

## 基于清晰半径的模糊点二次聚类算法

高翠芳\*, 胡 权

(江南大学 理学院, 江苏 无锡 214122)

(\* 通信作者电子邮箱 cuifang\_gao@163.com)

**摘 要:**针对模糊 C-均值(FCM)聚类算法在模糊边界上容易出现划分错误的问题,提出一种对模糊点进行二次处理的改进算法。该算法以各类中的数据分布密度为依据,首先利用清晰点构成超球体中心区域,然后基于中心区域的清晰半径定义一种新的相似性距离,并利用该距离对模糊点的隶属度进行二次计算,重新确定其类别归属。实验结果显示,改进算法能有效纠正分类错误,提高模糊点的清晰度,在密度差异较大的数据集上具有一定的应用潜力。

**关键词:**模糊聚类;模糊点;相似性距离;中心区域;二次聚类

**中图分类号:** TP311 **文献标志码:** A

### Second clustering algorithm for fuzzy points based on clear radius

GAO Cuifang\*, HU Quan

(School of Science, Jiangnan University, Wuxi Jiangsu 214122, China)

**Abstract:** Concerning the problem of wrong partition at fuzzy boundary in Fuzzy C-Means (FCM) clustering algorithm, an improved recalculation technique for fuzzy points was proposed. The new method took into account the data distribution characteristics in different classes. Firstly, it made the hyperspheres central regions by clear data, then defined a new similarity distance based on the clear radius of central region to recalculate the membership of fuzzy point, and finally reassigned the fuzzy points to right category. The experimental results show that the new algorithm can correct some wrong partition and improve the definition of fuzzy point, and also it is a promising algorithm for dataset with significant density differences.

**Key words:** fuzzy clustering; fuzzy point; similarity distance; central region; second clustering

## 0 引言

聚类是按照一定的相似性原则将数据对象的集合分为若干类或簇,使得同一类中的对象具有较高的相似度,而不同类中的对象之间差别较大<sup>[1-2]</sup>。传统的硬聚类算法通常得出确定的结果,就是把数据对象严格地划分到某个具体的类别中,具有非此即彼的性质。但是许多客观事物之间的界限往往不是很清晰,很多对象在性质和类属上存在中介性<sup>[3]</sup>。为了能真正反映数据对象和类之间的模糊划分关系,人们把 Zadeh 的模糊集理论<sup>[4]</sup>引入到传统硬聚类算法中,发展成为模糊聚类,其特点是数据对象不再属于某个确定的类,而是以不同的隶属度属于每个类。模糊聚类建立了对对象的不确定性描述,能够比较客观地反映现实世界,更加符合自然规律。

经典模糊 C-均值(Fuzzy C-Means, FCM)聚类算法<sup>[5]</sup>是目前研究最广泛的算法之一。它从一个初始划分开始,需要预先定义一个最优化聚类标准也就是目标函数,作为度量各类对象分布的代价函数,在隶属度约束条件下求得目标函数最小值,从而获得最优模糊划分结果。FCM 算法原理简单,计算效率高,相似性度量运用灵活,近年来国内外学者进行了持续不断的研究<sup>[6-9]</sup>。但是,该算法中每类只用一个简单的类中心来代表,不能反映不同区域中数据点分布的紧密程度,导致边界上一些较难确定的模糊点出现分类错误。密度指标能有效处理数据点分布不均的情况<sup>[10-11]</sup>,在一些引入密度概念的聚类算法中,可以用来确定聚类中心的位置<sup>[12]</sup>,或根据密

度差异形成不同的聚类区域<sup>[13-15]</sup>。因此,如果以各类的紧密程度为依据,对 FCM 中类属模糊的点进行重新划分,有望得到更符合实际结构的聚类结果。

本文针对 FCM 算法的聚类结果中存在的部分模糊点,提出了一种模糊点的二次聚类算法。该算法引入了基于清晰半径的模糊点二次处理算法,利用新的相似性距离,对模糊点的隶属度重新计算,可以使模糊点的隶属度更加清晰,从而重新确定其归属类别。

## 1 模糊 C-均值聚类算法

FCM 是用隶属度确定聚类程度的一种聚类算法<sup>[5]</sup>。它把  $n$  个向量  $x_i (i = 1, 2, \dots, n)$  分为  $c$  个类别,用每类的聚类中心代表该类,聚类中心为  $V = \{v_1, v_2, \dots, v_c\}$ 。设  $u_{ik}$  表示第  $i$  个数据点属于第  $k$  类的隶属度,隶属度越大表示数据对象属于该类的程度越高,隶属度越小表示属于该类的程度越小。FCM 算法的目标函数为:

$$J(U, V) = \sum_{i=1}^n \sum_{k=1}^c u_{ik}^m d_{ik}^2$$

其中:  $u_{ik} \in [0, 1]$ ,  $\sum_{k=1}^c u_{ik} = 1 (i = 1, 2, \dots, n)$ ;  $d_{ik} = \| \cdot \|$  为样本点与类中心之间的相似性距离;  $m$  是模糊参数。FCM 算法用迭代方法求得  $U$  和  $V$  的最优值,使得  $J$  取值最小,遵循如下步骤:

1) 确定聚类数目  $c$  和参数  $m$ ;

收稿日期: 2012-08-06; 修回日期: 2012-09-21。 基金项目: 中央高校基本科研业务费专项资金资助项目(JUSRP211A23, JUSRP31103)。

作者简介: 高翠芳(1974-), 女, 河北石家庄人, 讲师, 博士, 主要研究方向: 模式识别、生物信息学; 胡权(1991-), 女, 湖南邵阳人, 主要研究方向: 模式识别。

- 2) 随机初始化聚类中心  $V$ ;
- 3) 利用下面的迭代公式计算隶属度:

$$u_{ik} = \frac{(\|x_i - v_k\|)^{-\frac{2}{m-1}}}{\sum_{r=1}^c (\|x_i - v_r\|)^{-\frac{2}{m-1}}}$$

- 4) 利用下面的迭代公式计算聚类中心:

$$v_k = \sum_{i=1}^n u_{ik}^m x_i \cdot \left( \sum_{i=1}^n u_{ik}^m \right)^{-1}$$

5) 重复步骤3)、4), 进行反复迭代运算, 直到目标函数收敛到最小值, 完成模糊划分。

## 2 基于清晰半径的模糊点二次聚类算法

FCM 算法优化计算后可以得到所有数据对象的隶属度, 但是不可避免地在边界上出现了类属不明确(隶属度之间相差不大)的现象, 即有些数据点的分类比较模糊。因此, 对这些模糊点进行二次处理, 提出一种新的相似性距离算法, 把原来只用一个中心点代表的类扩展为用中心区域代表, 重新计算模糊点的隶属度。

如图1所示,  $v_1, v_2, v_3$  分别为三个聚类中心点,  $L_1, L_2, L_3$  为类中心区域的超球体半径, 用来反映各类数据点分布的紧密程度。假设  $a$  是其中某一个模糊点, 在 FCM 算法中,  $a$  对于  $v_1, v_2$  两个聚类中心的隶属度相差不大, 如果用一个类中心代表该类,  $a$  到两个类中心的距离  $d(a, v_1) < d(a, v_2)$ , 则  $a$  点应划分到  $v_1$  类中。但实际的数据分布特点是,  $v_1$  类中的结构较紧凑,  $v_2$  类中较松散, 故超球体半径  $L_2 > L_1$ , 此时数据点到类中心的距离已不能作为模糊点类别归属的准确依据。于是把对于模糊点  $a$  的算法修改为: 求出  $a$  到三个中心区域边缘的相似性距离(等于  $a$  到三个聚类中心的距离减去超球体半径), 如图中  $\tilde{d}_1, \tilde{d}_2, \tilde{d}_3$  的长度, 然后把这个新的相似性距离代入隶属度公式中, 重新求得  $a$  点相对于三类的隶属度, 最大隶属度对应的聚类中心即为新算法下模糊点  $a$  的类别归属。其中超球体半径可由各类中清晰点到其聚类中心的平均距离计算得到, 本文定义为清晰半径。

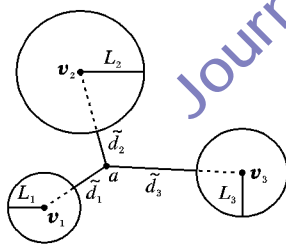


图1 基于类中心和清晰半径的两种相似性距离的差异

改进算法的主要步骤如下:

- 1) 根据 FCM 算法得到的初次聚类结果, 确定最大隶属度小于 0.65 的点为模糊点, 大于 0.9 的为清晰点, 其余为次清晰点。
- 2) 根据条件, 将原数据集划分为模糊点集  $\varphi$ 、清晰点集  $\tau$  和次清晰点集  $\psi$ 。
- 3) 计算清晰半径  $L_k$ , 即集合  $\tau$  中各清晰点到其所属聚类中心的平均距离, 清晰半径范围内的超球体构成中心区域。

$$L_k = \frac{1}{N_k} \sum_{i=1}^{N_k} d_{ik}$$

其中  $d_{ik} = \|x_i - v_k\|$ ,  $k = 1, 2, \dots, c$ ,  $x_i \in \tau$ ,  $N_k$  为第  $k$  类中清晰点的个数。

- 4) 利用如下所示的新的相似性距离公式计算集合  $\varphi$  中模糊点到中心区域边缘的距离:

$$\tilde{d}_{ik} = d_{ik} - L_k; k = 1, 2, \dots, c, x_i \in \varphi$$

- 5) 利用如下修正的隶属度公式计算集合  $\varphi$  中模糊点的新隶属度, 并对新隶属度进行归一化处理:

$$u_{ik} = \frac{[\tilde{d}_{ik}^2]^{-1}}{\sum_{r=1}^c [\tilde{d}_{ir}^2]^{-1}}$$

- 6) 最终隶属类别的确定: 集合  $\tau$  和集合  $\psi$  中的隶属度保持不变, 集合  $\varphi$  中模糊点的隶属度根据第 5) 步重新确定。
- 7) 综合集合  $\tau, \psi, \varphi$  中的聚类结果, 得出最终分类结果。

算法结构如图2所示。

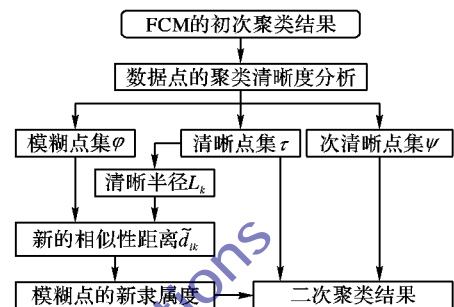


图2 模糊点二次聚类算法结构

本文算法对 FCM 的改进和区别主要体现在两个方面: 一方面引入了新的相似性距离公式, 对于模糊点来说, 本文算法的类由清晰半径确定的中心区域表示 (FCM 算法由一个简单的中心点表示), 基于不同的类中心 (点或区域) 产生了不同的距离公式; 另一方面, 清晰半径可以度量各类中数据点分布密度的不同, 实际上相当于引入了密度指标, 因此改进算法可以更好地适应数据分布结构不均匀的数据集。同时本文算法保留了 FCM 中一定数量的隶属度信息, 并完善了模糊点的二次处理结果, 使最终的划分矩阵更加清晰和准确。

## 3 实验与分析

用随机产生的二维人工数据对本文算法进行验证, 原始数据点分布如图3所示。数据集共包含 90 个数据点, 聚类数  $c = 3$ , 图3中的  $\square, \triangle$  和  $\circ$  分别表示第一、二、三类, 每类有 30 个数据点, 相邻两类之间的数据密度相差较大, 其中第一类和第三类中数据点分布较松散, 第二类分布较紧凑, 位于两类边界上的模糊点分类难度较大。

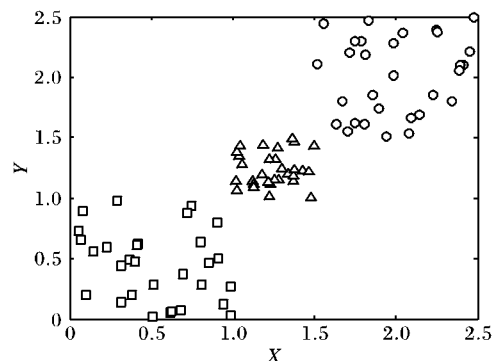


图3 原始数据点分布

图4是 FCM 算法在该数据集上得到的聚类结果, FCM 用中心点  $v_1, v_2, v_3$  分别代表三个类, 采用统一的划分标准 (即数据点到类中心的相似性距离) 对数据集聚类, 部分位于一、三两类边界上的点, 由于距离  $v_2$  点较近, 在没有考虑分布密度的情况下, FCM 算法把这些点错误地划分到了第二类中。

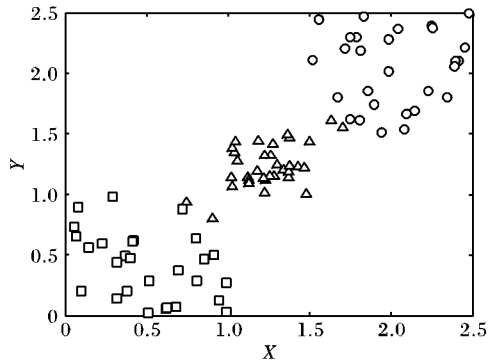


图4 FCM分类结果

根据FCM算法的初次结果,经过聚类清晰度分析以后,找出了分类边界上的模糊点( $A \sim H$ ),如图5,虚线范围是每类的中心区域,也就是本文所提出的新相似性距离的依据。从图5中可以看出,第一类和第三类的清晰半径明显大于第二类,与数据点的实际分布密度一致。用这三个中心区域代替中心点 $v_1, v_2, v_3$ ,采用本文算法对模糊点集 $\varphi$ 中的隶属度进行二次计算,得到数据集的重新分类结果,如图6。不难看出二次聚类后的结果,不但改正了FCM算法中存在的聚类错误,而且更加符合数据集的实际分布情况。

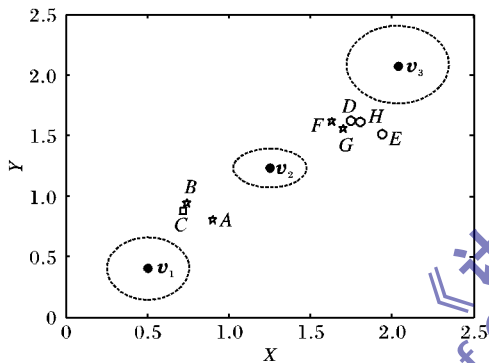


图5 模糊点以及各类的中心区域

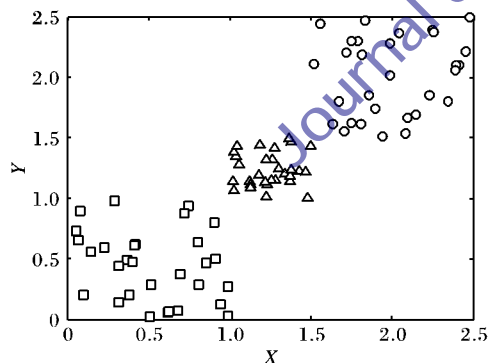


图6 新算法的分类结果

表1中给出了模糊点的两种相似性距离的具体数值,可以看出 $\tilde{d}_{ik}$ 小于相应的 $d_{ik}$ (到中心区域的距离小于到中心点的距离);更重要的是由于考虑了密度差异,部分点的 $\tilde{d}_{ik}$ 的最小值已经与 $d_{ik}$ 相差较大(如点 $A, B, F, G$ ),这种改变必将影响到隶属度的改变以及最终划分结果。

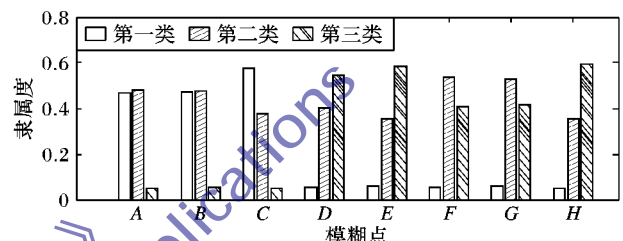
为了进一步考察模糊点的隶属度变化情况,分别采用两种相似性距离来计算隶属度,得到图7中的对比结果。可以看出,模糊点在FCM算法和本文算法下的隶属度区别非常明显。

图7(a)显示,初次结果中模糊点的隶属度都相差不大,一些距离 $v_2$ 较近的模糊点,如 $A, B, F, G$ 点,在FCM算法下的

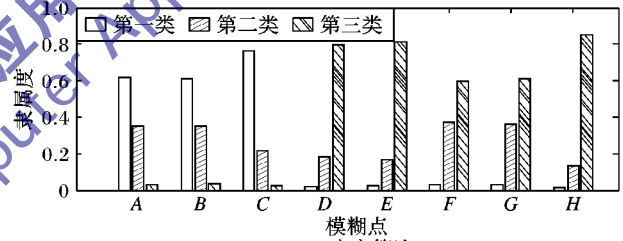
分类结果是错误的,但是经过对其隶属度的重新计算,实现了正确划分。还有些模糊点,如 $C, D, E, H$ 点,它们的分类本来就是正确的,通过二次聚类算法,可以使这些点的清晰度变强,最大隶属度均超过0.75。

表1 模糊点的两种相似性距离

模糊点 $x_i$	模糊点到类中心的距离			模糊点到中心区域的距离		
	$d_{i1}$	$d_{i2}$	$d_{i3}$	$\tilde{d}_{i1}$	$\tilde{d}_{i2}$	$\tilde{d}_{i3}$
A	0.5660	0.5590	1.7062	0.3081	0.4087	1.3946
B	0.5884	0.5859	1.7198	0.3305	0.4356	1.4083
C	0.5197	0.6442	1.7832	0.2618	0.4939	1.4716
D	1.7431	0.6283	0.5400	1.4852	0.4780	0.2284
E	1.8157	0.7425	0.5771	1.5578	0.5922	0.2655
F	1.6558	0.5343	0.6154	1.3979	0.3840	0.3038
G	1.6635	0.5512	0.6202	1.4057	0.4009	0.3086
H	1.7783	0.6687	0.5182	1.5204	0.5184	0.2066



(a) FCM算法



(b) 本文算法

图7 模糊点分别在FCM算法和本文算法下的隶属度

还采用了UCI机器学习库中的标准数据集Iris对本文算法进行检验,并与基础算法FCM以及基于密度的聚类算法(MDCA)<sup>[14]</sup>进行了性能比较,三种算法均采用欧几里得距离,结果如表2。其中Iris有4维,含有150个数据点,分为3个类别。

表2 三种聚类算法的运算时间和聚类准确率

聚类算法	运算时间/ms	准确率/%
FCM	8	86.00
MDCA	60	87.67
本文算法	24	89.33

本文算法是在FCM的基础上引入了清晰半径,本质上还是属于划分式聚类算法,因此保持了划分式聚类的速度优势;但是由于增加了数据集拆分、清晰半径计算、以及模糊隶属度计算等二次处理过程,在运行时间上比FCM算法有所增加。而基于密度的MDCA算法需要计算两两数据点之间的距离,比基于划分的聚类算法运行速度慢了很多。

表2中本文算法和MDCA算法的聚类准确率的提高幅度均不大,主要是由于Iris数据集的密度差别不够明显,通常情况下密度差异越大,基于密度的聚类算法的处理效果会越好。即使如此,对于没有密度差异的数据集,本文算法也是适用的。

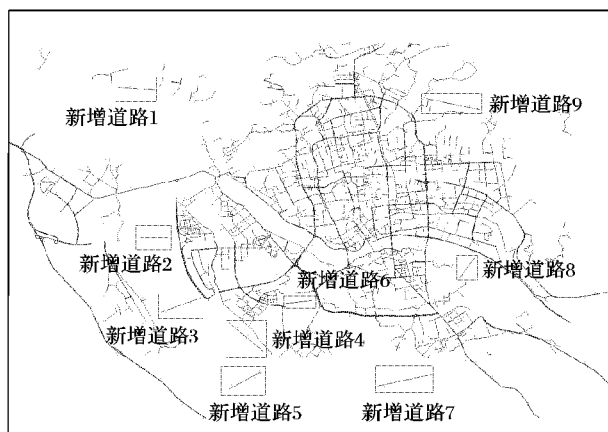


图9 新增道路自动发现效果

#### 4 结语

本文提出通过将浮动车的移动轨迹生成图与路网图层生成图进行图像配准,从而可快捷地发现新增道路,并自动地生成包含其位置和长度等信息的疑似新增道路报表方便后期处理。该方法有效解决了基于遥感影像的新增道路发现方法存在的背景图像干扰问题,同时也有效解决了目前利用浮动车的移动轨迹进行新增道路检测方法存在的需要人工提取新增道路信息的问题,具有良好的应用前景。

##### 参考文献:

- [1] 石善斌,吕志平,陈华远,等. 车载GPS道路测量数据处理技术[J]. 测绘科学技术学报, 2006, 23(4): 275-277, 283.
- [2] 朱丽云,温慧敏. 交通路网数据自动增量识别与技术更新[J]. 交通信息与安全, 2009, 27(2): 22-24, 55.
- [3] WEIS M, MÜLLER S, LIEDTKE C-E, *et al.* A framework for GIS and imagery data fusion in support of cartographic updating[J]. Information Fusion, 2005, 6(4): 311-317.

(上接第549页)

#### 4 结语

针对FCM算法中存在的部分模糊点,本文提出了一种对模糊点进行二次分类的新算法。该算法在FCM形成初始聚类的基础上,引入一种基于中心区域的相似性计算公式,对其中不能清晰确定类别归属的模糊点进行重新计算。本文算法以数据点分布的紧密程度为理论依据,具有较直观的几何意义,实验结果显示,它能有效纠正分类错误,提高模糊点的清晰度,对模糊点的隶属度处理提高了FCM算法的精确度。该算法适用于各类数据点密度结构不同的数据集,在模糊聚类中具有一定的推广性。在今后的研究工作中,可以尝试对其他模糊聚类算法中产生的模糊点进行二次处理,以期进一步提高基础算法的性能。

##### 参考文献:

- [1] HAN J, KAMBR M. 数据挖掘概念与技术[M]. 范明,孟小峰,译. 北京:机械工业出版社, 2002: 223-224.
- [2] OMRAN M G H, ENGELBRECHT A P, SALMAN A. An overview of clustering methods[J]. Intelligent Data Analysis, 2007, 11(6): 583-605.
- [3] 高新波. 模糊聚类分析及其应用[M]. 西安:西安电子科技大学出版社, 2004.
- [4] ZADEH L A. Fuzzy sets[J]. Information and Control, 1965, 8(3): 338-353.

- [4] ALBOODY A, SEDES F, INGLADA J. Post-classification and spatial reasoning new approach to change detection for updating GIS database[C]// ICTTA 2008: The 3rd International Conference on Information and Communication Technologies: from Theory to Applications. Piscataway: IEEE, 2008: 1-7.
- [5] 张韵,李清泉,曹晓航,等. 一种道路网信息几何差异检测算法[J]. 测绘学报, 2008, 37(4): 521-525.
- [6] ARAFAT S Y, BUTT A Y, LIAQAT N. Automatic road detection using MCSC[C]// The 14th IEEE International Multitopic Conference (INMIC). Karachi, Pakistan, IEEE, 2011: 126-131.
- [7] 刘志青,郭海涛,张保明,等. 一种基于直线特征的遥感影像和GIS数据自动整体配准方法[J]. 测绘科学技术学报, 2011, 28(2): 129-133.
- [8] MORIYA M, TAKEUCHI S, SHIKADA M, *et al.* Establishment of map update technique for local government by using GPS[C]// IGARSS 2008: IEEE International Geoscience and Remote Sensing Symposium. Piscataway: IEEE, 2008, 4: 447-450.
- [9] NIU Z, LI S N, POUSAID N. Road extraction using smart phones GPS[C]// COM. Geo 2011: The 2nd International Conference on Computing for Geospatial Research and Applications. New York: ACM, 2011: Article No. 22.
- [10] SUGAWARA H, SATO N, TAKAYAMA T, *et al.* Early evaluation of road width estimation on rapid road map survey system using GPS trajectories as collective intelligence[C]// NBIS 2011: The 14th International Conference on Network-Based Information Systems. Piscataway: IEEE, 2011: 547-552.
- [11] HUANG W D. Research on data process of GPS in road surveying[J]. Advanced Materials Research, 2012, 433-440: 6007-6013.
- [12] ZHAO Y, LIU J, CHEN R Q, *et al.* A new method of road network updating based on floating car data[C]// IEEE International Geoscience and Remote Sensing Symposium. Piscataway: IEEE, 2011: 1878-1881.

- [5] BEZDEK J C, EHRLICH R, FULL W. FCM: the fuzzy C-means clustering algorithm[J]. Computer and Geoscience, 1984, 10(2/3): 191-203.
- [6] XU R, WUNSCH D, II. Survey of clustering algorithms[J]. IEEE Transactions on Neural Networks, 2005, 16(3): 645-678.
- [7] YIN Z H, TANG Y G, SUN F C, *et al.* Fuzzy clustering with novel separable criterion[J]. Tsinghua Science and Technology, 2006, 11(1): 50-53.
- [8] WU K-L, YU J, YANG M-S. A novel fuzzy clustering algorithm based on a fuzzy scatter matrix with optimality test[J]. Pattern Recognition Letters, 2005, 26(5): 639-652.
- [9] ZHANG D Q, CHEN S C. Clustering incomplete data using kernel based fuzzy C-means algorithm[J]. Neural Processing Letters, 2003, 18(3): 155-162.
- [10] MAUCEFI C, HO D. Clustering by kernel density[J]. Computational Economics, 2007, 29(2): 199-212.
- [11] 石陆魁,何丕廉. 一种基于密度的高效聚类算法[J]. 计算机应用, 2005, 25(8): 1824-1826.
- [12] 薛磊,白康生,孙玉强,等. 两阶段模糊聚类算法在气测资料解释中的应用[J]. 计算机工程与设计, 2009, 30(4): 1027-1029.
- [13] 吕宗磊,王建东. 一种基于多维空间超球体的快速聚类算法[J]. 南京航空航天大学学报, 2006, 38(6): 706-711.
- [14] 王晶,夏鲁宁,荆继武. 一种基于密度最大值的聚类算法[J]. 中国科学院研究生院学报, 2009, 26(4): 539-548.
- [15] 马帅,王腾蛟,唐世渭,等. 一种基于参考点和密度的快速聚类算法[J]. 软件学报, 2003, 14(6): 1089-1095.