

基于属性论的无监督图像聚类

李根根, 冯嘉礼

(上海海事大学 信息工程学院, 上海 200135)

摘要: 提出了应用属性论来实现无监督图像聚类的方法。首先提取图像的底层特征, 再利用属性论方法中定性映射、转换程度函数理论对图像进行分类。由于一次聚类并未达到很好的效果, 因此采取了多次聚类的策略。实验结果表明, 本方法简单明了并且有效。

关键词: 属性论; 定性映射; 转换程度函数; 图像聚类; HSV

中图分类号: TP391

文献标识码: A

文章编号: 1674-7720(2012)02-0032-03

Unsupervised image clustering based on attribute method

Li Gengen, Feng Jiali

(College of Information Engineering, Shanghai Maritime University, Shanghai 200135, China)

Abstract: The paper proposes a method of using attribute to achieve image clustering. Firstly, image's features is extracted. Then using qualitative mapping and conversion degree function of attribute method to classify images. Because clustering image in one time do not have a good result, the paper chooses to cluster image three times. Experiments prove that this method is very simple, obvious and effective.

Key words: attribute method; qualitative mapping; conversion degree function; image clustering; HSV

随着多媒体技术的飞速发展以及图像获取与存储技术的提高, 图像数据库得到了快速发展。面对如此庞大的图像数据库, 各个领域的图像都非常多, 想要检索其中用户感兴趣的图像十分不容易。因此, 有必要对图像数据库中的图像进行聚类以减少图像检索的代价。俗话说“物以类聚”, 图像的聚类即是图像数据库中最为相似的部分划为一类, 使得类内的相似度尽可能高, 而类间的相似度尽可能低。

本文中的图像聚类采用图像的颜色以及轮廓两种特征。利用熟知的 HSV 模型来得到图像的颜色谱, 以此作为颜色特征。轮廓特征则利用属性论方法来获取。在提取了特征之后, 利用转换程度函数来完成聚类。

1 图像聚类

在聚类之前, 先来介绍本文所用到的一些理论知识。好的聚类结果离不开两个重要方面: 准确有效的图像特征以及合适的理论与方法。这两方面缺任何一个都有可能达不到预期的效果。

1.1 定性映射^[1-2]

众所周知, 事物有质和量的两种规定性, 质和量之间有质量互变规律: 只有超过度的范围的量变才导致质的变化, 而限于度的范围之内的量变不导致质变。这

一性质可用如下定理表示:

定理 1: 设 $a(u)$ 是事物 u 的一个属性, $x_0 \in [\alpha, \beta] \subseteq X$ 是 $a(u)$ 的一个量, $[\alpha, \beta]$ 是 x_0 所对应性质 $p(x_0, u)$ 的定性基准 (若 α 和 β 分别是性质 $p(u)$ 发生变化的两个关节点, 则称区间 $[\alpha, \beta]$ 为性质 $p(u)$ 的“度”或定性基准), $\xi = (\alpha + \beta) / 2$ 为 $[\alpha, \beta]$ 的球心, $r = (\beta - \alpha) / 2$ 为 $[\alpha, \beta]$ 的半径, 并设 $q(x, u)$ 是 x_0 变为 $x = x_0 + \Delta x$ 所对应的性质。若 $|x - \xi| < r$, 则有 $q(x, u) = p(x, u)$; 若 $|x - \xi| > r$, 则 $q(x, u) \neq p(x, u)$ 。

用逻辑语句则可表示为: if $x \in [\alpha, \beta]$, then $p(u)$, else $\neg p(u)$ 。其中, $\neg p(u)$ 表示非 $p(u)$, 即不具备性质 $p(u)$ 。同时, 若将定性基准 $[\alpha, \beta]$ 作为一个调控自变量 x 变化范围的参变量, 则最简性质命题判断可表达为下述映射形式:

定义 1: 设 $a(u)$ 是对象 u 的某个属性, $x \in X \subseteq R$ 是属性 $a(u)$ 的一个量值, $p(u) \in P_0$ 是属性 $a(u)$ 的某个性质, $[\alpha, \beta] \in \Gamma$ 是性质 $p(u)$ 的定性基准, 则称映射 $\tau: X \times \Gamma \rightarrow \{0, 1\} \times P_0$, 使得:

$$\tau(x, [\alpha, \beta]) = x \in [\alpha, \beta] = \begin{cases} p(u) & x \in [\alpha, \beta] \\ \neg p(u) & x \notin [\alpha, \beta] \end{cases} \quad (1)$$

式(1)为最简性质 $p(u)$ 的判断映射或定性映射。其中, \in 为问题: “ x 是否在 $[\alpha, \beta]$ 中?” 的算子, 或“ x 是否

图形、图像与多媒体

满足性质(命题) $p(u)$ 的定性基准?”的算子。由于对每一个属性 $a(u)$,输入的是 $a(u)$ 的一个量(特征)值 x ,输出是 $a(u)$ 的一个性质 $p(u)$ 或非 $p(u)$ 的 $q(u)$,因此,若强调式(1)是一个从量特征 x 到质特征 $p(u)$ 的一个转化过程,则符号 \in 又可以称为属性 $a(u)$ 的量——质特征转化算子,或(性)质特征抽取算子。

若用 $\tau_p(x,u)=\begin{cases} 1 & x \in [\alpha, \beta] \\ 0 & x \notin [\alpha, \beta] \end{cases}$ 表示性质命题 $p(u)$ 的真值,则式(1)可表示为:

$$\tau_p(x, [\alpha, \beta]) = x \in [\alpha, \beta] = \tau_p(x, u) \quad (2)$$

特别地,若用 $\tau_p(x)$ 代替 $\tau_p(x, u)$ 而不至引起误解或混淆,则式(2)可写为:

$$\tau(x, [\alpha, \beta]) = x \in [\alpha, \beta] = \tau_p(x) \quad (3)$$

若令 $\xi_i = (\beta_i + \alpha_i)/2$,因 (α_i, β_i) 是一个以 ξ_i 为球心,以 $\delta_i = (\beta_i - \alpha_i)/2$ 为半径的拓扑邻域 $N(\xi_i, \delta_i)$,则有:

$$\begin{cases} x \in (\alpha_i, \beta_i) = N(\xi_i, \delta_i) \Leftrightarrow |x - \xi_i| < \delta_i \\ x \notin (\alpha_i, \beta_i) = N(\xi_i, \delta_i) \Leftrightarrow |x - \xi_i| \geq \delta_i \end{cases} \quad (4)$$

于是,最简判断型定性映射式(2)可改写为 $\tau: X \times \Gamma \rightarrow \{0, 1\}i$,使得:

$$\tau(x, N(\xi_i, \delta_i)) = x \in N(\xi_i, \delta_i) = \tau_p(x) \quad (5)$$

其中, $\tau_p(x) = \begin{cases} 1, & |x - \xi_i| \leq \delta_i \\ 0, & |x - \xi_i| > \delta_i \end{cases}$ 。称式(5)是以邻域 $N(\xi_i, \delta_i)$ 为基准的性质 $p_i(x)$ 的定性映射。

1.2 转换程度函数^[3-6]

以成绩为例来讲述转换程度函数。比如 $x_1=60$ 分和 $x_2=75$ 分都是合格,但两者在对知识的掌握程度上却有所不同。尽管它们同属一个定性基准 $[60, 75]$,即: $x_1, x_2 \in [\alpha_i, \beta_i]$,经量—质特征转化后,它们对应的性质同属一个质特征类——合格,即: $p_i(x_1), p_i(x_2) \in p_i(u)$,但它们转化为 $p_i(u)$ 的程度 $\eta(p_i(x_1))$ 和 $\eta(p_i(x_2))$ 却会因 $x_1 \neq x_2$ 而产生很大的差异。

因此,需要引进一个能刻画这种差异的概念——转化程度函数来描绘这种因量不同而导致同质转化程度不同的差异。

质特征 $p_i(u)$ 的量 $x \in [\alpha_i, \beta_i]$ 越接近 $[\alpha_i, \beta_i]$ 的边缘,越容易发生质变。定性基准 $[\alpha_i, \beta_i]$ 的边界点 α_i 和 β_i 对应的两性质 $p_i(\alpha_i), p_i(\beta_i) \in p_i(u)$ 是 $p_i(u)$ 中最容易向其他质特征(或性质) $p_k(u)$ 转变的性质,故可称之为类 $p_i(u)$ 中的临界性质;而中点 ξ_i 对应的性质 $p_i(\xi_i)$ 则是 $p_i(u)$ 中最稳定、也最不容易转变为其他质特征、最能体现其质特征类 $p_i(u)$ 本质的性质,因此可设 $p_i(\xi_i)$ 为 $p_i(u)$ 的本征性质,并称 ξ_i 为 $p_i(u)$ 的本征点。若称 $k_1(x) = \frac{p_i(\xi_i) - p_i(x)}{p_i(\xi_i) - p_i(\alpha_i)}$ (其中 $x < \xi_i$)和 $k_2(x) = \frac{p_i(x) - p_i(\xi_i)}{p_i(\beta_i) - p_i(\xi_i)}$ (其中 $x > \xi_i$)为 x 对应的性质 $p_i(x)$ 偏离 $p_i(\xi_i)$ 的程度,则可称式(6)为 $p_i(x)$ 接近本征性质 $p_i(\xi_i)$

的程度,或 $p_i(x)$ 体现其质特征类 $p_i(u)$ (或 $p_i(\xi_i)$)的程度,或 x 转化为质特征类 $p_i(u)$ 的转化程度。

$$\eta(x) = \begin{cases} -(1 - k_1(x)) & x < \xi_i \\ 1 - k_2(x) & x > \xi_i \end{cases} \quad (6)$$

$k_1(x), k_2(x) \in [0, 1]$,故 $\eta_i(x) \in [-1, 1]$,则转化程度函数 $\eta_i(x)$ 可定义为:

定义2:称映射 $\eta: X \times \Gamma \rightarrow [-1, 1]i$ 为 $p_i(x)$ 体现其质特征类 $p_i(\xi_i)$ 的程度函数,如果对 $\forall (x, N(\xi_i, \delta_i)) \in X \times \Gamma$, $\exists \eta_i(x) \in [-1, 1]i$,得:

$$\eta(x, \xi_i, \delta_i) = |x - \xi_i| \leq \delta_i = \eta_i(x) \quad (7)$$

式(7)说明转换程度函数导致了定性基准的边界模糊化。

如果将转化程度函数 $\eta(x)$ 看作一个值域扩充为 $[-1, 1]$ 的定性映射,则定性映射 $\tau_i(x)$ 就可看作是一个仅取 $\{0, 1\}$ 的转化程度函数。反之,将式(7)的值域 $[-1, 1]$ 限制为 $\{0, 1\}$,并令所有不等于0的值 $\eta_i(x)$ 等于1,那么转化程度函数就退化为定性映射。由此可见,程度函数 $\eta_i(x)$ 不仅是 $\tau_i(x)$ 诱导出来的,而且它们之间还存在着非常深刻的本质联系,因此有时又将 $\eta_i(x)$ 称为定性程度函数。

由于 $\tau_i(x)$ 的值域 $\{0, 1\}$ 仅有两个值,因此 $\tau_i(x)$ 只能表达基于属性量—质特征转化的定性判断中的逻辑关系。而转化程度函数 $\eta_i(x)$ 的值域已扩充为区间 $[-1, 1]$,故能对该量—质转化中的程度差异进行数学表达。

由此可见,转化程度函数 $\eta_i(x)$ 的数学本质是将 $|x - \xi_i|$ 与 δ_i 进行比较, $\frac{|x - \xi_i|}{\delta_i}$ 比值的大小反映量 x 所对应的性质 $p_i(x)$ 与本征性质 $p_i(\xi_i)$ 的差异。

1.3 轮廓特征提取

对图像提取轮廓,首先将图像进行灰度化,然后获得图像的灰度直方图,再确定图像二值化的阈值,最后对图像进行二值化。

二值化后的图像只有两种颜色,从属性论的观点出发,每幅图片只有黑色和白色两种属性,它们所对应的量值分别为0和255。由于图像轮廓产生在这两种属性发生转变的临界部分,这样图像的轮廓就很容易提取出来了。将图像分成若干个区域,在这些区域中利用表示白色与黑色属性的定性映射做一个差,这个差值的绝对值不外乎两种情况:1和0,1表示发生了性质的转变,0表示未发生性质的转变。当每一个区域中有1值时,就将这个区域划为轮廓,否则不划为轮廓部分并将这个区域变成白色。

1.4 颜色特征提取

图像颜色的表达有许多种方式,如比较常见的RGB空间、HSV空间、HIS空间等。HSV空间具有如下特点:三个分量相对于人的视觉彼此独立,信息结构更紧凑;能够获得对彩色的直观表示;各彩色值根据主观评价均匀量化,彩色距离的大小与人眼的感觉一致;人对颜色的感觉

图形、图像与多媒体

Image Processing and Multimedia Technology

比较均匀。因此,本文以此空间来表示图像的颜色。HSV空间中,H表示色调,S表示饱和度,V表示亮度。

HSV空间中,根据参考文献[5],各分量的计算公式如下:

$$V = \max \quad (8)$$

$$S = (\max - \min) / \max \quad (9)$$

$$\text{其中, } h = \begin{cases} 5 + \frac{R-B}{R-G} & R = \max, G = \min \\ 1 - \frac{R-G}{R-B} & R = \max, B = \min \\ 1 + \frac{G-R}{G-B} & G = \max, B = \min \\ 3 - \frac{G-B}{G-R} & G = \max, R = \min \\ 3 + \frac{B-G}{B-R} & B = \max, R = \min \\ 5 - \frac{B-R}{B-G} & B = \max, G = \min \end{cases} \quad (10)$$

$$H = 60 \times h \quad (10)$$

该模型中,max、min表示R、G、B中最大的与最小的。当R=G=B时,S=0;当V=0时,表示当前的点是黑色。接着对计算出来的值进行归一化,归一化公式如下:

$$H = \begin{cases} 0 & h \in (316, 360] \cup [0, 20] \\ 1 & h \in (20, 40] \\ 2 & h \in (40, 75] \\ 3 & h \in (75, 155] \\ 4 & h \in (155, 190] \\ 5 & h \in (190, 270] \\ 6 & h \in (270, 295] \\ 7 & h \in (295, 316] \end{cases} \quad S = \begin{cases} 0 & s \in [0, 0.2] \\ 1 & s \in (0.2, 0.7] \\ 2 & s \in (0.7, 1] \end{cases} \quad (11)$$

$$V = \begin{cases} 0 & v \in [0, 0.2] \\ 1 & v \in (0.2, 0.7] \\ 2 & v \in (0.7, 1] \end{cases} \quad (11)$$

将得到的H、S、V的量化值合成一维,合成公式为:

$$1 = H \times Q_s \times Q_v + S \times Q_v + V \quad (12)$$

其中, Q_s 、 Q_v 分别表示S、V的量化级数,通过式(11)可知它们都取值为2。由式(12)很容易得到1的取值范围,然后再统计图像中各值的分布,即可得到图像的颜色谱。

1.5 聚类算法

将上面得到的颜色特征记为向量C,轮廓特征记为向量P。由于应用的不同,可能想要得到的结果也就不同。通常的做法是为每一种特征都赋一个权值,由于并无确定权值的具体算法,因此只能从反复实验中获得一组合理的。

聚类的步骤如下:

(1)由于开始时并无类别之分,因此就将第一个待归类的作为一个类别。

(2)若待匹配类别中只有一个元素,则可以用普通

的余弦公式来计算它们的相似度;如果元素在两个或者两个以上,就可以建立基准,利用如式(3)所示的转换程度函数公式来计算转换程度。

$$\text{sat}(x_j, x^*) = 1 - \frac{\sqrt{\sum_{j=1}^m |x_j - x^*(z_j)| \times |x_j - x^*(z_j)|}}{\sum_{j=1}^m \delta_j} \quad (13)$$

其中, x_j 为待聚类的特征, x^* 为已知类的本征值。

最终相似度由两种特征的相似度加权和得到。

(3)在对未知类图像归类后,对类别的基准以及本征值进行更新,即看每一维上的量值是否落在基准范围内,若不在,将其量值作为基准的端点。对于本征值,可再求次平均。

(4)如果未知图片与所有已知类别都不是很相似,也就是不满足聚类前规定的阈值,则将其划为新的一类,然后再回到步骤(2),如此循环下去,直到所有图片均归类。

2 实验结果及分析

本文选用Java和Eclipse来完成聚类。采用的图像库是哥伦比亚COIL100图像库,从中选取了20种不同的对象,每种对象45张图片。同类图片如图1所示,不同类型如图2所示。

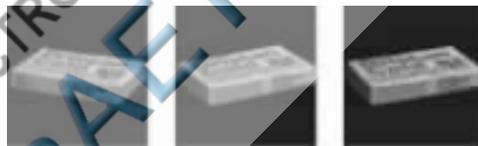


图1 同类图片样例



图2 不同类图片样例

本文对图像库做了多次聚类,即在每次聚类所得类的基础上再聚类。将利用本文方法所做的实验结果与参考文献[6]中方法的实验结果相比较,结果如表1所示。

表1 本文方法与参考文献[6]中方法的实验结果对比

	ARC	SOON	SVD	本文方法二次聚类	本文方法三次聚类
类别数	31	42	25	34	21
归类错误率/%	20	35	10	20	25

本文将属性论方法应用在图像聚类中,从表1中的结果可以看出该方法的有效性。但是本文提出的算法还存在很多不足,如权值以及相似度阈值的确定都比较依赖于人们的经验值或反复实验所获得的合理值,在无监督的聚类中,将这些值变为程序计算出来是很困难的。实验所用图像库是对物体做360°全方位拍摄所得的图片,本文只对图片做了一种变换(即对称变换)。下一步

的工作就是对图像做一个三维的变换来提高聚类的正确率以及减少聚类的次数。

参考文献

- [1] 李文佩. 基于定型映射和转化程度函数的汉字识别[D]. 上海:上海海事大学,2004.
- [2] 冯嘉礼. 定性映射、基准变换、转化程度函数、人工神经元和集合模糊化[J].模糊数学与系统,2004,18:53-56.
- [3] 冯嘉礼. 基于属性坐标学习和分析的核事故评估与决策模型[J].应用基础与工程科学学报,2006,5(14):1-8.
- [4] 冯嘉礼. 基于定性映射的程度函数及其诱导的模糊拓

扑空间[J].计算机科学,2003,30(5):106-108.

[5] 孙即祥. 图像处理[M].北京:科学出版社,2009.

[6] 谢从华,沈钧毅,宋余庆,等. 一种图像数据库聚类与归类方法的研究[J].控制与决策,2008,6(6):701-704.

(收稿日期:2011-09-25)

作者简介:

李根根,男,1987年生,硕士研究生,主要研究方向,模式识别与智能系统。

冯嘉礼,男,1948年生,教授,主要研究方向:人工智能。

