仿射传播算法在图像聚类应用中的实现与分析*

赵 健, 唐 洁, 谢 瑜
(西北大学 信息科学与技术学院, 西安 710069)

摘 要: 近年来,基于划分的聚类算法被广泛应用于数据和图像聚类中。针对应用最为广泛的 k-均值算法在图像聚类中存在的聚类速度慢、效果差等问题,提出一种仿射传播算法应用于图像聚类中。提取图像中颜色、形状和纹理等特征向量,利用仿射传播算法对综合特征向量模型进行聚类,最后将仿射传播算法和 k-均值算法对 MIT 图像的聚类作了对比分析。仿真实验表明,仿射传播算法在速度和聚类效果上均优于已有的 k-均值算法,在准确性和实时性方面均能达到较好的效果。

关键词: 仿射传播算法; 图像聚类; 相似度距离; 灰度共生矩阵
中图分类号: TP391 **文献标志码**: A **文章编号**: 1001-3695(2012)10-3980-03
doi:10.3969/j.issn.1001-3695.2012.10.101

Implementation and analysis of affine propagation algorithm in image clustering applications

ZHAO Jian, TANG Jie, XIE Yu
(*School of Information & Technology, Northwest University, Xi'an* 710069, *China*)

Abstract: Recently, the clustering methods based on partitioning are widely used in the field of data and image clustering. For the most widely k-means being slow and poor effect, this paper presented an improved algorithm based on affinity propagation, which was applied to image clustering. First, it presented the method combing color, shape and texture feature for efficient image retrieval. Then, on the basis of comprehensive characteristic model, it introduced a novel clustering method based on affinity propagation algorithm, Finally, it compared the results of affinity propagation with k-means in the MIT image database. The simulation experimental results show that the proposed method is superior to the traditional k-means clustering algorithm in the speed and effect of clustering. In addition, it is effective in exactness and real-time property.

Key words: affinity propagation algorithm; image clustering; similarity metrics; gray-scale co-occurrence matrix

0 引言

随着科学技术的发展和推广应用,大量的各种类型的信息在全球得到了采集、传输、流通和应用,其中图像信息就是一种极其重要的信息资源。近年来信息需求的增加,以各种方式获取的图像信息数量飞速增长。鉴于此,如何快速、准确、高效地从浩如烟海的图像信息源中获取有效的信息变得极为重要。

图像聚类就是在给出的图像集合中,根据图像的内容,在无先验知识的条件下,将图像分成有意义的类。对于图像聚类,最引人注目的特征属性是颜色、纹理和形状等。k-means 聚类[1]是一种应用最广泛的聚类算法,它能有效地处理大数据集,迭代速度快;其缺点是聚类数是预先设定的,聚类效果与初始聚类和事件的顺序有关。

基于以上原因,本文使用颜色直方图、边界方向直方图及灰度共生矩阵作为图像的综合特征,以减少误匹配,提高聚类的精确度。采用最新提出的仿射传播聚类(AP)算法[2]对图像进行聚类以克服了 k-means 的缺点,无须选择初始聚类中心,提高聚类速度。

1 聚类算法

1.1 k-均值算法

k-均值算法是一种应用最广泛的聚类算法,它能有效地处理大数据集,迭代速度快。通过该聚类算法,可以对图像库中的图像进行事先的分类处理,并建立这些类的索引表,加快检索效率。k-均值算法的思路如下:

设有 N 个样本点的数据集 $R=\{x_1,x_2,\cdots,x_N\}$,聚类为类,$N\geqslant k$。

a)任选 k 个聚类中心 $Z=\{z_1,z_2,\cdots,z_k\}$。

b)将每个待分类的样本按照最近准则分类到以初始聚类中心为标准的各类中去。

c)计算分类后各类的中心,形成新的聚类中心:$y_i=\frac{1}{N}\sum_{X_i\in R}X_i, i=1,2,\cdots,k$。

d)检测 $z_1,z_2,\cdots,z_k$ 是否等于 $y_1,y_2,\cdots,y_k$,若相等,则算法结束;否则用 y_i 代替 z_i,返回 b)。

1.2 仿射传播算法

仿射传播聚类是一种新的聚类算法,其优势体现在处理类

收稿日期: 2012-02-28; **修回日期**: 2012-03-30 **基金项目**: 陕西省教育厅科技立项项目(2010JK847);西北大学研究生重点课程项目基金资助项目(09YKC21)

作者简介:赵健(1973-),男,河北河间人,教授,硕导,双博士(后),主要研究方向为图像处理、数字水印、模式识别;唐洁(1986-),女,陕西三原人,硕士,主要研究方向为信号与信息处理(zjctec@nwu.edu.cn);谢瑜(1974-),女,安徽灵璧人,工程师,学士,主要研究方向为信号与信息处理.

数很多的情况时运算速度快[3,4]。

a)按照式(1)计算数据集中两两数据间的相似度距离,为使均方误差达到最小,在每个元素前加上负号成为负数,得到一个相似度矩阵 S 作为输入。如对点 x_i 和 x_k,有

$$S(i,k) = -\| x_i - x_k \|^2 \tag{1}$$

b)设置偏向参数。偏向参数 $p(i)$(通常是负数)表示数据点 i 被选作聚类中心的倾向性,而且增大或减小 p 可以增加或减少 AP 输出的聚类数目。在无先验知识时,将所有数据点都视为潜在的类代表,则 $p(i)$ 设定为 S 中元素的中值[5]。

c)AP 算法的核心内容,在数据点之间有两类消息反复传递,反映了消息的收集与积累的过程:吸引度 $r(i,k)$ 表示数据点 k 适合作为数据点 i 的类代表的吸引程度;归属度 $a(i,k)$ 表示数据点 i 选择点 k 作为它的类代表的归属程度。消息的迭代过程如下:

初始化归属度 $a(i,k)=0$。吸引度的更新规则为

$$r(i,k) \leftarrow s(i,k) - \max_{k'\text{s.t.}\,k'\neq k} \{a(i,k') + s(i,k')\} \tag{2}$$

吸引度更新让所有的候选类代表参与竞争,而归属度更新从数据点中搜集证据,决定候选类代表是否能成为一个合适的类代表。其更新规则为

$$a(k,k) \leftarrow \min\{0, r(k,k) + \sum_{i'\text{s.t.}\,i'\notin\{i,k\}} \max\{0, r(i',k)\}\} \tag{3}$$

为了限制吸引度不断增大,对于总和设定一个阈值,防止它成为 0。则自归属度 $a(k,k)$ 的更新规则为

$$a(k,k) \leftarrow \sum_{i'\text{s.t.}\,i'\neq k} \max\{0, r(i,k)\} \tag{4}$$

在消息传递中,阻尼因子是另一个重要的参数。在每次循环迭代中,$r(i,k)$ 和 $a(i,k)$ 的更新结果都是由当前迭代过程中更新的值和上一步迭代的结果加权得到的。加权更新式为

$$r_{\text{new}}(i,k) = \lambda r_{\text{old}}(i,j) + (1-\lambda) r_{\text{new}}(i,j) \tag{5}$$

$$a_{\text{new}}(i,k) = \lambda a_{\text{old}}(i,j) + (1-\lambda) a_{\text{new}}(i,j) \tag{6}$$

最终产生了 m 个高质量的类代表,同时能量函数也得到了最小化。将各数据点分配给最近的类代表所属的类,则找到的 m 个聚类即是聚类结果。

2　图像特征提取

2.1　颜色直方图

颜色是描述图像内容最直接的视觉特征。其中颜色直方图除了简单外,还具有对平移和旋转不敏感的优点。因此,目前大部分的检索系统都采用颜色直方图作为图像的基本特征。本文采用在 RGB 空间提取颜色直方图[6]。

首先提取图像的红、绿、蓝三个颜色分量的直方图,并对其进行归一化处理。然后对每一个颜色分量平均划分为 16 个区域,这样三个分量可形成 48 个区域,则颜色特征可以表示为一个 48 维的特征向量 Hc。

2.2　边界方向直方图

边缘是图像最基本的特征,可以体现图像的空间信息,因此本文采用基于边缘检测的方法。其中图像锐化可以加强边缘特征,使边缘易于识别。本文中主要采用 Canny 边缘算子的锐化方法[7]。Canny 算法首先采用二维高斯函数对图像进行平滑;然后细化梯度幅值图像的屋脊带,只保留幅值的局部极大值,精确定位边缘;最后采用双阈值法从候选边缘点中检测和连接出最终边缘。使用 Canny 算子提取图像边界后,对边界方向以 5°为范围划分,形成一个 72 级的边界方向直方图 Hi,并按照式(7)(8)进行归一化和平滑处理,使其具备尺度和旋转不变性。

$$Hi = \frac{Hi}{S} \tag{7}$$

$$Hs = \frac{\sum_{j=i-k}^{i+k} Hi}{2k+1} \tag{8}$$

其中,参数 k 决定平滑度。这样,就可以得到一个 72 维的特征向量 Hs。

2.3　灰度共生矩阵

纹理通常定义为图像的某种局部性质,或是对局部区域中像素之间关系的一种度量,本文采用的灰度共生矩阵[6] $P_{i,j}$ 反映图像灰度分布关于方向、局部邻域及变换幅度的综合信息,需从 $P_{i,j}$ 中进一步提取描述图像纹理特征。

首先计算四个方向上的 $P_{i,j}$,角度分别为 0°、45°、90°、135°。然后对四个方向上的 $P_{i,j}$ 按照式(9)~(12)计算能量、熵、惯性矩、相关性四个纹理参数。

能量是图像灰度分布均匀性的度量,从图像整体来观察,纹理较粗,此时能量 E 较大;反之,细纹理的 E 较小。能量的定义式为

$$E = \sum_{i=0}^{N-1}\sum_{j=0}^{N-1} p_{i,j}^2 \tag{9}$$

图像的对比度可理解为图像的清晰度,即纹理的清晰程度。在图像中,纹理的沟纹越深,图像的视觉效果越清晰,则对比度 I 越大。惯性矩的定义式为

$$I = \sum_{i=0}^{N-1}\sum_{j=0}^{N-1} (i-j)^2 p_{i,j} \tag{10}$$

相关性是用来衡量灰度共生矩阵的元素在行的方向或列的方向的相似程度。四个方向的角度分别为 0°、45°、90°、135°。相关性的定义公式为

$$C = \frac{\sum_{i=0}^{N-1}\sum_{j=0}^{N-1} ijp_{i,j} - \mu x \mu y}{\sigma x^2 \sigma y^2} \tag{11}$$

熵是图像所具有的信息量的度量。若图像没有任何纹理,则灰度共生矩阵几乎为零阵,则熵值 H 接近为零;若充满细纹理,则 H 最大;若分布着较少的纹理,则 H 较小。熵的定义式为

$$H = \sum_{i=0}^{N-1}\sum_{j=0}^{N-1} p_{i,j} \log p_{i,j} \tag{12}$$

最后计算出能量、对比度、惯性矩和熵的四个方向上的均值 μ、标准差 σ,并按照式(13)进行归一化,使其具有相同的权重,最终形成 16 维特征向量 Ht。高斯归一化式为

$$Ht = \left(\frac{Ht-\mu}{3\sigma} + 1\right)/2 \tag{13}$$

3　算法的实现过程与结果分析

对于任意两幅图像 t_1 和 t_2,按照第 2 章所述方法对图像库中所有图像提取颜色直方图 Hc、边界方向直方图 Hs 和灰度共生矩阵 Ht。然后按照式(14)计算任意两幅图像之间的综合相似度距离,记为 D_{ij},并将其作为相似度矩阵中的元素,其中值作为偏向参数 p,最后用 AP 聚类算法进行聚类。

$$D(t_1,t_2,t_3) = \frac{w_c D_c + w_s D_s + w_t D_t}{w_c + w_s + w_t} \tag{14}$$

其中:w_c、w_s 和 w_t 分别表示对颜色、形状和纹理特征的加权,且

$w_c + w_s + w_t = 1$。

在 AP 算法聚类完成之后,对其中的每一类计算该类图像的平均颜色直方图、平均边界方向直方图和灰度共生矩阵。

计算示例图像的颜色直方图、边界方向直方图和灰度共生矩阵;计算其与图像库中每一类的相似度距离,取距离最近的一类,并计算示例图像与最近一类中的每一幅图像的相似度距离;如果距离小于设定的阈值,则将此图像检出,作为检索结果。之后再取距离次近的一类,输出与示例图像相关的图像。

用查全率与查准率这两项重要指标来对图像聚类进行有效性的评价。

$$\text{查准率} = \frac{\text{查到的相关图像数目}}{\text{查到的所有图像数目}} \times 100\%$$

$$\text{查全率} = \frac{\text{查到的相关图像数目}}{\text{图像库中与目标图像相关的图像数目}} \times 100\%$$

为了验证本文所述方法的有效性,采用 MIT 图像库中的 3 000幅图像进行实验,其中包括动物、植物、风景、建筑和人物等,图 1 是测试集的部分图像。测试环境在 Windows XP 环境下,并采用 MATLAB 7.1 软件仿真实现。

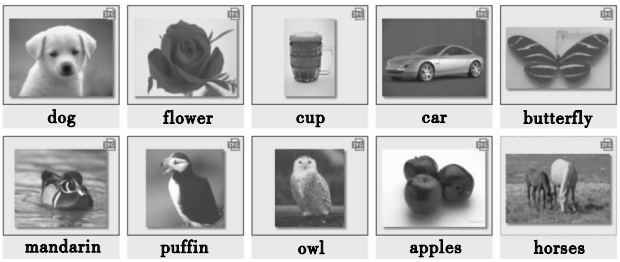

图1 测试集部分图像

提取出图像库中每幅图像的特征后,分别采用 k-均值和仿射传播算法对图像进行聚类,结果如图 2 和 3 所示。

图2 k-均值算法的聚类结果

图3 仿射传播算法的聚类结果

通过图 2 和 3 的比较可以看出,仿射传播聚类算法对图像特征向量聚类方式的结果检索,按照先从上到下、再从左到右的顺序排列,相似性越大的图像位置越靠前。

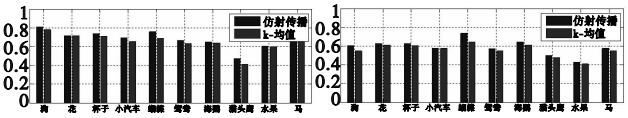

图4 测试集中部分图像的检索结果

图 4 是测试集中部分图像的检索结果的查准率和查全率的统计直方图。从图 4 中可以看出,通过仿射传播算法对图像进行聚类得到的准确率与 k-均值相比有所提高,可以有效提高检索的准确率。

从表 1 中可以看出,本文提出的方法相对于 k-means 聚类算法,检索时间、查全率和查准率都有较大提高。由此证明了本文所采用的聚类方法在查准率、查全率和检索时间上具有更好的效率;而且本文采用的颜色、形状和纹理的综合特征,相对于只使用一种特征进行检索,具有更高的稳定性、检索精确度及更好的区分度。

表 1 AP 与 k-均值的检索方法对比实验结果

比较项	k-均值	AP
聚类数目	54	59
平均查到图像数目	41.7	17.7
平均相似图像数目	15.2	13.5
平均查准率/%	40.25	77.24
平均查全率/%	72.42	73.08
执行时间/s	3 301.754	3 287.462
迭代次数	5	81

4 结束语

本文采用更高效率的 AP 聚类算法和区分度更好的颜色、形状和纹理的综合特征,对 MIT 图像库进行聚类并获得较好的检索结果。AP 聚类算法运算速度快,聚类结果更加准确,并且无须指定聚类中心,特别是当聚类目标数量较大时,优势更加明显。经过实验证明,本文的方法在查准率、查全率上均优于 k-均值算法。

参考文献:

[1] KANUNGO T, MOUNT D M, NETANYAHU N S. An efficient k-means clustering algorithm: analysis and implementation[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2002, 24(7): 881-892.

[2] FREY B J, DUECK D. Clustering by passing messages between data points[J]. Science, 2007, 315(5814): 972-976.

[3] MEZARD M. Where are the exemplar[J]. Science, 2007, 315(5814): 949-951.

[4] 王开军,张军英,李丹,等. 自适应仿射传播聚类[J]. 自动化学报, 2007, 33(12): 1242-1245.

[5] JIA Y Q, WANG J D, ZHANG C S, *et al*. Finding image exemplars using fast sparse affinity propagation [C]//Proc of the 16th ACM International Conference on Multimedia. New York: ACM, 2008: 639-642.

[6] YANG Fang-yu, WANG Xiang-yang. Color image retrieval based on multiple features of image edges[J]. Computer Science, 2010, 37(2): 256-260.

[7] 杨芳宇,王向阳. 一种基于边缘综合特征的彩色图像检索算法[J]. 计算机科学, 2010, 37(2): 256-260.