基于粗糙集的图像分类挖掘

李龙澍[1,2],邹 武[1,2]
(1.安徽大学 计算智能与信号处理教育部重点实验室,安徽 合肥 230039;
2.安徽大学 计算机科学与技术学院,安徽 合肥 230039)

摘 要:随着图像数据库的不断增长,传统方法的对基于内容的图像数据的分类挖掘越来越显得不足,使用粗糙集方法,利用先验知识可以提高图像分类的准确率。文中从图像的色彩特征的角度出发,通过多种方法尽可能多地提取图像的色彩特征信息,同时按照数据挖掘的一般步骤提出具体应用于图像分类的挖掘算法,且通过实验证明该算法的可行性和优越性。

关键词:CBIR;粗糙集;数据挖掘;直方图
中图分类号:TP391.41 **文献标识码**:A **文章编号**:1673-629X(2009)04-0143-03

Image Classification Mining Based on Rough Set

LI Long-shu[1,2],ZOU Wu[1,2]
(1.Ministry of Educational Key Laboratory of Intelligent Computing & Signal Processing,
Anhui University, Hefei 230039,China;
2.Computer Science and Technology School, Anhui University, Hefei 230039,China)

Abstract:With the growing image database, the traditional methods of content-based image classification data mining is increasingly inadequate. The use of rough set methods and prior knowledge can improve the accuracy of image classification. This article from the color characteristics of the image point of view, through a variety of methods to extract as much as possible image of the color feature information and data mining in accordance with the general steps to make specific request for images of the mining algorithms, and through experiments prove the feasibility of the algorithm and superiority.

Key words:CBIR;rough set;data mining;histogram

0 引 言

基于内容的图像检索(CBIR)比基于文本的图像检索发展要晚得多,而且技术也没有基于文本的成熟,发展缓慢。传统的直方图方法是最常用基于内容的图像检索方法之一。直方图算法简单,效果显著,计算效率高,对镜头位置不敏感,对图像变形、旋转、尺度具有不变性,针对颜色直方图,通常采用直方图相交法[1]来寻找与查询图像最相似的也就是距离最近的图像。由于直方图距离比较方法使用某一灰度下的像素的总量这样的一相单一的统计指标,两两比较同一灰度,没有考虑每一个灰度值在图像中形成的具体形状与空间分布因素,而是将其视为一个单一整体,忽略内部差异,比较的粒度较粗糙,没有能够反映出同一色彩在空间分布的不同,因而造成空间信息的丢失。为了克服颜色直方图没有包含空间信息的问题,王小玲[2]提出了主要面积直方图法和平均面积直方图法,从一定的范围和程度上提高了图像检索分类的准确度。

但是应用局部特征检索某一类自然景物对象,以雪山与瀑布为例,由于颜色相似,误识率就比较高。应用风景图像的先验知识可以提高对风景图像检索的准确率,文中采用粗糙集方法对图像进行分类,突破了传统模板方法的局限性,不需要预先定义模板,特征提取简单易行。系统具有学习能力,能够自动发现用于分类的潜在知识;其次由于粗糙集具有处理模糊与不确定性知识的能力,因此系统能够处理外观差异较大的风景图像,而且有抗噪声能力。

1 相关概念

粗糙集作为一种处理不精确、不确定与不完全数

收稿日期:2008-07-27
基金项目:安徽省自然科学基金项目(050420204);安徽省高校拔尖人才基金项目(05025102)
作者简介:李龙澍(1956-),男,教授,博士生导师,研究方向为智能软件和知识工程。

据的新的数学理论，最初是由波兰数学家 Z. Pawlak 于 1982 年提出的[1]。其主要思想是，在保持信息系统分类能力不变的前提下，通过属性约简，导出问题的决策或分类规则。目前，粗糙集理论已经被应用于机器学习与知识发现、故障诊断、控制算法获取、数据挖掘等各种应用领域，并取得了很大的成功[2~4]。

定义 1 信息系统：称 4 元有序组 $S = < U, A, V, f)$ 为信息系统，其中 U 为所考虑对象的非空有限集合，称为论域；A 为属性的非空有限集合；V 为属性 a 的值域；$f: U \times A \rightarrow V$ 是一个信息函数。

定义 2 置信度：规则的置信度即可信度 $= P(A/\boldsymbol{B}) = \frac{P(A,\boldsymbol{B})}{P(\boldsymbol{B})}$ 其中 A 表示某一规则的决策值，$\boldsymbol{B}$ 表示该规则的条件属性向量，表示在 $\boldsymbol{B}$ 的条件下产生决策 A 的条件概率，是对规则的准确度的衡量。

定义 3 支持度：支持度和置信度是一对概念，支持度 $P(A,\boldsymbol{B})$ 表示决策规则中条件属性 $\boldsymbol{B}$ 和决策属性 A 同时出现的概率，是规则的重要性的标志[5]。

颜色特征提取在基于内容的图像检索中重要性很高，且直方图方法在基于内容的图像检索中获得了广泛的应用，色彩、纹理、形状等特征都可以用直方图来表示。

定义 4 直方图：$H(I)$ 表示图像 I 的颜色等级 C_i 的像素的数目：

$$H(I) = \{N(I, C_i) \mid i = 1 \cdots n\}$$

定义 5 相似度：用来计算图像 Q 与图像 I 的颜色直方图的距离，通过计算的距离得知图像间的相似度。如果相似度达到规定的阈值就可以认为这两幅图像是相近的，在图像分类中可以归为一类。

$$d(H(I), H(Q)) = [\sum_{i=1}^{n} \mid H(I_i) - H(Q_i) \mid^{\lambda}]^{1/\lambda}$$

定义 6 主要面积直方图：主要面积直方图法中 $H * c(I)$ 由颜色等级 C_i 所形成不连通的最大面积构成[6,7]。

$$H * c(I) = \{\mathrm{Max}(N(I, C_i)) \mid i = 1, 2, \cdots, n\}$$

定义 7 平均面积直方图：平均面积直方图是利用每一种颜色在不同连通区域的面积特征建立起来的。

$$H * c(I) = \{\frac{N(I, C_i)}{D(I, C_i)} \mid i = 1, 2, \cdots, n\}$$

$N(I, C_i)$ 是颜色等级 C_i 的像素总和，$D(I, C_i)$ 是 C_i 的不连通区域的数目。$H * c(I)$ 即表示颜色空间 C 的颜色等级的平均像素数目。

2 图像分类的粗糙集算法

基于内容的图像分类是数据挖掘的应用之一，其算法从总体上按照数据挖掘算法的一般步骤如下：

(1)数据选择。一定比例等分训练数据集 $U1$ 和测试数据集 $U2$，一般按 2:1 或 3:1 等分。

(2)数据集成。分析训练集图像，提取图像的特征属性，图像内容特征集构成条件属性集，类别属性作为决策属性构成决策表。

(3)数据预处理。进行属性约简和属性值约简。

(4)数据分类知识生成。提取决策规则，产生规则的可信度和支持度。

(5)数据测试。生成测试图像的内容特征，根据决策规则，符合规则的可信度最大的分类即为图像的分类，如果多则规则的可信度相同，即选择支持度最大的规则产生的分类决策。

图像特征的提取是基于内容的图像分类的关键性因素。传统的直方图法计算根本不能体现色彩的空间分布特征，但从色彩角度可以获取大量的信息。

王小玲提出的基于主要面积直方图与传统的直方图方法相反，它可以体现一定的空间特征，但却没有从全局考虑，没有考虑图像某一色彩信息量很大却分呈散开状的情况。

基于平均面积直方图在空间上从全局考虑，所以将三者统一考虑到条件属性集中，那么从色彩和空间分布得到的信息就比任何一种方法独立考虑要详细得多了。

3 实 验

风景图像分类问题一直是计算机视觉研究的开放性课题。风景图像的分类可以提供对图像内容的理解，因此是图像检索中一个非常值得研究的问题。应用传统的直方图的方法，以雪山与瀑布为例，由于颜色相似，误识率就比较高。应用风景图像的先验知识可以提高对风景图像检索的准确率。本节基于上节的图像分类的粗糙集算法对太阳、瀑布、山脉、花草四类风景图像分类进行研究[8]。

3.1 数据选择

将数据库图像 U（包含 1500 条数据）分为 2 个子集 $U1$ 和 $U2$，分别包含 1000 条训练集与 500 条测试集数据。

3.2 数据集成

图像的色彩的特征提取是图像分类的关键性因素。小面积色彩特征对图像分类的影响不大，所以应用将传统的直方图获得的第一、二主要色彩，主要面积直方图获得的第一、二主要面积色彩和平均面积直方图获得的第一、二平均色彩结合在一起得到图像的条件属性集。同时，可以考虑第一、二主要面积的位置增

加图像的空间特征信息[9]。

决策属性就是图像进行分类的类型如太阳、瀑布、山脉和花草等四类风景图像。数据集成结果见表 1(因为空间原因,决策关系呈纵向排列)。

表 1　图像分类信息系统表

	图像 1	图像 2	图像 3	图像 4	…
第一主要色彩	白	白	红	红	
第二主要色彩	绿	青	绿	白	
第一主面积色彩	白	白	红	红	
第二主面积色彩	绿	绿	白	白	
第一主面积位置	中	中	中	中	
第二主面积位置	左	右	左	左	
第一平均色彩	白	白	白	红	
第二平均色彩	绿	青	红	白	
分类决策	瀑布	瀑布	花草	太阳	

3.3　数据预处理

对数据进行清理和充实等预处理工作。对于基于粗糙集的图像分类主要是应用经典粗糙集理论的约简知识对图像数据先后进行属性约简和属性值约简。应用启发式属性重要性进行约简,结果显示表 1 的条件属性是最简约简。

对图像数据进行编码,数据库中字段的不同取值转换成数字形式将有利于搜索如表 2 对决策属性进行编码。

表 2　编码表

决策属性编码	编码含义
1	太阳
2	瀑布
3	山脉
4	花草

3.4　数据分类知识生成

将 1000 幅图片进行分类,计算可信度和支持度,将最小可信度的阈值定为 0.01,得到的分类知识有 51 条,部分结果如表 3 所示。

表 3　图像分类决策表

	规则 1	规则 2	规则 3	规则 4	规则 5	规则 6
第一主要色彩	白	白	白	红	红	…
第二主要色彩	绿	绿	绿	白	白	
第一主面积色彩	白	白	白	红	红	
第二主面积色彩	绿	绿	绿	白	白	
第一主面积位置	中	中	中	中	中	
第二主面积位置	左	左	左	左	左	
第一平均色彩	白	白	白	红	红	
第二平均色彩	绿	绿	绿	白	白	
分类决策	瀑布	山脉	太阳	花草	太阳	
可信度	0.96	0.03	0.01	0.60	0.44	
支持度	0.02	0.0006	0.0002	0.01	0.009	

3.5　数据测试与总结

将经过数据清洗和数据转换得到的测试图像的特征数据根据算法步骤 5 的基本原则,以规则的可信度为主要决策因素,以支持度为次要的决策因素。结果如表 4 所示。

表 4　图像分类测试表

类别(总体条数)	瀑布(125)	花草(125)	太阳(125)	山脉(125)
分类瀑布(条数)	瀑布(115)	瀑布(1)	瀑布(5)	瀑布(9)
分类花草(条数)	花草(0)	花草(120)	花草(6)	花草(3)
分类太阳(条数)	太阳(4)	太阳(2)	太阳(106)	太阳(6)
分类山脉(条数)	山脉(6)	山脉(2)	山脉(8)	山脉(107)
正确率	0.92	0.96	0.85	0.86

实验结果表明,在文中提取的特征基础之上,应用粗糙集方法对 4 类风景图像进行分类,取得了较好的检索效果,分类正确率达到 85% 以上。反馈式或增加式的粗糙集自动学习方法可以进一步提高检索系统的适应能力。

4　结束语

结果显示文中采用粗糙集方法对图像进行分类,突破了传统模板方法的局限性,可以处理分类过程中的模糊与不确定性知识,增强了对风景图像外观变化以及噪声的鲁棒性[10]。特征的提取方式上,将传统的直方图方法和主要面积以及平均面积直方图综合,且从最大影响分类色彩的角度出发,省略与分类关联不大的色彩,提高了算法的效率。

参考文献:

[1] Swain M J, Ballard D H. Color indexing[J]. International Journal of Computer Vision, 1991, 7(1): 11 - 32.

[2] Lin T Y, Cercone N. Rough Sets and Data Mining[M]. Boston: Kluwer Academic Publisher, 1997: 47 - 76.

[3] Moxon B. Defining Data Mining[M]. [s. l.]: Miller Freeman, Inc., 1996.

[4] 王小玲. 基于内容的图像检索技术研究[D]. 上海: 上海交通大学, 2006.

[5] 涂占新. 数据挖掘方法及其应用展望[J]. 中南财经政法大学学报, 2003, 2(2): 117 - 120.

[6] 阮秋琦. 数字图像处理学[M]. 北京: 电子工业出版社, 2001.

[7] 章毓晋. 图像处理和分析[M]. 北京: 清华大学出版社, 1999.

[8] 韩祯祥, 张　琦, 文福拴. 粗糙集理论及应用[J]. 信息与控制, 1998, 2: 37 - 44.

[9] 王　珏, 苗夺谦, 周育健. 关于 Rough Set 理论与应用的综述[J]. 模式识别与人工智能, 1996, 9: 337 - 344.

[10] 苗夺谦. Rough set 理论及其在机器学习中应用研究[D]. 北京: 中国科学院自动化研究所, 1997.