流形距离的自动免疫克隆聚类图像分割算法

邓晓政，焦李成

(西安电子科技大学智能感知与图像理解教育部重点实验室 西安 710071)

【摘要】聚类算法在对图像进行分割的过程中要面对如何自动确定聚类类别数、如何克服图像特征点分布复杂的流形结构、如何减少算法的运行时间。针对这些问题，提出了流形距离的自动免疫克隆聚类图像分割算法。自动免疫克隆聚类算法可以自动确定聚类个数，不需要人为事先给定，并且确保全局收敛；使用流形距离可以反映空间分布复杂的流形数据；使用超像素而非像素来降低图像分割的时间等问题。通过对4组人工数据集和4幅自然图像进行实验，对比k-means算法、GCUK算法，结果表明该方法优势比较明显，具有一定的实用性和先进性。

关 键 词 聚类; 图像分割; 免疫克隆; 流形

中图分类号 TP391.4 文献标志码 A doi:10.3969/j.issn.1001-0548.2014.05.019

Automatic Immune Clonal Clustering Method Using Manifold Distance for Image Segmentation

DENG Xiao-zheng and JIAO Li-cheng

(Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education of China, Xidian University Xi'an 710071)

Abstract There are several difficulties in using a partitional clustering algorithm to deal with image segmentation problem including choosing the correct number of clusters without any prior knowledge, measuring the image datasets with complicated manifold structures and reducing the computation time. In this paper, an automatic immune clonal clustering method using manifold distance is applied to image segmentation. This method can automatically determine the number of clusters, measure the complicated manifold dataset by using manifold distance, and less computation time by using super-pixels instead of pixels. Experimental results on four artificial data sets and four Berkeley images show that the novel method outperforms the k-means algorithm and the GCUK algorithm.

Key words clustering; image segmentation; immune clonal; manifold

图像分割技术是图像处理和模式识别领域的关键和基础[1]。图像分割是将待分割的图像分成不同的区域，每个单独的区域是同质的(灰度、颜色或纹理相近)，但是任意两个相邻的区域不是同质的一个过程[1]。目前图像分割在国防、农业、采矿、医疗等各个领域都起着重要的作用。近年来国内外专家学者提出了很多不同的图像分割方法：直方图阈值法、基于区域的合并和分裂法、基于模型的方法、基于聚类的方法等[2]。对于图像，特别是彩色图像，颜色空间是天然的特征空间，使用聚类方法在特征空间聚类是最直接有效的方法。

聚类是将一个无类标属性的数据集划分成若干子集，每一个子集称为一个类簇，相同的类簇之间是相似的，而不同的类簇之间是不相似的。聚类方法可以分为层次聚类(hierarchical)或划分聚类(partitional)，也可分为硬聚类(crisp)或者模糊聚类(fuzzy)。本文使用的是硬划分聚类。

现有的聚类算法中，k-means算法和FCM算法是最经典常见的聚类算法[3]，优点是算法操作简单，运行快捷，适用于大规模图像数据集。但是缺点也较为明显，主要有初始选取的聚类中心对最终结果影响较大，极易陷入局部最优值，且只适合球形和超球分布的数据集。而近年来涌现的基于进化计算的图像分割聚类算法[4-6]，具有良好的全局搜索能力，不容易陷入局部最优值，但是它们大多使用欧几里德距离，限制了其仅适合球形分布的图像数据。文献[7]提出了基于流形距离的人工免疫聚类算法，克服了欧几里德距离的缺点，对流形结构分布的数

收稿日期：2014 – 01 – 03，修回日期：2014 – 07 – 08
基金项目：国家自然科学基金(61001202)；国家教育部博士点基金(20100203120008)
作者简介：邓晓政(1982 –)，男，博士，主要从事人工免疫系统、智能图像处理方面的研究.

据有良好的效果。虽然很多彩色图像数据呈现此类数据结构[8-9]，但是不能直接使用该算法对图像进行分割，因为流形距离计算复杂度较高，几乎无法计算像素点对间的流形距离。另外，该算法需要人工提前输入正确的聚类类别数，也是其不足之处。

针对这样的研究背景和问题，本文提出了流形距离的自动免疫克隆聚类图像分割算法(简称AICCMD)。优点如下：1) 使用超像素而非像素作为聚类操作单元，可以大大降低算法运行的时间和空间复杂度；2) 使用免疫克隆优化算法是因为它比一般的进化算法寻优能力更强；3) 设计了一个使用流形距离的聚类有效性指标来自动确定复杂流形数据的聚类类别数。

1　流形距离

在聚类问题中，一个合适的距离测度对聚类算法至关重要。常用的欧式距离无法反应数据的全局分布情况。从图1可以形象地反映出点 X 与点 Z 的欧式距离小于点 X 与点 Y 的距离，这容易导致将点 X 与点 Z 错误地划分成一类。于是，文献[7]提出了流形距离。

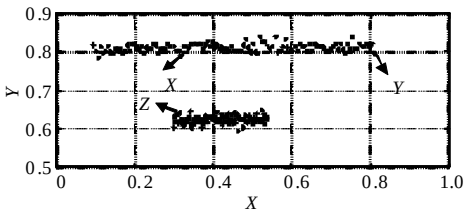


图1　欧氏距离的不足

根据定义，点 x_i 与点 x_j 的流形距离为：

$$M(x_i,x_j)=\min_{p\in P_{ij}}\sum_{k=1}^{|p|-1}S(p_k,p_{k+1}) \tag{1}$$

式中，按照图论理论，P_{ij} 为两点之间所有路径的集合；p 表示其中一条路径，长度为 $|p|-1$；$S(p_k,p_{k+1})$ 表示两点之间的边权值，具体为：

$$S(p_k,p_{k+1})=\mathrm{e}^{\alpha\cdot\mathrm{dist}(p_k,p_{k+1})}-1 \tag{2}$$

式中，$\alpha>0$；$\mathrm{dist}(p_k,p_{k+1})$ 为两点之间的欧式距离。很明显，这个距离测度满足自反性、非负性、对称性、三角不等性。相同流形上的两点可以用许多较短边相连，而不同流形上的两点却要用较长边相连，从而可以很好地反映数据的全局一致性。

2　免疫克隆聚类算法

免疫克隆选择算法借助生物学免疫系统的机理，比一般的遗传算法有更好的全局搜索能力，且兼顾局部寻优，具有巨大的解决工程问题的潜力[10-11]。在优化问题中，抗原指问题的目标函数，抗体指针对目标函数的候选解，抗体−抗原亲和度指候选解代入目标函数的值。本文为了方便起见，根据k-medoids聚类算法，将抗体表示成一个整数串，每个基因位上的值表示每个类簇的最佳代表样本点的序号。故编码长度为聚类的类别数。对抗体群 $\boldsymbol{A}(k)$ 依次执行如下操作：

克隆操作：$\boldsymbol{A}(k)\xrightarrow{\text{克隆操作}}\boldsymbol{A}'(k)$

其中，$\boldsymbol{A}(k)=[\boldsymbol{a}_1(k),\boldsymbol{a}_2(k),\cdots,\boldsymbol{a}_n(k)]$

$$\boldsymbol{A}'(k)=[\boldsymbol{a}_{11}(k),\boldsymbol{a}_{12}(k),\cdots,\boldsymbol{a}_{1q_1}(k)]+[\boldsymbol{a}_{21}(k),\boldsymbol{a}_{22}(k),\cdots,\boldsymbol{a}_{2q_2}(k)]+\cdots+[\boldsymbol{a}_{n1}(k),\boldsymbol{a}_{n2}(k),\cdots,\boldsymbol{a}_{nq_n}(k)]$$

克隆表示对抗体的原样复制，q_i 表示对抗体 $\boldsymbol{a}_i$ 的复制个数。

变异操作：$\boldsymbol{A}'(k)\xrightarrow{\text{变异操作}}\boldsymbol{A}''(k)$

该操作是以概率将抗体某些基因位上的值进行变动，可以产生新的候选解。本文中每个基因位的变异范围为[1,N]，N为数据集大小。

交叉操作：$\boldsymbol{A}''(k)\xrightarrow{\text{交叉操作}}\boldsymbol{A}'''(k)$

为了充分利用种群中各个抗体的信息，对 $\boldsymbol{A}''(k)$ 进行均匀交叉操作。

选择操作：$\boldsymbol{A}(k)\cup\boldsymbol{A}'''(k)\xrightarrow{\text{选则操作}}\boldsymbol{A}(k+1)$

该操作分别对抗体各自克隆后的子种群经过变异交叉操作，选择出优秀的抗体，形成新种群。具体地，对 $[\boldsymbol{a}''_{i1}(k),\boldsymbol{a}''_{i2}(k),\cdots,\boldsymbol{a}''_{iq_i}(k)]$ 选择出亲和度最大的抗体 $\boldsymbol{a}_i^*(k)$，则取代 $\boldsymbol{a}_i(k)$ 的概率为：当满足 $f(\boldsymbol{a}_i(k))<f(\boldsymbol{a}_i^*(k))$ 时，概率为1；当 $f(\boldsymbol{a}_i(k))\geqslant f(\boldsymbol{a}_i^*(k))$ 且 $\boldsymbol{a}_i(k)$ 不是种群最优个体时，概率为 $\exp\left(-\dfrac{f(\boldsymbol{a}_i(k))-f(\boldsymbol{a}_i^*(k))}{\alpha}\right)$；当 $f(\boldsymbol{a}_i(k))\geqslant f(\boldsymbol{a}_i^*(k))$ 且 $\boldsymbol{a}_i(k)$ 是种群最优个体时，概率为0。

抗体表示各类簇最佳代表样本点，以流形距离为距离测度对数据集进行划分。则抗体 $\boldsymbol{a}$ 的亲和度值定义为：

$$f(\boldsymbol{a})=\frac{1}{\mathrm{eps}+\sum_{i=1}^{\mathrm{num}}\sum_{x_j\in C_i}M(x_j,\mu_i)} \tag{4}$$

式中，eps 表示一个较小的常数；μ_i 为类簇 C_i 的最佳代表点；$M(x_j,\mu_i)$ 为两点的流形距离。

3　流形距离的聚类有效性指标

前面介绍的免疫克隆聚类算法仅仅适用于聚类

类别数已知的情况，而不能由计算机自行判断正确的聚类类别数。因为由式(4)的分母可知，随着类别数的增大，整个函数值变得越来越大。为了自动确定聚类类别数，需要借助聚类有效性指标，它是一个统计学数学函数。本文设计一个可以评价流形结构分布的聚类有效性指标。该指标借助了CS指标[12]的框架：

$$\mathrm{CS}(\mathrm{num})=\frac{\sum_{i=1}^{\mathrm{num}}\left\{\frac{1}{N_i}\sum_{x_j\in C_i}\max_{x_k\in C_i}\left\{e(x_j,x_k)\right\}\right\}}{\sum_{i=1}^{\mathrm{num}}\left\{\min_{j=1,2,\cdots,\mathrm{num},\,j\neq i}\left\{e(v_i,v_j)\right\}\right\}} \tag{5}$$

式中，e 代表欧式距离；num 表示聚类类别数；v_i 表示类簇 C_i 的聚类中心；N_i 表示 C_i 的样本点个数。不难看出，式(5)的分子表示类簇内的紧凑程度，分母表示类簇间的分离程度，整个指标越小，聚类结果越好。

为了适应空间分布复杂的流形数据集，引入流形距离 M。但是利用计算聚类中心之间的流形距离来得到类簇间的分离程度是不合适的。因为引入聚类中心相当于添加了新的数据点，则流形距离每次都要重新计算，不但增加了计算复杂度，而且会影响数据的全局一致性；另外对于一个类簇位于另一个类簇的凸包情况，这种计算毫无意义。所以进一步改进CS指标来计算类簇间的分离程度：

$$M(C_i,C_j)=\frac{1}{N_i\times N_j}\sum_{x_p\in C_i}\sum_{x_q\in C_j}M(x_p,x_q) \tag{6}$$

式中，N_i 和 N_j 分别表示 C_i 和 C_j 的样本点个数；$M(x_p,x_q)$ 表示两点之间的流形距离。

综上所述，流形距离的聚类有效性指标(简称MDCVI)为：

$$\mathrm{MDCVI}(\mathrm{num})=\frac{\sum_{i=1}^{\mathrm{num}}\left\{\frac{1}{N_i}\sum_{x_j\in C_i}\max_{x_k\in C_i}\left\{M(x_j,x_k)\right\}\right\}}{\sum_{i=1}^{\mathrm{num}}\left\{\min_{j=1,2,\cdots,\mathrm{num},\,j\neq i}\left\{\frac{1}{N_i\times N_j}\sum_{x_p\in C_i}\sum_{x_q\in C_j}M(x_p,x_q)\right\}\right\}} \tag{7}$$

4 AICCMD方法实现策略

AICCMD方法在运行过程中由于要事先计算数据集点对之间的流形距离，其核心是使用迪杰斯特拉算法求解最短路径，所以针对图像数据集，如Berkeley图像，大小为321×481，如果对图像的每个像素点直接进行操作，则时间复杂度和空间复杂度是无法想象的。为了解决这一棘手的难题，本文采用SLIC超像素算法[13]对待分割图像进行预分割，得到了许多大小相近，内部较为匀质的小封闭区域，这些区域称为超像素。将超像素块看成是聚类原型再进行聚类，可大大降低运算的时间和空间复杂度，且可以降低噪声点对分割结果的影响。

算法具体实现如下：

1) 对待分割图像使用SLIC算法进行过分割，得到超像素块作为聚类操作单元；

2) 计算点对之间的流形距离，并输入最大聚类类别数 $\mathrm{num}_{\max}$；

3) 对聚类类别数 $\mathrm{num}=2$ 到 $\mathrm{num}=\mathrm{num}_{\max}$，重复运行步骤4)～步骤10)；

4) 按照当前的聚类类别数，随机初始化抗体种群 $\boldsymbol{A}(k)$，$k=0$；

5) 对抗体种群 $\boldsymbol{A}(k)$ 进行克隆操作，得到种群 $\boldsymbol{A}'(k)$；

6) 对种群 $\boldsymbol{A}'(k)$ 进行变异操作，得到 $\boldsymbol{A}''(k)$；

7) 对种群 $\boldsymbol{A}''(k)$ 进行交叉操作，得到 $\boldsymbol{A}'''(k)$；

8) 对 $\boldsymbol{A}(k)\bigcup\boldsymbol{A}'''(k)$ 利用式(4)计算亲和度值，进行选择操作，得到种群 $\boldsymbol{A}(k+1)$；

9) 判断是否满足算法迭代次数，如果不满足，则 $k=k+1$，返回步骤5)；如果满足，则输出最佳抗体及其聚类结果；

10) 按照本文提出的聚类有效性指标式(7)，计算聚类结果的有效性；

11) 选取具有最优聚类有效性指标值的聚类结果，其聚类类别数就是最佳聚类类别数；

12) 输出图像分割结果。

5 实验分析

5.1 人工数据集聚类实验

在图像分割实验前，为了理论上验证AICCMD算法的优势，首先对4个典型的人工数据集进行聚类，并对比经典的k-means算法(需要人工指定聚类类别数)、GCUK算法(一个经典的自动确定聚类类别数的遗传聚类算法)。实验中，各算法参数设置如下：AICCMD迭代次数2 000次，$\mathrm{num}_{\max}=8$，抗体种群大小40，抗体复制个数为5，变异概率0.5，交叉概率0.1；k-means迭代次数100次；GCUK种群大小50，交叉概率0.8，变异概率0.001，迭代次数2 000次。各个算法分别运行10次。

表1 数据集描述

数据集名称	数据集大小	聚类类别数
dataset1	266	3
dataset2	400	2
dataset3	400	4
dataset4	785	6

表1描述了4个人工数据集，分别给出了这4个人工数据集的大小和正确的聚类类别数。

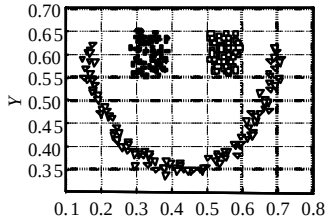
5.1.1　定量比较分析

表2为3种方法得到的聚类类别数均值和聚类结果的ARI值(adjusted rand index)[14]，它是一个定量评价聚类算法性能的指标，取值范围[0,1]，越大说明聚类效果越好，由表2看出，AICCMD算法明显优于其他算法。

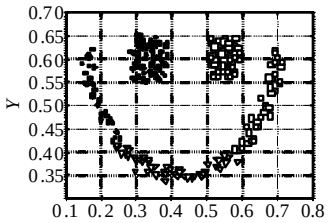
表2　不同算法在4个数据集的性能比较

数据集	AICCMD		k-means		GCUK	
	聚类类别数	ARI指标	聚类类别数	ARI指标	聚类类别数	ARI指标
dataset1	3.0	1.00	3.0	0.40	7.1	0.62
dataset2	2.0	1.00	2.0	0.19	4.3	0.24
dataset3	4.0	1.00	4.0	1.00	4.0	1.00
dataset4	6.0	1.00	6.0	0.56	2.4	0.26

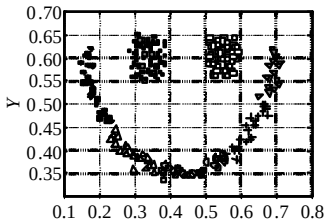
5.1.2　定性分析



a. AICCMD



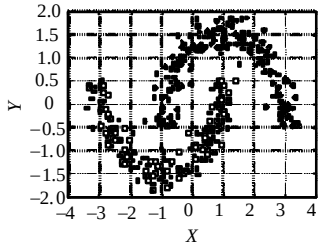
b. k-means



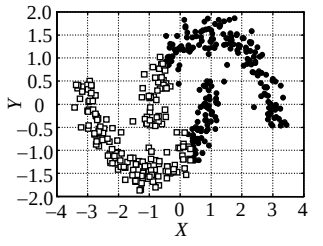
c. GCUK

图2　3种方法在dataset1上得到的聚类典型结果

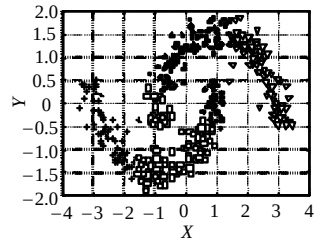
图2～图5直观展示了不同方法在4个人工数据集上得到的聚类结果。从k-means聚类结果看，虽然它需要人工输入正确聚类类别数，但是除了对dataset3这种球形分布的数据集效果好之外，对其他3个人工数据集效果都不好。GCUK算法也是除了对dataset3效果好之外，其他的3个人工数据集效果都不好。AICCMD则全部得到了正确的聚类结果。这就说明k-means算法和GCUK算法只能处理空间分布呈球形的数据集，而对复杂流形结构的数据集无能为力。AICCMD在无需人工事先指定类别数的情况下既能处理球形数据集，也能处理复杂流形结构的数据集。



a. AICCMD

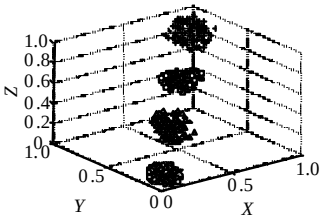


b. k-means

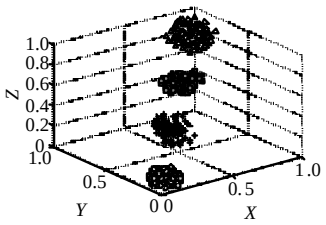


c. GCUK

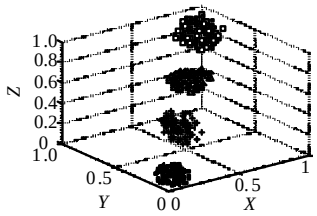
图3　3种方法在dataset2上得到的聚类典型结果



a. AICCMD

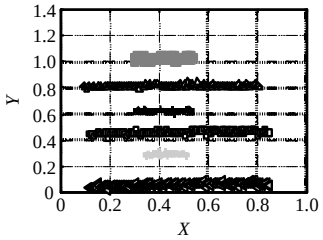


b. k-means

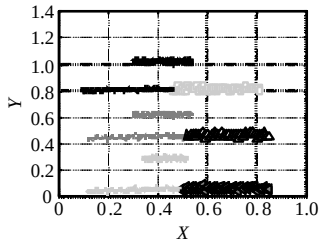


c. GCUK

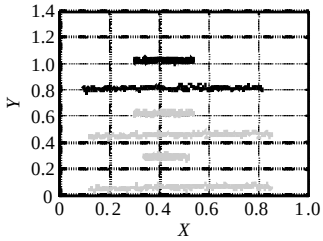
图4 3种方法在dataset3上得到的聚类典型结果



a. AICCMD



b. k-means



c. GCUK

图5 3种方法在dataset4上得到的聚类典型结果

5.2 图像分割实验

本文选取伯克利图像数据库(Berkeley image database)的4幅具有代表性的图像作为待分割图像，大小均为321×481。算法的各个参数和人工数据集相同。使用RGB及HSI颜色空间作为特征空间。采用SLIC超像素方法对图像进行预分割。实验中，超像素数量为800左右，数量太少会导致超像素内包含不同的目标，太多则使运算时间太长。

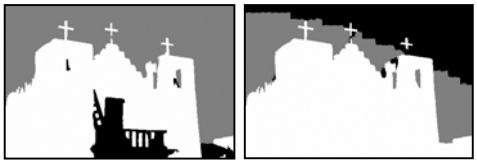
对于图6a所示原图，该图包含3类目标：天空、建筑物的墙体和建筑物的楼梯。图6b为SLIC算法得到的超像素结果。图6c～图6e分别为AICCMD算法、k-means算法、GCUK算法的分割结果。如图6d所示，k-means算法(人工给定聚类数目为3)分割效果很差，将楼梯这一目标错分成墙体；图6e GCUK算法虽然自动分为了3类，但是将墙体的一部分十字架错分成天空。图6c的AICCMD算法不但分割类别数正确，而且几乎没有错分。

对于图7a所示原图，目视包含2类目标：天空、沙漠中的金字塔。图7b为SLIC算法得到的超像素结果；图7d的k-means算法对2类目标都有严重错分；图7e的GCUK算法自动分成了3类，对天空和金字塔都产生错分；图7c的AICCMD算法分成了2类，分割视觉效果良好。

对于图8a所示原图，该图包含了3类目标：海星、海床上的绿苔、灰白色的礁石。图8b为SLIC算法得到的超像素结果；图8d的k-means算法将海星错分成了绿苔，并将礁石错分成海星；图8e的GCUK算法自动分成了2类；图8c的本文方法不但分对了类别数，而且错分很少。

对于图9a所示原图，包含4类目标：黑色背景、绿叶、红花、黄色花蕊。图9b为SLIC算法得到的超像素结果；图9d的k-means的结果对绿叶和花蕊有较严重错分；图9e的GCUK将图像分为4类，效果较k-means有所提高；图9c的本文方法自动分成了5类，但是视觉效果较好，特别是对绿叶、红花的分割。

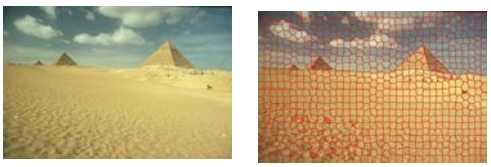
a. 原图 b. SLIC结果



c. AICCMD d. k-means

e. GCUK

图6 图像1分割结果对比

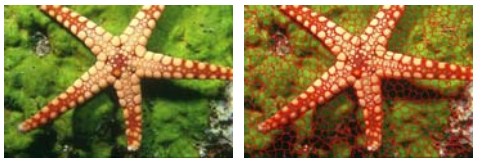


a. 原图 b. SLIC结果

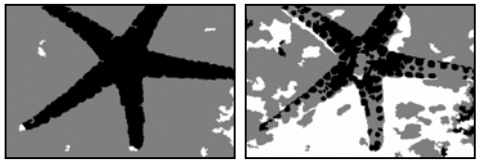
c. AICCMD　　d. k-means

e. GCUK

图7　图像2分割结果对比



a. 原图　　b. SLIC结果

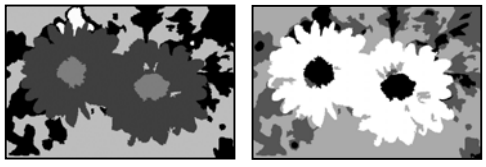


c. AICCMD　　d. k-means

e. GCUK

图8　图像3分割结果对比

a. 原图　　b. SLIC结果



c. AICCMD　　d. k-means

e. GCUK

图9　图像4分割结果对比

6　结　论

本文提出了一个流形距离的自动免疫克隆聚类图像分割算法。通过在典型人工数据集和Berkeley图像上分别进行测试，并对比经典的k-means算法和GCUK算法，根据定性和定量的分析，得出本文方法优势明显，具有一定的实用价值和推广价值。今后将考虑如何在图像的特征向量中加入空间位置信息，以进一步提高算法的分割性能。

参　考　文　献

[1] CHENG H D, JIANG X H, SUN Y, et al. Color image segmentation: Advances and prospects[J]. Pattern Recognition, 2001, 34(12): 2259-2281.

[2] TRANOS Z, OLUDAY O O, SUNDAY O O, et al. Image segmentation, available techniques, development and open issues[J]. Canadian Journal on Image processing and Computer Vision, 2011, 2(3): 20-29.

[3] JAIN A K. Data clustering: 50 years beyond k-means[J]. Pattern Recognition Letters, 2010, 31(8): 651-666.

[4] HALDER A, PATHAK N. An evolutionary dynamic clustering based on colour image segmentation[J]. International Journal of Image Processing, 2011, 4(6): 549-556.

[5] LIU Ruo-chen, JIAO Li-cheng, ZHANG Xiang-rong, et al. Gene transposon based clone selection algorithm for automatic clustering[J]. Information Science, 2012(204): 1-22.

[6] SANGHAMITRA B, UJJWAL M. Genetic clustering for automatic evolution of clusters and application to image classification[J]. Pattern Recognition, 2002, 35(6): 1197-1208.

[7] 公茂果, 焦李成, 马文萍. 基于流形距离的人工免疫无监督分类与识别算法[J]. 自动化学报, 2008, 34(3): 367-375.
GONG Mao-guo, JIAO Li-cheng, MA Wen-ping. Unsupervised classification and recognition using an artificial immune system based on manifold distance[J]. Acta Automatic Sinica, 2008, 34(3): 367-375.

[8] CHANG Hong, YEUNG D Y. Robust path-based spectral clustering[J]. Pattern Recognition, 2008, 41(1): 191-203.

[9] JUNG C, JIAO Li-cheng. Image segmentation via manifold spectral clustering[C]//IEEE International Workshop on Machine Learning for Signal Processing. [S.l.]: IEEE, 2011: 1-6.

[10] 柴争义, 陈亮, 朱思峰, 等. 认知无线网络中基于免疫克隆优化的功率分配[J]. 电子科技大学学报, 2013, 42(1): 36-40.
CHAI Zheng-yi, CHEN Liang, ZHU Si-feng, et al. Power allocation of cognitive wireless network based on immune clonal optimization[J]. Journal of University of Electronic Science and Technology of China, 2013, 42(1): 36-40.

[11] LOU H, MAO C, WANG D, et al. PWM optimization for three-level voltage inverter based on clonal selection algorithm[J]. IET Electric Power Application, 2007, 1(6):

870-878.
[12] CHOU C H, SU, M C, LAI E. A new cluster measure and its application to image compression[J]. Pattern Analysis and Applications, 2004, 7(2): 205-220.
[13] ACHANTA R, SHAJI A, SMITH K, et al. SLIC superpixels compared to state of the art superpixel methods[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34(11): 2274-2282.
[14] HANDL J, KNOWLES J. An evolutionary approach to multiobjective clustering[J]. IEEE Transactions on Evolutionary Computation, 2007, 11(1): 56-76.

编 辑 税 红