

文章编号:1001-9081(2005)07-1592-03

基于映射位集合的遥感图像关联规则挖掘

黄端琼,陈崇成,黄洪宇,樊明辉

(福州大学 福建空间信息工程研究中心,福建 福州 350002)

(lexin_223@163.com)

摘 要:提出 MBSA 算法,采用 Java 中的 TreeMap 的映射技术和压缩的 BitSet 来存储大量的布尔变量值,并且该算法只扫描一次事务数据库,用 BitSet 的逻辑“与”操作来代替数据库的扫描,有效提高了运行速度。将该算法应用到遥感图像挖掘中,提取遥感图像中红、绿、蓝波段与农作物产量之间的关联,为提高农作物产量提供有益的支持。

关键词:遥感图像;数据挖掘;关联规则;位集合;TreeMap

中图分类号: TP311;TP75 **文献标识码:** A

Map-based BitSet association rule mining of remote sensing image

HUANG Duan-qiong, CHEN Chong-Cheng, HUANG Hong-yu, FAN Ming-hui

(Spatial Information Research Center of Fujian, Fuzhou University, Fuzhou Fujian 350002, China)

Abstract: MBSA (Map-based BitSet Association Rule) algorithm was presented which used TreeMap class and a compressed BitSet class in Java to store Boolean values. MBSA algorithm scanned the transaction database only once and further database scans were replaced by BitSet logical AND operation, which efficiently speeded up the computation. MBSA algorithm had been applied to mine the association rules of red, green and blue bands associated with crop yield from remote sensing image of crop. It is useful for improve crop production.

Key words: remote sensing image; data mining; association rule; BitSet; Tree Map

0 引言

随着空间对地观测技术的发展,人们已获取并积累了同一地区多平台、多时相、多尺度、多源、多分辨率的海量遥感信息,成为国家空间数据基础设施(NSDI)的重要组成部分,也是数字城市、数字地球的重要信息源。由于理论基础和技术水平限制,遥感数据中隐含的丰富知识远没有得到充分的发掘和利用。将数据挖掘技术应用于遥感影像库,能够挖掘隐藏在遥感影像中丰富的时间、空间和光谱知识和规则,为智能信息处理服务。国内已开始一些研究,如文献[1]研究了卫星遥感数据的关联规则挖掘及其在土壤侵蚀和退耕还林上的应用,选择 Landsat 卫星 ETM 数据和地形高程数据(DEM)的派生数据进行关联规则挖掘。

关联规则是数据挖掘的一个重要研究方向,其算法主要包括 Apriori 算法及其种种变型^[2-5],主要的缺点在于需要产生大量的候选集,并且需要多次扫描数据库。Apriori TID^[6]是对 Apriori 的一种扩展,不依赖于原有的数据集,而是以现有的候选集来代替每次的交易。Park 等提出的 DHP 算法^[7]利用频繁项集的 Apriori 属性,使用散列表提高关联规则挖掘的效率,但是存放项集计数值的散列表与存放候选集的散列表之间存在内存争用问题。Brin 等提出的 DIC 算法^[8],动态地评估以被计数的所有项集的支持度。Han 提出了 FP-growth^[9]算法。该算法只进行 2 次数据库扫描,不会有庞大的候选集产生,减少了内存临时空间的占用,但是随着频繁模式树的增长可能使效率变得很差。

为了克服 Apriori 算法生成大量候选集的特点,解决 DHP 算法中存放项集计数值的散列表与存放候选集的散列表之间存在内存争用问题,本文提出一种新的 MBSA (Map-based BitSet Association Rule) 挖掘算法。该算法采用基于位集合和映射的概念,将大数据库压缩到位集合中,对其进行位集合的逻辑“与”操作来提高挖掘效率,无需生成候选集,且只需一次遍历数据库,有效地提高了计算速度。

1 位集合构建

1.1 问题描述

假设 $I = \{i_1, i_2, \dots, i_m\}$ 是项的集合,另设任务相关的数据 D 是数据库中事务的集合,其中每个 T 是项的集合,使得 $T \subseteq I$ 。每个事务有一个标识符,称为 TID。设 A 是一个项集,事务 T 包含 A 当且仅当 $A \subseteq T$ 。关联规则形如 $A \Rightarrow B$,其中 $A \subset I$, $B \subset I$ 并且 $A \cap B = \emptyset$ 。并不是所有的规则都是有用的,需要的是实用的和可靠的规则。关联规则的评价标准主要是支持度和可信度。只有那些超过最低支持度阈值和最低可信度阈值的规则称为强关联规则。

支持度 s : 设 D 中有 $s\%$ 的事务同时支持物品集 A 和 B ,可信度 c : 设 D 中有 $c\%$ 支持 A 的事务的同时也支持 B 。可信度是对关联规则准确度的衡量,支持度是对关联规则重要性(适用范围)的衡量。支持度说明这条规则在所有事务中有多大的代表性,支持度越大,关联规则越重要,应用越广泛。用 $P(A)$ 表示事务中出现物品集 A 的概率, $P(B | A)$ 表示物品集 A 的事务中,出现物品集 B 的概率,它们的表达式分别为:

收稿日期:2004-12-16 基金项目:福建省教育厅项目(K04001);福建省青年科技人才创新项目(2004J005)

作者简介:黄端琼(1982-),女,福建南安人,硕士研究生,主要研究方向:空间数据挖掘、可视化技术; 陈崇成(1968-),男,福建闽清人,副教授,博士,主要研究方向:空间信息集成技术、空间数据挖掘、虚拟地理环境、环境与自然资源遥感研究与开发。

$$s(A \Rightarrow B) = P(A \cap B)$$

$$c(A \Rightarrow B) = P(B | A)$$

给定一个数据库 D , 挖掘关联规则问题就是在数据库 D 中找出满足用户指定的最小支持度 (\min_sup) 和最小可信度 (\min_conf) 的所有规则。挖掘任务可分为两个子问题: 首先, 找出事务数据库中所有支持度大于最小支持度的频繁项集; 然后, 在频繁项集中产生所有大于等于最小可信度的关联规则。相对来说, 第二个子问题比较容易, 目前大多数研究主要集中于第一个子问题。

1.2 位集合构建算法

通过例子来说明如何构建位串集合, 其算法主要分两步。假定事务数据库为 TDB 如表 1, 最小支持阈值为 3。

扫描数据库 TDB, 根据属性列的个数生成相应的位集合, 对每条交易进行以下操作:

每个位集合的相应位置如果为 1, 则设置为 1, 默认为 0。当检查完所有记录后, 就生成 5 个 1-项集位集合。如 $A = \{1, 0, 1, 1, 1, 0\}$, $C = \{1, 0, 0, 1, 0, 0\}$, 依此类推。

把整个数据库存储在 N 个 (N 为 1-项集数目) 位集合中, 由于位集合所占的存储空间小, 大大地压缩数据库的大小。对数据库进行扫描后, 将事务数据库分割成 N 个位集合, 处理结果如表 2 所示。

2 MBSA 算法实现

构建完位集合, 就可以完全不参考原数据库, 通过对位集合进行简单的位统计就能挖掘频繁项集。位统计就是对每个位置进行位 1 的统计。

首先, 对 1-项集的位集合位置进行统计, 从表 3 中可以看出, 频繁集为 ($A:4$), ($C:2$), ($D:4$), ($E:4$), ($F:5$), 建立一个频繁项集 $TreeMap$ 数组的数据结构, 将字段名称或者字段名称组合作为映射的关键字, 映射内容就是该字段或者字段组合在数据库中出现的位集合。

表 1 示例数据库 TDB

ID	A	C	D	E	F
1	1	1	0	1	1
2	0	0	1	0	0
3	1	0	0	1	1
4	1	1	0	0	1
5	1	0	1	1	0
6	0	0	1	1	0

表 2 构建 1-项集的位集合

1-Item	Bitset
A	{1, 0, 1, 1, 1, 0}
C	{1, 0, 0, 1, 0, 0}
D	{0, 1, 0, 0, 1, 1}
E	{1, 0, 1, 0, 1, 1}
F	{1, 0, 1, 1, 0, 0}

表 3 1-项集的位集合位统计表

ID	Item					
	A	C	D	E	F	
1	1	1	0	1	1	
2	0	0	1	0	0	
3	1	0	0	1	1	
4	1	1	0	0	1	
5	1	0	1	1	0	
6	0	0	1	1	0	
Bit Count	4	2	3	4	3	

接着对满足支持度的 1-频繁项集进行位集合的逻辑“与”操作, 如果组合满足最小支持度, 则加入 2-频繁项集中。依此类推, 对 k -频繁项集进行上面的操作直到找不到满足最小支持度的频繁项集。

上述挖掘算法能挖掘出所有频繁项集, 并且使用压缩技术存储的位集合, 对它们使用位集合的逻辑“与”操作, 可以不必反复扫描数据库。

MBSA 算法描述如下:

Algorithm: MBSA algorithm

Input: Database D ; minimum support threshold \min_sup ;

minimum confidence threshold \min_conf

Output: The complete set of frequent itemsets

Method:

/* 扫描数据库, 构建所有位集合 */

For($i = 1$; $i \leq \text{rowcount}$; $i++$)

/* rowcount: 事务数据库中交易数量 */

For($j = 1$; $j \leq N$; $j++$) /* N : 事务数据库中属性列数量 */

if 第 j 列的第 i 行为 1

bit[j] 的第 i 个位置设为 1;

else

bit[j] 的第 i 个位置设为 0;

end

/* 生成频繁项集 */

Input: 频繁项目集 L_k

Output: L_{k+1}

Begin:

for $i = 1$ to ($|L_k| - 1$) {

for $j = i + 1$ to $|L_k|$ {

join($L_k[i]$, $L_k[j]$)

if bitset(j) 位统计 $\geq \min_sup$

$L_{k+1} = L_{k+1} \cup \text{bitset}(i, j)$;

}

}

/* 输出完整的频繁项集 */

Generate pattern from L_1 to L_k

end

3 算法比较分析

3.1 算法比较

为了验证算法 MBSA 算法的性能, 针对同样的数据库, 在相同的硬件和软件环境下, 采用 Apriori 算法和 MBSA 算法进行测试比较。在相同的支持度和可信度下, 两个算法进行执行时间比较。测试程序由 Jbuilder 9.0 编制, 测试在 P4 1.6 GHz, 256M 内存的微机上进行。测试数据记录数 5 000 条。

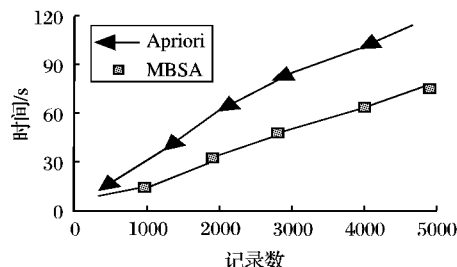


图 1 不同记录数下 MBSA 算法与 Apriori 算法执行时间比较

3.2 算法优势分析

与其他挖掘频繁项集算法相比, MBSA 算法的挖掘效率体现在下面几个方面:

1) 只需对数据库进行一次扫描, 无需像 Apriori 算法每次产生 k -项集扫描一次数据库;

2) MBSA 算法不产生候选项集因而也就无需对候选集进行修剪;

3) 不需要判断要连接的两个长度为 k 的项目的 $k-1$ 项

是否相同,可以直接进行拼接;

4) MBSA 算法仅需要有限的内存空间。与其他频繁模式增长方法如 FP-Tree 相比,它能全部装入内存。尽管较频繁的位集合的长度比较长,Java 语言中的 BitSet 类以简练和低内存消耗著称,可以全部装入内存。

4 MBSA 算法在遥感图像数据挖掘中的应用

4.1 数据源

遥感图像根据传感器不同分为诸多类型,如美国陆地资源卫星 TM、法国 SPOT、气象卫星 AVHRR 等。本文以 TM 的遥感图像为例开展遥感图像数据挖掘研究,TM 包含了 7 个波段,即 B (Blue), G (Green), R (Red), NIR (Reflect-Infrared), MIR (Mid-Infrared), TIR (Thermal-Infrared) 及 2 个 MIR (Mid-Infrared),每个波段的辐射分辨率范围均为 8 比特(像元取值 0~255)。本文选择实验区农作物收割前的 TM 遥感图像(数据来源见文献[10]),与之相对应的同一地区产量图进行关联规则挖掘。产量图是实际产量的 8 比特的灰度图,栅格化

后取值范围为 0~255。

本文采用 TM 遥感图像的 3 个可见光波段 B (Blue), G (Green) 和 R (Red), 而产量图只包含一个属性产量 Y (Yield)。从图像中提取四个属性 B (Blue), G (Green), R (Red) 和 Y (Yield), 全部属性都是量化数据。图像经过 3×3 重采样,图像大小为 200×200 像元,利用遥感图像处理软件 ERDAS IMAGINE 8.7 导出 4 个属性表。

4.2 属性分割

遥感图像通常包含的像元是以百万为单位的,如果得到很具体的规则形如 B=80, G=90, R=110→Y=150, 将导致规则的冗余并且没有意义。对属性的分割既不能太大,也不能太小。分割过大就会隐藏有意义的规则;如果分割过小会导致缺乏足够的支持度,丢失一些有用的规则。本文利用遥感图像处理软件 ERDAS IMAGINE 8.7 的直方图工具观察各个属性直方图范围,将四个属性维进行离散化,可以避免分割过大或者过小,表 4 为四个属性值的划分结果。

表 4 四个属性值的划分结果

Red			Green			Blue			Yield		
R-low	R-mid	R-high	G-low	G-mid	G-high	B-low	B-mid	B-high	Y-low	Y-mid	Y-high
[0, 25]	[26, 50]	[51, 255]	[0, 85]	[86, 170]	[171, 255]	[0, 40]	[41, 80]	[81, 255]	[0, 100]	[101, 180]	[181, 255]

属性分割完成后,得到 12 个原始项,分别是 {B-low (Blue 波段小), B-mid (Blue 波段中等), B-high (Blue 波段大), G-low (Green 波段小), ……

4.3 图像数据的布尔转换

关联规则算法只能处理布尔型的数据,挖掘开始前,需将遥感图像转化为事务数据库的形式。事务数据一般由事务的标识符和事务项集组成,本研究将遥感图像中的每个像元作为一次事务,不同波段相同像元的分割的属性就构成事务的项集。与真正的事务数据唯一的不同的是,每次交易项集的数目是相同的。

4.4 关联规则的提取及其分析

满足最小支持度和最小可信度的规则为强关联规则,根据领域知识可知,可见光 Blue, Green, Red 为参数数据,而 Yield 为结果数据,所以形如 {Blue, Green, Red} => Yield 的规则才是有意义的,其他的规则不用提取。实验中设定支持度为 10%,可信度为 40%,由于篇幅有限,仅列出下面的强关联规则如表 5。

表 5 关联规则(样本)

ID	Rules
1	R-low => Y-high, 支持度为 29.46%, 可信度为 73.65%
...	...
20	R-low, G-high => Y-high, 支持度 20.03%, 可信度为 72.29%
...	...
93	R-low, G-high, B-mid => Y-high, 支持度为 19.04%, 可信度为 72.28%

由表 5 可以看出,当 Red[0,25], Green[171,255], Blue[41,80]时,产量值[181,255],即产量高。这种对应关系表明可以通过调整土壤的湿度、施肥量等来改变农作物的长势,

提高农作物产量,也可以通过遥感进行农作物估产。

参考文献:

- [1] 马超飞, 刘建强. 遥感图像多维量化关联规则挖掘[J]. 遥感技术与应用, 2003, 18(4): 243-247.
- [2] AGRAWAL R, SRIKANT R. Fast Algorithms for Mining Association Rules[A]. Proceeding of the 20th International Conference on Very Large Databases[C]. Santiago, Chile, 1994. 487-499.
- [3] AGRAWAL R, IMIELINSKI T, SWAMI A. Mining association rules between sets of items in large databases[A]. Proceedings of the ACM SIGMOD Conference on Management of Data[C], 1993. 207-216.
- [4] STRIKANT R, AGRAWAL R. Mining Generalized Association Rules[A]. Proceeding of the 21st VLDB Conference[C]. Zurich, Switzerland, 1995.
- [5] ANTHONY K, TUNG H, LU HJ, et al. Efficient Mining of Inter-transaction Association Rules[J]. IEEE Transactions on Knowledge and Data Engineering, 2003, 15(1).
- [6] 颜雪松, 蔡之华. 一种基于 Apriori 的高效关联规则挖掘算法的研究[J]. 计算机工程与应用, 2002, 38(10): 209-211.
- [7] PARK JS, CHEN M, YU PS. Using a Hash-based Method with Transaction Trimming for Mining Association Rules[J]. IEEE Transactions on Knowledge and Data Engineering, 1997, 19(5).
- [8] BRIN S, MOTWANI R, ULMAN J, et al. Dynamic Itemset Counting and Implication Rules for Market Basket Data[A]. ACM SIGMOD Conference on Management of Data[C], 1997.
- [9] HAN J, PEI J, YIN Y. Mining Frequent patterns without candidate generation[A]. Proceedings of International Conference on Management of Data (SIGMOD'00)[C], 2000. 1-12.
- [10] TM image DataSets website[EB/OL]. <http://midas10.cs.ndsu.nodak.edu/data/image>, 2003-10.
- [11] ZAKI M, PARTHASRATHY S, OGIHARA M. New Algorithms for Fast Discovery of Association Rules[A]. 7th International Workshop Research Issues in Data Engineering[C], 1997.