

```
(1)  L=∪k Lk; // L 是频繁项集集合,
(2)  AR=Φ; // AR 是频繁关联规则集合
(3)  for all λk∈L(λk 是 L 的元素,是一个 k-频繁项集,大小为 n){
(4)    for all αk (αk 是 λk 的非空真子集){
(5)      if(αk→βm 的置信度>=min_Conf)  { //这里,m+k=n,
//其中 αk→βm 是一关联规则
(6)        AR=AR∪(αk→βm);
(7)      }
(8)    }
(9)  }
(10)  return AR;
```

4 算法性能评测

为了检测算法性能，笔者利用动态加权关联规则算法分析教学资源库使用日志的数据进行分析，日志的数据由{term，major，source_type，source_name，Sex}5 个属性组成，在加权过程中，权值的设置具有灵活性，主要根据专家的经验，同时充分考虑了用户分析问题侧重面，如用户更关注专业年级的学生使用教学资源的习惯，可以赋给 term，major，source_type 属性更高的权值。各属性分配权值如下：

{term，major，source_type，source_name，Sex}

0.6，0.7，0.7，0.1，0.2，

将 min_sup 设置为 0.2、min_conf 设置为 0.8 的情况下，频繁集维数为 2，根据动态加权的原则，min_sup=0.05，min_conf=0.2，算法对教学资源库使用日志的数据进行了分析，挖掘出所有频繁项集，并且生成关联规则。部分挖掘结果如表 1 所示。

表 1 关联规则表

关联规则	动态加权支持度	置信度
计算机系→程序设计	0.067 4	27.611
CAD/CAM→特种加工	0.071 5	26.534
珠宝专业→首饰	0.051 3	25.527
服装专业→服装	0.058 0	27.543

由此可看出不同专业学生使用习惯。挖掘出的关联规则用语义形式表达，容易被人所理解，也符合实际情况，对改进系统智能水平有很强的指导性意义。

为了测试动态加权关联规则算法，将不同数量记录的数据运用算法运行在硬件环境下，效果如图 1 所示。

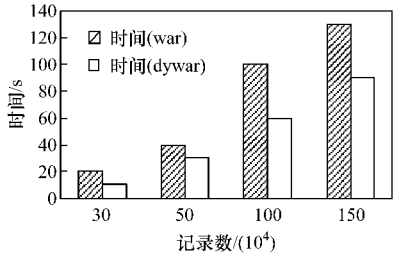

图 1 算法效率比较

可以看出，随着处理记录数的增大，改进后的算法运行效率提高更加明显，节约的时间更多。所以，算法总体耗时更多。动态加权关联规则算法在保证算法向下封闭性基础上，充分体现不同项目重要性的不同，运行效率得到了提高。

5 结束语

本文分析传统的数据挖掘算法关联规则的不足，提出在关联规则中引入基于动态的加权概念，给出了基于动态加权的关联规则分析算法。此算法保证了向下封闭性，给出了理论上的证明，同时充分体现了不同项目的重要性的不同。实验表明，动态加权关联规则算法准确性和效率都得到有效提高，发现规则时间缩短，得到的关联规则能满足用户需求。

今后的研究工作将根据实际使用效果如何优化算法，同时，基于动态加权关联规则挖掘方法中权重在一些典型应用的确定上原则进行深入研究，进一步提高使用效果。

参考文献

[1] 欧阳为民，郑　程，蔡庆生，等. 数据库中加权关联规则的发展[J]. 软件学报，2001，12(4)：612-619.

[2] 毛国君，段立娟，王　实，等. 数据挖掘原理与算法[M]. 北京：清华大学出版社，2007.

[3] Kamber M. Data Mining：Concepts and Techniques[M]. San Francisco，USA：Morgan Kaufmann Publishers，2001.

[4] 杜　北，李伟华，史豪斌. 一种新的模糊加权关联规则挖掘算法[J]. 计算机工程，2008，34(20)：218-220.

[5] 张智军，方　颖，许云涛. 基于 Apriori 算法的水平加权关联规则挖掘[J]. 计算机工程与应用，2003，39(14)：197-199.

编辑：陈　文