

◎图形、图像、模式识别◎

图像纹理联合关联规则挖掘研究

王佐成^{1,2}, 张飞舟¹, 薛丽霞²WANG Zuo-cheng^{1,2}, ZHANG Fei-zhou¹, XUE Li-xia²

1. 北京大学 遥感与地理信息系统研究所, 北京 100871

2. 重庆邮电大学 软件学院, 重庆 400065

1. Institute of Remote Sensing and Geographic Information System, Peking University, Beijing 100871, China

2. Software Institute, Chongqing University of Posts and Telecommunications, Chongqing 400065, China

E-mail: zhangfz@pku.edu.cn

WANG Zuo-cheng, ZHANG Fei-zhou, XUE Li-xia. Research on image combined texture association rules mining. *Computer Engineering and Applications*, 2009, 45(26): 147-149.

Abstract: The representation method of texture features by image combined texture association rules and the algorithm of mining image combined texture association rules are proposed. Based on the concept of image combined texture association rule, by noise reduction and data mining preprocessing, multi-dimension image combined texture association rules data mining based on mask of texture image counting can be accomplished. Experiments testify that image combined texture association rules can represent the texture features perfectly and texture image segmentation can be achieved.

Key words: image texture; data mining; association rule; combined texture association rules

摘要: 提出以纹理联合关联规则来表达图像纹理特征以及挖掘纹理联合关联规则的算法。在纹理关联规则定义基础上, 通过图像降噪预处理和数据挖掘预处理, 采用模板统计挖掘方法挖掘低维和高维图像纹理联合关联规则。实验表明联合关联规则能够较好地表达图像纹理特征, 可以据此进行纹理分割。

关键词: 图像纹理; 数据挖掘; 关联规则; 联合关联规则

DOI: 10.3778/j.issn.1002-8331.2009.26.043 文章编号: 1002-8331(2009)26-0147-03 文献标识码: A 中图分类号: TP391

1 引言

从图像中像素的光谱特征来看, 构成纹理图像的各个像素、各个纹理基元之间都具有关联关系, 并且这些关联关系具有某种程度的重复性, 这是构成纹理的基础, 也是关联规则挖掘能够用于图像的前提。目前已经提出许多图像关联规则挖掘算法, Ordonez 提出了一种不依赖于特定应用领域知识的通用图像挖掘方法^[1]。该算法只能产生图像对象共同出现的关联规则, 仅适用于相同类型且只包含少量对象的图像。Osmar 提出大型图像数据库的空间特征关联规则挖掘算法^[2]。Ishwar 通过对图像的低层可视特征和高层概念间进行关联规则挖掘以获取对大型图像集的自动注释^[3-4]。在这些关联规则挖掘算法中, 都试图从图像中挖掘出一维或者多维频繁模式。然而, 由于图像的随机和模糊性的存在, 仅仅一种频繁模式难以准确描述某一纹理图像, 因此, 采用联合关联规则来对纹理进行描述, 并且提出联合关联规则挖掘算法。

2 图像关联规则定义

图像关联规则以可视特征、图像对象、对象空间关系作为特征表示图像, 对应于事务数据库中的项, 图像中包含的对象是可以重复的。另外, 图像具有不同的分辨率, 以及图像中关联关系比事务数据关联关系复杂, 包含不同层次特征间的关联。一方面, 可以把每一个图像看作一个事务, 从中找出不同图像间出现频率高的模式。此时, 需要考虑图像中每个对象的特征, 如颜色、形状、纹理、关键字和空间位置等可能存在大量的关联, 多个重复出现对象的重复模式以及图像对象间存在的重要空间关系等。另一方面, 如果图像数据挖掘深入到像素级, 则需要将一个像素和其邻域看作一个事务, 从中找出在图像中重复出现的模式。在纹理图像中, 这种模式实际上就是纹理基元。纹理基元有大小之分, 这就要求在多个层次上多分辨率情况下进行挖掘。这些使得图像数据库中的关联规则与传统事务数据库中的关联规则在挖掘方法和表现形式方面都有较大区别, 图像

基金项目: 重庆市教育委员会科学技术研究项目(Chongqing Municipal Education Commission Science Research Program, No.KJ080521)。

作者简介: 王佐成(1973-), 男, 博士后, 副教授, 主要研究方向为图像处理、软件工程和地理信息系统; 张飞舟(1966-), 男, 副教授, 主要研究方向为智能控制等; 薛丽霞(1976-), 女, 博士, 副教授, 主要研究方向为遥感图像处理与地理信息工程。

收稿日期: 2009-03-18

修回日期: 2009-04-20

关联规则定义也变得更复杂。

对于图像关联规则定义有较多研究^[5-6],均参考事务数据库中的关联规则进行。定义了根像素、项、项集、事务、关联规则。一条关联规则表达了图像的局部结构,形式如下:

$$(X_1, Y_1, I_1) \wedge \cdots \wedge (X_m, Y_m, I_m) \Rightarrow (X_{m+1}, Y_{m+1}, I_{m+1}) \wedge \cdots \wedge (X_{m+n}, Y_{m+n}, I_{m+n})(s\%, c\%) \quad (1)$$

式中, (X_m, Y_m, I_m) 表示 (X_m, Y_m) 处像素的灰度特征为 I_m , $s\%$ 和 $c\%$ 分别为支持度和置信度。

例如下面这条关联规则表示了二值图像中,一个像素宽的垂直条带的右边通常为一个像素宽的白色条带。

$$(0, 1, 1) \wedge (0, 0, 1) \wedge (0, -1, 1) \Rightarrow (1, 0, 0) \wedge (1, 1, 0) \wedge (1, -1, 0)(s\%, c\%) \quad (2)$$

支持度和置信度表明了这种情况出现的可能性。

从图像关联规则定义中,可以看出,根据该方法所形成的数据挖掘搜索空间十分庞大。对图像来说,矩阵的表现形式远比事务型表现形式高效,因此,相对于将图像进行关系模式化然后挖掘而言,将图像关系模式化融入挖掘算法然后对图像进行挖掘才是一个可行的研究方向。提出一种先通过图像降阶,突出纹理特征,然后从 2×2 模板开始挖掘 4 维至多维关联规则,由于纹理规则前件和后件的对等关系,只通过支持度来作为规则兴趣度量。

3 图像数据预处理

3.1 图像降噪预处理

由于实际数据常常存在不完整性、噪声和不一致性,预处理就变得很重要。有两种数据预处理技术用于图像数据的数据挖掘中,即数据清洗和数据变换^[7-8]。该文采用常用的方法进行图像降噪处理,包括两步:(1)图像去噪。经过去噪处理后,可去掉图像中的大多数背景信息和噪声。(2)图像增强。在图像的生成、传输或变换过程中,由于多种因素的影响,总会造成图像质量的下降,图像增强的目的是采用一系列技术改善图像的效果或将图像转换成更适合处理的形式^[9]。采用直方图均衡化处理方法实现图像的增强。同时,在此之前应该完成去噪处理,否则噪声也会同时得到增强。

3.2 图像数据挖掘预处理

针对纹理特征描述情况,构造关联规则必须考虑到一条或者若干条关联规则能够表达纹理特征。在规则纹理情况下,可以考虑用一条或者若干条关联规则表达纹理基元。然而,当纹理基元太大的情况下,规则的维数将会很大,这样使得挖掘搜索空间大大复杂化。因此必须考虑如何减少能够表达纹理的关联规则搜索空间。

在关联规则定义中,一个具有 G 种灰度值的 $n \times n$ 邻域中,可能产生 $n^2 G$ 个不同的项。可以考虑两种方法来减小项的数量:一种是减少灰度值 G 的阶数;另一种是缩小能够表达纹理基元的 $n \times n$ 的邻域范围。

经过图像降阶能够抽取纹理特征,从而缩小纹理关联规则挖掘搜索空间。然而,降阶虽然能够降低图像关系模式化中形成的项的数量,但是对表达纹理的关联规则维数却不能降低。

因此,提出一种四叉树抽取纹理特征的方法来降低表达纹理的关联规则的维数。

四叉树抽取纹理特征的方法是表示栅格图像的逆过程。该方法分为三步:(1)对图像进行扫描,获取最小纹理基元表达需要的像素邻域 $n \times n$ 。如果基元不是方阵,可以通过扩展的方式得到方阵。(2)将图像纹理基元方阵四等分,检查其格网灰度值,求取格网的像素点灰度平均值,如果像素点的灰度值与平均值的差异超出预先确定的阈值,则该格网需要继续分割。直到分割成的格网内的像素点灰度值与平均值差值小于阈值。(3)表示纹理基元。对每个最后分割出来的格网取其灰度平均值作为其灰度值并构造四叉树。针对模糊和随机纹理挖掘,有很多时候不能从图像中直接发掘基元,此时必须根据图像降阶的情况组合多模板采用统计挖掘方法来获得感兴趣的基元模式。

4 图像纹理联合关联规则挖掘

4.1 基于降阶图像的模板统计挖掘方法

可以构造 2×2 或者 3×3 甚至更高维的模板来进行纹理挖掘,模板维数的选择可以根据图像纹理描述符纹理二阶矩 W_{ij} 所描述的图像均匀性或平滑性并结合上章论述的四叉树对纹理的抽象情况来决定。当纹理二阶矩相对较小,并且四叉树对纹理的抽象处理中进行了较多平滑处理时,可以选择较小模板,反之选择较大模板进行统计挖掘。模板的大小对挖掘效率的影响很大,复杂度为 $O(n^2)$ 。因此在实际挖掘中,尽可能选用较小模板,即使对较大纹理基元进行了分割,通过下文提出的联合多维关联规则同样也能够表达纹理基元和纹理特征。

首先取 2×2 模板,模板中灰度值由 0 至 3 组成,可见总模板数为 $4 \times 4 \times 4 \times 4 = 256$ 个。在用模板进行图像扫描时,可以采用非重叠窗口的方式进行,非重叠窗口是指:作边长为 $M=2$ 的方形窗口,它们互相不重叠。也可以采用重叠窗口的方式进行,重叠窗口的选取是指:以 i 行 j 列像素为中心的,边长为 M 的窗口,其下一个窗口是以 i 行 $j+1$ 列像素为中心,边长为 M 的窗口,它们有互相重叠的部分。扫描时,图像模板窗口从图像的左上角开始进行。

对每个模板窗口在图像中的匹配情况进行计数,在计数中,采用一种高效的计数方法在程序中实现:由于采用的是 2×2 模板窗口,首先定义一个四维数组,其下标既作为模板种类的标志,又作为在扫描图像的过程中模板计数的指示。这样避免了窗口匹配时的大量的关系运算,大大提高了效率。

4.2 高维联合关联规则挖掘

通过联合关联规则对图像纹理表达进行实验分析,可以看出,四维联合关联规则能够较好表达图像纹理。但当图像纹理基元较大时,挖掘模板大小会明显影响联合关联规则对纹理的表达精度。根据 Apriori 挖掘算法的思想,借助 Apriori 性质,通过连接步和挖掘统计剪枝步来获得更高维联合关联规则。

由于纹理数据挖掘的特殊性,纹理关联规则的维的扩展不能象事务关联规则一样一项一项地扩展。在纹理挖掘中,是从 2×2 的模板开始挖掘获取四项集,在进行规则维扩展时,采取

模板连接的方式,下面以例说明。两个频繁四项集模板如图 1(a)和(b)所示,其模板 a 自身连接方式如图 2 所示,模板 a 和 b 互相连接的方式以此类推。



图 1 实例模板

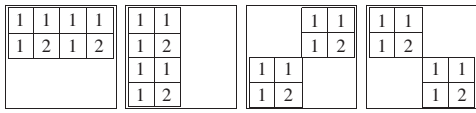


图 2 模板 a 自身连接图

按照这种模板连接方式从四项集得到 8 项集 C_8 , 还可以扩展到 16 项集 C_{16} 、32 项集 C_{32} 甚至更高维项集。得到候选集后,进行扫描图像,根据支持度阈值,可以完成剪枝步,得到频繁 8 项集 L_8 或者 L_{16}, L_{32} 。在进行剪枝步时,仍然采用 Apriori 性质压缩候选项集,提高效率。整个算法描述如下:

算法:采用 Apriori 性质完成模板 mask 连接和剪枝步。

输入:纹理图像;最小支持度 min_sup 。

输出:经过压缩的候选集。

方法:

- (1) procedure apriori(L_{k-1}, min_sup)
- (2) for each itemset $l_1 \in L_{k-1}$
- (3) for each itemset $l_2 \in L_{k-1}$
- (4) $\{c=l_1 \cup l_2 \mid //mask\ join\ step, generate\ candidates$
- (5) if has_infrequent_submask(c, L_{k-1}) then
- (6) delete c ;
- (7) else add c to C_k ;
- (8) return C_k ;

算法中,第(4)步完成模板连接,根据方位来完成连接。第(5)至(7)步完成剪枝步,根据频繁模板 L_{k-1} 完成判断是否剪枝。

5 实验结果与分析

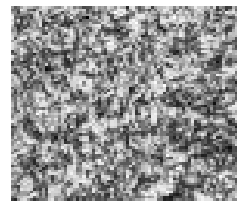
对各种规则和不规则纹理图像进行实验,图 3(a)是原始图像,通过基于降阶图像的模板统计挖掘方法,得到这 256 个模板出现的频数,如图 3(b)所示。图 3(c)是对(b)中数据进行曲线模拟。

经过挖掘,总共四项集共 5 476 项,以 0.02 为兴趣度阈值,得到 13 种频繁模式,频繁模式和出现频数见表 1。

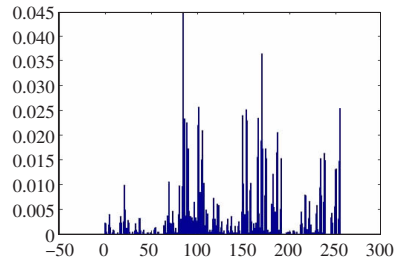
这 13 条关联规则实际上是图像中频繁出现的纹理情况,它们共同表现出纹理的特征,同时用这 13 条关联规则构成联合关联规则来表现纹理特征。

对图 3(a)进行高维联合关联规则挖掘,挖掘 L_8 ,取 $min_sup=0.02$,得到 C_8 如表 2。

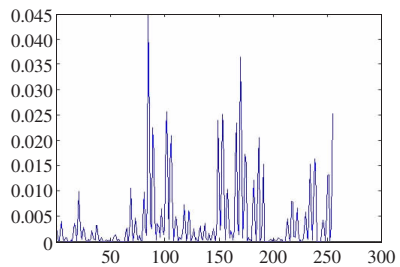
其中,规则描述中的第三项,如(I,II)表示模板方位关系为 1,2 象限方位。通过对相似和相异的纹理图像挖掘,其频繁模式拟合的曲线有明显区别,比较过程在此不再赘述。实验结果表明挖掘出的联合关联规则能够很好表达各种纹理。



(a)原始图像



(b)模板频数统计



(c)对(b)的曲线模拟

图 3 模板频数统计表示

表 1 频繁模式列表

序号	1	2	3	4	5	6	7
频繁模式	1,1,1,1	1,1,1,2	1,1,2,1	1,2,1,1	1,2,1,2	1,2,2,2	2,1,1,1
支持度	0.044 7	0.023 2	0.022 5	0.021 9	0.025 6	0.020 8	0.023 9
序号	8	9	10	11	12	13	
频繁模式	2,1,2,1	2,1,2,2	2,2,1,2	2,2,2,2	2,3,2,3	3,3,3,3	
支持度	0.025 0	0.022 8	0.023 4	0.036 3	0.020 5	0.025 4	

表 2 L_8 列表

序号	关联规则
1	(1,1,1,1)→(1,1,1,1)(I,II)(3.275 1%)
2	(1,1,1,1)→(1,1,1,2)(I,IV)(2.054 6%)
3	(1,1,1,2)→(1,2,1,1)(II,III)(2.023 4%)

6 结论与展望

论文针对图像纹理数据挖掘的关键问题,如图像纹理关联规则定义、纹理图像预处理、纹理图像挖掘模型进行了研究,提出采用联合关联规则表达纹理并验证其可行性。最后使用统计挖掘方法对纹理图像中频繁模式进行挖掘,得到能够表达图像纹理的联合关联规则,并通过实验验证。该挖掘方法可在不同的分辨率下对纹理图像进行关联规则挖掘。即可先在粗分辨率下进行图像挖掘,对发现的频繁模式再进行细分辨率下的挖掘,从而减少算法复杂度。

参考文献:

[1] Ordonez C, Omiecinski E. Discovering association rules based on image content[C]// Proceedings of the IEEE Advances in Digital Libraries Conference, Baltimore, 1999, 5: 38-49.