

doi:10.3969/j.issn.1002-0802.2017.12.011

一种改进的 K -means 聚类算法在图像分割中的应用^{*}

任恒怡, 贺 松, 陈文亮

(贵州大学 大数据与信息工程学院, 贵州 贵阳 550025)

摘 要: K -means 聚类算法是图像分割中比较常见的一种方式。它是一种无监督学习方法, 能够从研究对象的特征中发现关联规则, 因而具有强有力的处理方法。但是, 由于该算法对噪声的敏感性 K 值及初始类心的不确定性, 使其在图像分割中存在缺陷, 于是提出了一种改进的 K -means 聚类算法来提高分割的效果。首先对图像进行平滑滤波处理, 再根据相应条件找到特征向量作为初始类心, 最后进行聚类操作。实验表明, 本算法能够有效提取目标对象, 提高图像分割的效果。

关键词: K -means 聚类算法; 平滑滤波; 欧式距离; 图像分割

中图分类号: TP391.9 **文献标志码:** A **文章编号:** 1002-0802(2017)-12-2704-04

Application of Modified K -means Clustering Algorithm in Image Segmentation

REN Heng-yi, HE Song, CHEN Wen-liang

(College of Big Data and Information Engineering, Guizhou University, Guiyang Guizhou 550025, China)

Abstract: K -means clustering algorithm is a common way in image segmentation, and as a kind of unsupervised learning method, could discover the association rules from the characteristics of the research object, and thus is of strong processing ability. However, due to its sensitivity to the noise and uncertainty of the initial centroid, this algorithm still has some defects in the image segmentation. Based on this, a modified K -means clustering algorithm is proposed to improve the segmentation results. Firstly, the image is smoothed and filtered, and then the feature vector is found as the initial centroid according to the corresponding conditions; finally, the clustering operation is performed. Experiments show that this algorithm could effectively extract the target object and improve the effect of image segmentation.

Key words: K -means clustering algorithm; smooth filtering; Euclidean distance; image segmentation

0 引 言

图像分割是从图像中把人们感兴趣的目标对象用有效手段提取出来。它是图像分析和理解的基础^[1]。目前, 主要的分割方法主要有区域分割、边缘分割、阈值分割等^[2]。

聚类分析是从研究对象中发现关联规则的一种半监督学习方法, 具有强有力的处理信心^[3]。它把图像中的各个像素用相应的特征向量表示, 根据特征空间的相似性进行分割, 并映射到原图像空

间, 主要有 K 均值、模糊均值 K 均值聚类算法^[4]。

1 传统的 K -means 聚类算法

K -means 聚类算法是一种简洁、快速的算法, 能够用于处理大数据集, 并呈现出可伸缩、高效的特点^[5]。它的结果类比较密集, 类与类之间有明显的区别, 是解决聚类问题的经典算法。但是, 它也存在局限性, 主要有以下几点:

(1) K -means 的初始聚类中心是随机抽取的,

^{*} 收稿日期: 2017-08-16; 修回日期: 2017-11-10 Received date: 2017-08-16; Revised date: 2017-11-10

基金项目: 贵州省工程技术中心建设项目 (黔科合 G 字 [2014]4002 号)

Foundation Item: Construction Project of Guizhou Provincial Engineering and Technology Centre (No.[2014]4002)

对于不同的初始聚类中心具有不同的效果;

(2) K -means 必须事先定好聚类数目, 但聚类个数 K 难以确定;

(3) K -means 对噪声和孤立点比较敏感, 对少量数据往往会对均值产生较大影响;

(4) K -means 要在类的均值被定义的前提下才能使用, 对符号属性的数据不适用^[6]。

2 改进的 K -means 聚类算法

传统 K -means 聚类算法的局限性, 对图像分割的效果会产生较大影响。本文提出改进的 K -means 聚类算法, 主要是为了弥补传统算法对噪声和孤立点敏感、 k 值不确定以及初始类心随机性的问题^[7]。具体流程如图 1 所示。



图 1 改进 K -means 聚类算法流程

2.1 平滑滤波

在图像采集、传递、获取的过程中, 由于所在的环境复杂多变, 受光照、电磁波等可见不可见的噪声干扰, 对图像的分割造成了很大影响, 使得分割结果不甚理想^[8]。采用平滑滤波处理能够尽量降低由噪声引起的影响^[9]。

本文主要对图像进行平滑处理, 使其保持总体特征。令 S_{xy} 表示彩色图像中的一组坐标, 定义为以 (x,y) 为中心的邻域, 邻域中 RGB 向量的平均值为:

$$\vec{C}^{(x,y)} = \frac{1}{K} \sum_{(s,t) \in S_{xy}^{C(s,t)}} \quad (1)$$

其中, K 是邻域中像素中的数量。满足向量加法性质:

$$\vec{C}^{(x,y)} = \begin{bmatrix} \frac{1}{K} \sum_{(s,t) \in S_{xy}^{R(s,t)}} \\ \frac{1}{K} \sum_{(s,t) \in S_{xy}^{G(s,t)}} \\ \frac{1}{K} \sum_{(s,t) \in S_{xy}^{B(s,t)}} \end{bmatrix} \quad (2)$$

使用上述滤波器模板可以对每幅分量图执行邻域平均, 且邻域平均的平滑可以在每个图像平面的基础上执行, 滤波效果如图 2 所示。

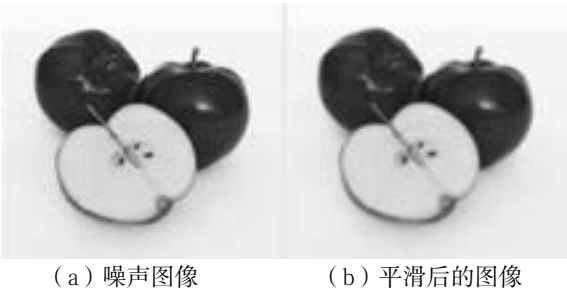


图 2 平滑滤波结果对比

2.2 确定初始类心

步骤 1: 假设聚类的个数为 K , T 表示聚类中心, 则 $T=\{T_1,T_2\cdots T_k\}$, 类的集合为 $Y=\{Y_1,Y_2\cdots Y_k\}$ 。首先, 确定两个初始聚类中心, 分别为 T_1 和 T_2 。假设图像的特征向量个数为 N 个, 维度为 M , 那么特征向量集为 $X=\{X_1,X_2\cdots X_k\}$ 。用欧式距离公式计算任意两个特征向量的距离 D_{ij} , 则:

$$D_{ij} = |X_i - X_j| = \sqrt{(X_{i1} - X_{j1})^2 + \cdots + (X_{im} - X_{jm})^2} \quad (3)$$

$(1 < i < N, 1 < j < N, \text{且} i \neq j)$

由计算结果查找最大的 D_{ij} , $D_{ij}=\max_{i,j}|D_{ij}|$, 确定特征向量 X_i 和 X_j 作为初始聚类中心。

步骤 2: 由步骤 1 确定了两个初始聚类中心, 假设已确定 k 个聚类中心, 那么第 $k+1$ 个聚类中心 T_{k+1} 就是特征向量集中其余 $N-k$ 个特征向量与前 k 已确定的聚类中心距离累加和最大的特征向量, 公式如下:

$$S_i = \sum_{j=1}^k |X_i - T_j| |X_i - C_j| = \sqrt{(X_{i1} - T_{j1})^2 + \cdots + (X_{im} - T_{jm})^2} \quad (4)$$

由计算结果查找最大的 S_i , $S_i=\max_i|S_i|$, 那么特征向量 X_i 作为第 $k+1$ 的初始类中心, 即 $T_{k+1}=X_1$ 。重复步骤 2, 直到 K 个初始聚类中心都找到为止。

2.3 进行 K -means 聚类

2.3.1 K -means 聚类算法的步骤

K -means 算法的原理是先从数据样本中随机抽取 K 个作为初始聚类中心, 再计算样本到聚类的最近距离并归类, 然后计算新的聚类的均值作为新的聚类中心, 最后重复以上操作直到相邻聚类没有变化, 则聚类达到最佳^[10]。

具体步骤如下:

步骤 1: 任取 K 个作为初始聚类中心, 设 K 个中心分别为 $T_1(1)$ 、 $T_2(1)$ 、 $T_k(1)$;

步骤 2: 所有样本 X 按式 (5) 分类, 使样本 X 的样本分别与相应的类心 $T_1(n), T_2(n) \cdots T_k(n)$ 的子集 $Y_1(n), Y_2(n) \cdots Y_k(n)$ 相对应;

$$D_1 = \min \{D_j\} \rightarrow \begin{matrix} X \in Y_i(n) \\ j \in N_k \end{matrix}, N_k = \{1, 2, 3, \cdots k\} \quad (5)$$

步骤 3: 计算 X 和 $T_j(n)$ 之间的距离:

$$D_j = \|X - T_j(n)\| \quad (6)$$

步骤 4: 由式 (7) 计算各子集 $Y_i(n)$ 的新类心 $T_i(n+1)$, N 是集合 $y(n)$ 中的元素个数。其中, $T_i(n+1)$ 是属于 Y_i 的 X 的均值;

$$T_i(n+1) = \frac{1}{N_i} \sum_{X \in Y_i(n)} X \quad (7)$$

步骤 5: 当式 (8) 成立时结束操作, 回到步骤 2 继续操作;

$$T_i(N+1) = T_i(N), 1 = N_k \quad (8)$$

K -means 聚类算法在图像分割中的应用, 使得所分割的各区域具有差异性^[11], 但其内部区域又有相似性, 其分割效果取决于初始类心的选取和分类数的确定^[12]。

2.3.2 K -means 聚类算法的要点

(1) 相似性度量

本文主要采取欧式距离作为样本的相似性度量用来计算样本之间的距离。

$$D(X_i, X_j) = \|X_i, X_j\| = (X_i, X_j) T(X_i, X_j) = \sqrt{\sum_{k=1}^n (X_{ik} - X_{jk})^2} \quad (9)$$

当样本间的距离越小时, X_i 和 X_j 相似性越大, 存在的差异性越小^[13]。反之, 样本间的距离越大, X_i 和 X_j 相似性越小, 差异性越大。

(2) 聚类中心终止条件

K -means 聚类算法在每次迭代的过程中都要判定样本的分类是否正确^[14]。若不正确, 则需要进行调整直至结束, 然后再进行聚类中心的修改并进入下一次迭代过程, 直到满足条件:

①每个对象与聚类相对应, 不能同时分配给不同对象;

②聚类的中心固定, 不再发生变化;

③误差平方和准则函数的局部最小^[15]。

2.3.3 聚类性能评价

假设数据集 X 含有 k 个聚类子集, 分别为

$X_1, X_2, \cdots X_n$, 每个聚类子集的样本数量为 $n_1, n_2, \cdots n_k$, 每个聚类子集的聚类中心为 $Z_1, Z_2, \cdots Z_k$, 那么误差平方和准则函数公式是:

$$E = \sum_{i=1}^k \sum_{p \in X_i} \|P - Z_i\|^2 \quad (10)$$

3 算法实现及仿真结果分析

为了证明改进 K -means 聚类算法的有效性, 本文对两种算法分别进行仿真实验并对结果进行对比, 结果如图 3 所示。通过仿真结果可以看到, 改进的 K -means 聚类算法图 3 (c) 对图像分割效果较传统的 K -means 聚类算法图 3 (b) 有明显改进, 避免了噪声对分割造成的影响, 有效提取了需要进行分割的区域。



(a) 原图

(b) 传统 K -means 聚类结果



(c) 改进 K -means 聚类结果

图 3 聚类结果对比

4 结 语

本文提出的 K -means 聚类算法有效改进了原始 K -means 算法, 先通过平滑滤波进行图像去噪处理, 再用欧式距离计算特征向量之间的距离来确定初始聚类类心, 最后通过 K -means 聚类进行分割问题, 使得图像呈现出更好的效果。

参考文献:

- [1] 李苏梅, 韩国强. 基于 K -均值聚类算法的图像区域分割方法 [J]. 计算机工程与应用, 2008, 44(16): 163-167.
LI Su-mei, HAN Guo-qiang. Research on Image Segmentation based on K -means Clustering Algorithm [J]. Computer Engineering and Applications, 2008, 44(16): 163-167.

- [2] 王爱莲, 伍伟丽, 陈俊杰. 基于 K-means 聚类算法的图像分割方法比较及改进 [J]. 太原理工大学学报, 2014, 45(03): 372-375.
WANG Ai-lian, WU Wei-li, CHEN Jun-jie. Comparison and Improvement of Image Segmentation based on K-means [J]. Journal of Taiyuan University of Technology, 2014, 45(03): 372-375.
- [3] 周新建, 涂弘斌. 基于改进的 K-means 聚类图像分割算法 [J]. 无损检测, 2007, 29(05): 258-261.
ZHOU Xin-jian, TU Hong-bin. Image Segmentation based on a Modified K-means Algorithm [J]. Non-Destructive Testing, 2007, 29(05): 258-261.
- [4] 黄韬, 刘胜辉, 谭艳娜. 基于 K-means 聚类算法的研究 [J]. 计算机技术与发展, 2011, 21(07): 54-57.
HUANG Tao, LIU Sheng-hui, TAN Yan-na. Research of Clustering Algorithm based on K-means [J]. Computer Technology and Development, 2011, 21(07): 54-57.
- [5] 黄宇, 付琨, 吴一戎. 基于 Markov 随机场 K-Means 图分割算法 [J]. 电子学报, 2009, 37(12): 2700-2704.
HUANG Yu, FU Kun, WU Yi-rong. Image Segmentation Method Using K-Means based on Markov Random Field [J]. Acta Electronica Sinica, 2009, 37(12): 2700-2704.
- [6] 李栋, 刘萌萌, 郭莎. 基于改进的 K-means 算法在图像分割中的应用 [J]. 电脑知识技术, 2016, 12(08): 166-168.
LI Dong, LIU Meng-meng, GUO Sha. Application of Improved K-means Algorithm in Image Segmentation [J]. Computer Knowledge and Technology, 2016, 12(08): 166-168.
- [7] 李翠, 冯冬青. 基于改进 K-均值聚类算法的图像分割算法研究 [J]. 郑州大学学报, 2011, 43(01): 109-113.
LI Cui, FENG Dong-qing. Study on Image Segmentation Algorithm based on Improved K-means Clustering Algorithm [J]. Journal of Zhengzhou University, 2011, 43(01): 109-113.
- [8] 吕明磊, 刘冬梅, 曾智勇. 一种改进的 K-means 聚类算法的图像检索方法 [J]. 计算机科学, 2013, 40(08): 285-288.
LV Ming-lei, LIU Dong-mei, ZENG Zhi-yong. Novel Image Retrieval Method of Improved K-means Clustering Algorithm [J]. Computer Science, 2013, 40(08): 285-288.
- [9] 史习云, 薛安荣, 刘艳红. 改进 K-means 聚类算法在图像检索中的应用研究 [J]. 计算机工程与应用, 2011, 47(10): 193-196.
SHI Xi-yun, XUE An-rong, LIU Yan-hong. Application Study of Improved K-means Clustering Algorithm in Image Retrieval [J]. Computer Engineering and Applications, 2011, 47(10): 193-196.
- [10] 刘小丹, 牛少敏. 一种改进的 K-means 聚类彩色图像分割方法 [J]. 湘潭大学自然科学学报, 2012, 34(02): 90-93.
LIU Xiao-dan, NIU Shao-min. Improved K-means Clustering for Color Image Segmentation Method [J]. Natural Science Journal of Xiangtan University, 2012, 34(02): 90-93.
- [11] 赵凤娇, 贺月姣. 基于改进的 K-means 聚类算法水下图像边缘检测 [J]. 现代电子技术, 2015, 38(18): 89-91.
ZHAO Feng-jiao, HE Yue-jiao. Underwater Image Edge Detection based on Improved K-means Clustering Algorithm [J]. Modern Electronics Technique, 2015, 38(18): 89-91.
- [12] 毛秀, 冒纯丽, 丁岳伟. 基于密度和聚类指数改进的 K-means 算法 [J]. 电子科技, 2015, 28(11): 47-50.
MAO Xiu, MAO Chun-li, DING Yue-wei. Improved K-means Algorithm based on Density and Clustering Index [J]. Electronic Science and Technology, 2015, 28(11): 47-50.
- [13] 韩凌波. 基于密度的 K-means 初始聚类中心选取算法 [J]. 电子科技, 2015, 28(07): 105-107.
HAN Ling-bo. K-means Initial Clustering Center Selection Algorithm based on Density [J]. Electronic Science and Technology, 2015, 28(07): 105-107.
- [14] 楚晓丽. K-Means 聚类算法和人工鱼群算法应用于图像分割技术 [J]. 计算机系统应用, 2013, 22(04): 92-94.
CHU Xiao-li. K-Means Clustering Algorithm and Artificial Fish Swarm Algorithm Applied in Image Segmentation Technology [J]. Computer Systems and Applications, 2013, 22(04): 92-94.
- [15] 杨明川, 吕学斌, 周群彪. 不完全 K-means 聚类与分类优化结合的图像分割算法 [J]. 计算机应用, 2012, 32(01): 248-251.
YANG Ming-chuan, LV Xue-bin, ZHOU Qun-biao. Image Segmentation Algorithm based on Incomplete K-means Clustering and Category Optimization [J]. Journal of Computer Applications, 2012, 32(01): 248-251.

作者简介:



任恒怡 (1992—), 女, 在读硕士, 主要研究方向为数字图像处理、大数据分析;

贺 松 (1974—), 男, 硕士, 副教授, 主要研究方向为数字图像处理、医疗大数据;

陈文亮 (1990—), 男, 在读硕士, 主要研究方向为数字图像处理、数据挖掘与分析。