

一种新的在图像关联规则挖掘中产生频繁项集的方法

杜 琳¹, 陈云亮^{1,2}, 朱 静¹

(1 中国地质大学 计算机学院, 湖北 武汉 430074;

2 华中科技大学 计算机学院 武汉光电国家实验室, 湖北 武汉 430074)

摘 要: 提出了一种进行图像关联规则提取时产生频繁项集的方法——频繁项树. 为便于频繁项树的运用, 使用了 bSQ 的图像数据格式来重新组织图像数据, 并在此基础上提出了频繁项树的截断、半深度优先、图像掩模和多层次灰度范围自动生成等优化技术, 降低了算法的时间和空间复杂度, 使其具有较高的运行效率和实用价值.

关键词: 数据挖掘; 关联规则; 图像; 频繁项树; bSQ

中图分类号: TP301.6

文献标识码: A

文章编号: 1000-7180(2011)07-0121-04

A New Approach to Generate Frequent Itemsets in Mining Image Association Rules

DU Lin¹, CHEN Yun-liang^{1,2}, ZHU Jing¹

(1 Computer Science Department, China University of Geosciences, Wuhan 430074, China;

2 Wuhan National Laboratory for Optoelectronics, Huazhong University of Science & Technology, Wuhan 430074, China)

Abstract: This paper proposes an approach to generate frequent itemsets in mining image association rules —— Frequent Item Tree. We utilize bSQ image format to re-organize the image data to apply frequent item tree. Moreover, this paper propose several optimization techniques, including frequent item tree pruning, semi-depth-first search, image mask, and multi-level gray generation, to decrease the time and space complexity.

Key words: data mining; association rule; image; frequent item tree; frequent itemsets; bSQ

1 引言

关联规则的挖掘已成功地运用到大型图像数据库中^[1]. 关联规则挖掘有两种主要方法: (1) 从大型图像集合中挖掘; (2) 在图像集合及相关联的字符数据的混合体中进行挖掘.

事务数据库中的关联挖掘是大型数据库中关联规则挖掘的典型情况. 在这种情况下, 可以通过考察事物的数据来产生关联规则. 数据是明确的, 并且对每个组成项都有具体明确的数值, 一个客户的事务(交易)将包括这些项的子集. 但在图像数据库中, 人工标识所有图像实际上是不可能的^[2], 在执行挖掘之前, 只能依赖自动的或半自动的图像内容来分析, 产生的描述可以是颜色、纹理、形状和大小等.

频繁项集(Frequent Itemset)在某些文献中又被称作大项集(Large Itemset), 其概念是针对事务数据库的. 在事务数据库中, 项的集合被称为项集(Itemset), 如果项集的出现频率(即包含项集的事务数)大于或等于最小支持度与事务数据库中事务总数的乘积, 则把该项集称为频繁项集(大项集).

频繁项集的产生是关联规则挖掘中的最关键和最困难的步骤^[3].

2 图像数据格式

一般情况下, 图像数据可以分为两类: 栅格数据、矢量数据^[3]. 文中主要对栅格数据图像进行多层次关联规则的提取.

为了便于对图像进行多层次的关联规则的挖

收稿日期: 2010-08-25; 修回日期: 2010-09-07

基金项目: 国家自然科学基金项目(60603074); 国家“九七三”重大项目(2004CB318203)

掘,文中使用了一种新的图像数据格式 bSQ(bit Sequential). 将原来图像中每个像素点的(0~255)灰度值转化为 8 位的二进制,由此,可按 bit 位提取每个像素点的当前 bit 值,将原来由 0~255 灰度值构成的一个图像文件转化为 8 个由 0、1 灰度值构成的图像文件,其分别代表每个像素点在相对应的 bit 位上的值. 文中给出了一个 bSQ 数据格式的简单例子:

$$\begin{bmatrix} 127 & 254 \\ 14 & 193 \end{bmatrix} \rightarrow \begin{bmatrix} 01111111 & 11111110 \\ 00001110 & 11000001 \end{bmatrix}$$
$$\rightarrow \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

3 频繁项树

3.1 图像频繁项集

针对图像而言,可以定义图像关联规则挖掘中图像频繁项集的概念. 文中针对栅格图像数据进行图像数据挖掘,利用图像中各像素的灰度值,希望挖掘出来的规则是形如 $Image1[192, 255] \Rightarrow Image2[128, 255]$ 这样的规则. 因此,可将图像频繁项集作出定义.

图像频繁项集:由两图(或多图)灰度值范围所

组成的,且满足此灰度值范围的像素点数大于或等于最小支持度与图像总像素点数的乘积,这些灰度值范围所组成的项集就被称为图像频繁项集.

3.2 图像频繁项树

图像频繁项树(简称:频繁项树)是文中提出的一种运用逐层搜索的办法提取和保存图像频繁项集的数据结构. 频繁项树是建立在 bSQ 图像数据格式的基础上的一种树形结构. 在对两幅图像进行关联规则提取时,从数据结构的角度来说,频繁项树是一种四叉树. 其 4 个分支分别对应两个 bSQ 图像文件中,相同层次的灰度值 0(第一个 bSQ 文件某层的灰度值)0(第二个 bSQ 文件某层的灰度值,下同)、01、10、11. 对于每一个频繁项树的结点,我们保存了要进行关联规则提取的每个 bSQ 文件当前层的灰度值(只能是 0 或 1)、当前频繁项集的支持度、以及支持当前频繁项集的像素点在图像中的位置等信息. 频繁项树中每个结点的基本结构和一个简单的频繁项树的例子,如图 1、图 2 所示.

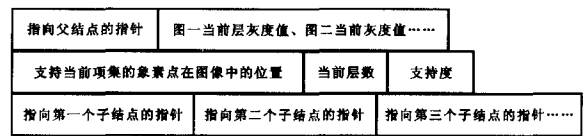


图 1 频繁项树中每个结点的基本结构

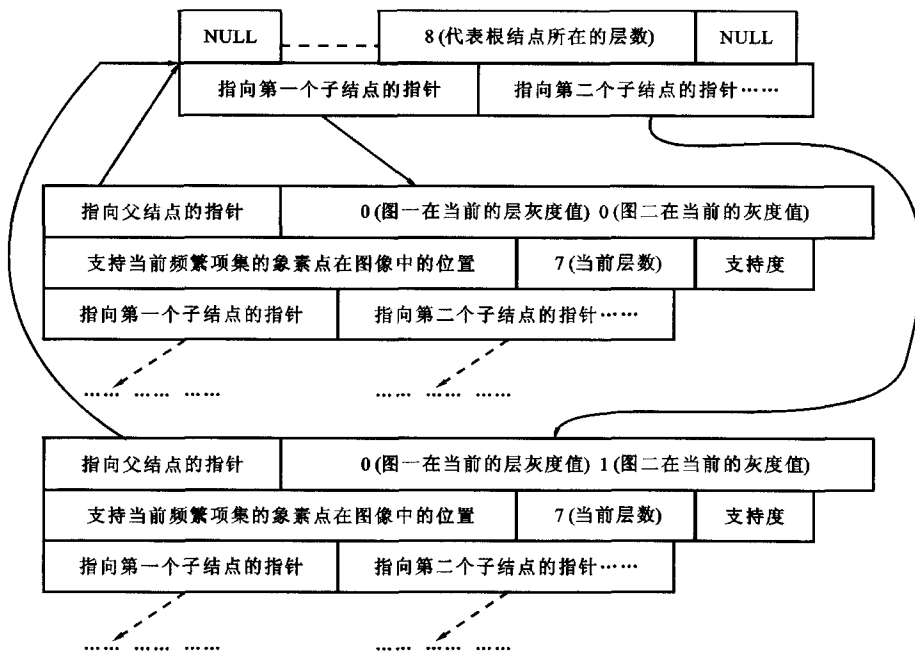


图 2 频繁项树的一个简单例子

4 利用频繁项树进行图像关联规则提取算法

一般情况下,关联规则的挖掘过程可分为两个

步骤:(1)发现所有的频繁项集,也就是支持度大于给定最小支持度的项集;(2)对于每个频繁项集,产生所有比最小置信度大的规则.

其中,第一步发现所有的频繁项集是关联规则

挖掘的关键所在,其总体性能也是由这一步决定的^[2].

针对图像进行的关联规则挖掘算法同样也要经过这两个步骤,频繁项树就是针对图像,沿用了Apriori算法逐层搜索来找到频繁项集的方法^[4].对于转化为bSQ格式的图像数据,每个层次都有可能产生若干个频繁项集,每个频繁项集都是以结点的形式存储在频繁项树中.由于要挖掘出低层的关联规则,因而对于每一个已存在于频繁项树中的频繁项集,还要根据下一个层次的bSQ格式的图像数据提取更低层的频繁项集,直到最后无法再找到出满足支持度要求的频繁项集.在以频繁项树的方法找到所有的频繁项集后,针对每个满足要求的频繁项集计算其各种置信度,并最终得到满足支持度和置信度的规则.其具体做法如下^[1]:

(1)将原图像数据转化为bSQ格式图像数据.

(2)对于bSQ数据,由高到低以递归的方法逐层找到频繁项集,并将频繁项集的相关信息(如:图像灰度值,支持当前频繁项集的像素点的位置,当前频繁项集的支持度等)存于频繁项树中.频繁项树中各个层次符合支持度要求的频繁项集组成了所有的图像频繁项集,越靠近根结点的是越抽象的频繁项集.其递归算法如下:

```

Procedure CreateFITree(depth)
{
  Get bSQ data;
  For (I=0; I<PixelNum; I++)
  {
    if ((GreyValue of Image1 = 0) & (GreyValue of
      Image2 = 0) & ...) count0++;
    if ((GreyValue of Image1 = 0) & (GreyValue of
      Image2 = 1) & ...) count1++;
    .....
  }
  Calculate Support of FrequentItem0, FrequentItem1,
  ..., FrequentItemN;
  For (I=0; I<N; I++)
  {
    If (Support of FrequentItemI >= Support)
      Add FrequentItemI into FITree
    Else RuleNode = NULL;
  }
  depth--;
  For (I=0; I< Num of FITree; I++)
  {
    If (RuleNode <> NULL) CreateFITree(depth)
  }
}

```

```

}
}

```

(3)在形成的频繁项树中,提取出符合要求(包括支持度和层次要求等)的所有频繁项集.

(4)根据原始图像(非bSQ数据格式),针对每一个频繁项集计算其各种置信度(两幅图像将计算2个,三幅图像将计算6个),并得到最终满足要求的各种图像关联规则.

5 相关优化技术

5.1 频繁项树的截断

频繁项树的截断是跟据Apriori性质,也即其反单调性实现的.在针对图像形成频繁项树时,是运用的逐层搜索的方法^[3-4].低层频繁项集是高层频繁项集的细化,若高层频繁项集已经无法满足支持度的要求,其细化的低层频繁项集更不可能满足相关要求.基于此,在形成频繁项树时,若当前频繁项集不满足支持度,在频繁项树中将当前结点赋为NULL,从而对频繁项树在当前位置进行截断,无需就当前频繁项集进一步细化.

5.2 半深度优先

在对树形结构进行生成和遍历时,一般使用深度优先或广度优先的方法.同样,在生成和遍历频繁项树时,也需要用到这样一种方法.若使用深度优先的方法,对于每个频繁项集结点将逐一生成其子频繁项集结点,这样,能生成多少个子频繁项集结点,就需要对原bSQ图像数据进行多少次扫描,影响了效率;若使用广度优先的方法,用递归的方法则由较为困难.因此,文中使用了一种半深度优先的生成树和遍历树的方法.其针对频繁项树中的每个结点,先以广度优先的方法,只对图像进行一次扫描,生成其所有的子频繁项集结点,而后,再以深度优先的方法,对其子频繁项集结点逐个以递归的方法继续进行运算,从而生成整棵频繁项树.

5.3 图像掩模

在产生图像所有频繁项集(形成频繁项树)时,低层次的频繁项集必须是建立在高层次频繁项集基础上的.因此,在频繁项树形成的过程中,要产生新的频繁项集的结点,必须是针对那些满足其上层频繁项集结点的像素点.通过设置图像掩模的方法来解决这一问题.对于每一个满足支持度的频繁项集,用图像掩模的方法保存支持当前频繁项集的像素点在图像中的位置,而后在生成下层频繁项集时,则只需要针对这些像素点进行计算,虽然稍微加大了空

间复杂度,但算法的效率有了较大提高.

5.4 多层次灰度范围自动生成

在进行图像关联规则的提取时,难点之一是如何以尽可能小的时间复杂度和空间复杂度来保存和计算支持关联规则的图像灰度值的取值范围.文中采取的方法是在频繁项树中仅保存支持当前频繁项集的各图在当前层的灰度值(只可能为 0 或 1).在频繁项树形成后,从频繁项树中提取所有符合要求的频繁项集时,由当前频繁项集结点保存的灰度值,倒推其父结点、祖结点(最终到根结点)等保存的灰度值(即支持当前频繁项集的各图在上面若干层的灰度值),利用位运算进行组合,形成支持当前频繁项集的图像灰度值取值范围(对于每一图像都包括最小灰度值和最大灰度值),最终完成图像关联规则提取.

通过实验说明了使用规则树进行图像多层次关联规则提取的方法的有效性.

6 结束语

文中提出了一种进行图像关联规则提取时产生频繁项集的方法——频繁项树.利用频繁项树进行图像关联规则提取非常简单,在设定好初始参数(包括置信度和支持度、规则挖掘的深度)的情况,可以无需人工介入,全自动的提取出各种符合要求的图像规则.为便于频繁项树的运用,使用了 bSQ 的图像数据格式来重新组织图像数据,并在此基础上提出了频繁项树的截断、半深度优先、图像掩模和多层次灰度范围自动生成等优化技术,降低了算法的时间和空间复杂度,使其具有较高的运行效率和实用价值.

参考文献:

- [1] William Perrizo, Qing Ding, Amalendu Roy. Deriving high confident rules from spatial using peano count trees [J]. Springer-Verlag, LNCS 2118, 2001.
- [2] Lee A-J-T, Lin-W-C Wang C S. Mining association rules with multi-dimensional constraints[J]. The Journal of Systems and Software, 2006 (79):79 - 92.
- [3] Wang J, Han J, BIDE: efficient mining of closed sequences [C]// Proc of the IEEE International Conference on Data Engineering. Boston, Massachusetts: IEEE, 2004: 79 - 90.
- [4] Wang Y, Image indexing and similarity retrieval based on spatial relationship model[J]. Information Sciences, 2003,(154):39 - 58.
- [5] Qin Ding, Qiang Ding, William Perrizo. Association Rule Mining on Remotely Sensed Images Using Ptrees [C]// Proceedings of 15th International Conference on Computer Applications in Industry and Engineering (CAINE'02), San Diego, CA: IEEE,2002: 138-141.
- [6] Zhang J, Hsu W, Lee M L. An information-driven framework for image mining[C]// Proceedings of 12th International Conference on Database and Expert Systems Applications (DEXA), Munich, Germany: IEEE, 2001.

作者简介:

杜琳女,(1978—)博士研究生,讲师.研究方向为图形图像处理;

陈云亮男,(1979—),博士研究生.研究方向为图像数据存储、人工智能;

朱静女,(1974—),讲师.研究方向为图像处理.

(上接第 120 页)

- [2] Suzuki N., Kurotaki S., Suzuki M., et al. Implementing and evaluating stream applications on the dynamically reconfigurable processor. [C]// 12th Annual IEEE Symposium on Field-Programmable Custom Computing Machines, Napa, California: IEEE, 2004, 328 - 329.
- [3] 刘洋,尹蕾,李广军. 基于流水线技术的可重构体系结构的研究与设计[J]. 微电子学,2008,38(4):593-595.
- [4] 刘洋,李广军. 基于函数级原型的流水线可重构结构的研究[J]. 小型微型计算机系统,2009,(6):1237-1239.
- [5] 黄乐天,李广军. 一种用于数字信号处理的可重构基核

研究[J]. 微电子学与计算机,2009,26(3):56-59.

- [6] 潘伟,刘欢,李广军. 一种新型高速低成本可重构 FFT 处理器结构[J]. 微电子学与计算机,2008,25(11):29 - 32.

作者简介:

黄乐天男,(1984—),博士研究生.研究方向为通信中的信号处理、通信专用集成电路设计、可重构计算.

李广军男,(1950—),教授、博士生导师.研究方向为通信中的信号处理、通信专用集成电路设计.