聚类中心的初始化方法 [1]

裴继红 范九伦 谢维信 *

(西安电子科技大学电子工程学院 西安 710071)
*(深圳大学校长办公室 深圳 518060)

摘 要 本文对用于聚类中心初始化的势函数的几个参数选择问题进行了讨论，给出了这些参数的两种形式。同时提出了一种新的使用密度函数法进行聚类中心初始化的方法，进行了一组对比实验，得到了令人满意的结果。

关键词 聚类，初始化，势函数，密度函数，非线性优化
中图号 TP391.4

1 引 言

聚类方法是模式分类和系统建模的基本方法之一。聚类的目的就是根据某种准则，将样本空间中的样本数据集合划分为可以用来表示系统行为的一些子集。聚类算法，尤其是模糊C-均值 (FCM) 聚类算法 [1]，更是得到了深入的研究和广泛的应用。FCM 算法是一个使目标函数 (1) 式最小化的迭代收敛优化过程。

$$J=\sum_{k=1}^{n}\sum_{i=1}^{C}\mu_{ik}^{m}\|x_k-v_i\|^2, \tag{1}$$

其中 n 是样本集合中样本的个数，C 是要划分的类数，x_k 是第 k 个数据样本，v_i 是第 i 个类的类中心矢量，μ_{ik} 是第 k 个数据样本属于第 i 个类的隶属程度，m 是一个大于 1 的常数 (m 是控制隶属度模糊性的一个指数)。给定类数 C 和初始中心位置 v_i, $i=1,\cdots,C$, FCM 算法收敛到目标函数 (1) 式的一个局部极小点或一个鞍点 [2]。

像大多数非线性优化问题一样，聚类结果的好坏受初始值选择的影响很大。不合适的初始值，可能导致结果收敛到一个不希望的极小点或者导致聚类过程收敛很慢。在聚类优化过程中，目标函数的每个局部极小点周围都有一个吸引域，如果选择的初始值处在吸引域中距离吸引子很近的位置，则优化过程很快收敛到该极值点；反之，收敛速度就很慢。如果初始值落在吸引域以外，则优化过程可能收敛到其他局部极小点上。

比较常用的初始化方法是在数据样本的特征空间 R^s 中随机选取 C 个矢量 ($v_i\in R^s$, $i=1,\cdots,C$) 作为初始中心进行聚类。这样选取的初始聚类中心使得聚类结果进入全局最优的概率较小，同时还易出现死点问题。所谓死点就是由于不合适的初始化，使得某些初始化中心值总是在竞争中失败，而出现空类的现象。采用随机选择 C 个样本点作为初始聚类中心，可以解决死点问题，但是不会提高聚类结果进入全局最优的概率。一般认为比较合适的聚类中心是出现在样本点比较密集的地方。根据这一观点，Yager 和 Filev[3] 提出了一种称为爬山法的初始化聚类中心方法。该方法是先在数据样本空间中构造一个网格，然后根据样本点到每个网格的距离，分别算出每个网格点的势函数值，如果网格点周围样本点多，

[1] 1997-09-11 收到，1998-07-01 定稿
国家自然科学基金资助项目

则势值就高。然后将势值最高的网格点选为第一个初始聚类中心。一旦第一个初始聚类中心选定以后，所有网格点的势根据网格点距离第一个初始聚类中心的距离做相应的调整，离第一个初始聚类中心越近的网格点，其势减小得越大。而下一个初始聚类中心选在调整后的势值最大的网格点上，然后再对所有网格点的势进行调整，寻找下一个初始聚类中心，$\cdots$。该方法尽管简单、有效，但是其计算量却是随样本维数呈指数增长，例如在样本空间每一维，取 10 条网格线，那么数据样本为 2 维、 3 维、 4 维 $\cdots$ 时，要计算的网格点数分别为 $10^2, 10^3, 10^4 \cdots$。 Chiu[4] 提出了一种改进的爬山法估计聚类中心。其方法的关键为：将计算每个网格点的势，改为计算每个数据样本点上的势，这样要计算的势点就等于样本点的数目，而与样本的维数无关。另外一个优点是消除了由构造网格而引起的计算精度与计算复杂度之间的矛盾。在 Chiu 的势函数方法中，几个关键性参数的选取不很合理，给使用造成了一定的困难，同时由于采用指数运算，使得运算量较大。本文对 Chiu 方法中参数的选取进行了研究，同时提出了一种使用密度函数的初始化方法。密度函数方法性能和势函数法相同，但是计算简单，运算量较势函数小得多。

2　聚类中心初始化的势函数法和密度函数法

对于一个 S 维空间的具有 n 个样本的数据样本集合 $\{x_1, x_2, \cdots, x_n\}$， Chiu 定义的样本点 x_i 处的势函数为

$$P_i^{(0)} = \sum_{j=1}^{n} e^{-\alpha \|x_i - x_j\|^2}, \tag{2}$$

其中

$$\alpha = 4/r_\alpha^2, \tag{3}$$

r_a 是一个正常数，表示邻域半径，在邻域半径 r_α 之外的数据点对势的计算影响很小。从以上关系式，我们看到，点 x_i 周围聚集的样本点越多，则点 x_i 的势就越高。令 $P_1^* = \max\{P_i^{(0)}, i = 1, \cdots, n\}$，同时取对应的 x_1^* 为第一个初始聚类中心位置，然后根据 (4) 式调整每个样本点的势：

$$P_i^{(1)} = P_i^{(0)} - P_1^* e^{-\beta \|x_i - x_1^*\|^2}, \tag{4}$$

其中

$$\beta = 4/r_\beta^2, \tag{5}$$

r_β 是一个正常数。令 $P_2^* = \max\{P_i^{(1)}, i = 1, \cdots, n\}$，取相应的 x_2^* 为第二个初始聚类中心位置。势函数调整的一般关系式如下：

$$P_i^{(k)} = P_i^{(k-1)} - P_k^* e^{-\beta \|x_i - x_k^*\|^2}, \quad k = 1, \cdots, c-1, \tag{6}$$

其中 $P_k^* = \max\{P_i^{(k-1)}, i = 1, \cdots, n\}$，对应的样本点 x_k^* 取为第 k 个初始聚类中心位置。

在上述方法中， Chiu 取定 r_α 、 r_β 为常数，但是，由于样本特征选择的不规范性，常常导致样本空间的不规范性，因此， r_α 、 r_β 的选择应该与数据样本集合的分布特性有关。

为此本文给出两种邻域半径，形式如下：

$$r_f = \frac{1}{2} \vee_{k=1}^{n} \wedge_{i=1}^{n} \|x_i - x_k\| = \frac{1}{2} \min\{\max\{\|x_i - x_k\|, i = 1, \cdots, n\}, k = 1, \cdots, n\}, \quad (7)$$

$$r_m = \frac{1}{2} \sqrt{\frac{1}{n(n-1)} \sum_{k=1}^{n} \sum_{i=1}^{n} \|x_i - x_k\|^2}, \quad (8)$$

其中 n 是数据集合的样本个数，$\max\{\cdot\}$ 、$\min\{\cdot\}$ 是集合求大求小函数．r_f 表示处于样本集合最中间的样本到距离它最远的样本之间的距离的二分之一，r_m 表示的是 n 个样本的均方根距离的二分之一．在具体应用中，可以选 $r_\alpha = r_\beta = r_f$ 或 $r_\alpha = r_\beta = r_m$ 代入 (3)，(5) 式中进行求解．

在 Chiu 给出的初始化方法中，势函数是以指数运算为基础的．这在样本量较大的情况下影响了速度，为此，定义样本点 x_i 处的密度函数如下：

$$D_i^{(0)} = \sum_{k=1}^{n} \frac{1}{1 + f_d \|x_i - x_k\|^2}, \quad (9)$$

其中

$$f_d = 4/r_d^2, \quad (10)$$

其中 r_d 可以取 (7) 式中的 r_f 或 (8) 式中的 r_m，是邻域密度有效半径．由 (9) 式，在 x_i 周围样本点越密集，则 $D_i^{(0)}$ 值越大，故 $D_i^{(0)}$ 可以用来表示在样本空间中样本点的密集程度．与势函数法相似，令 $D_1^* = \max\{D_i^{(0)}, i = 1, \cdots, n\}$，对应的 x_1^* 取为第 1 个初始聚类中心．求后续初始聚类中心的密度调整关系式如 (11) 式：

$$D_i^{(k)} = D_i^{(k-1)} - D_k^* \frac{1}{1 + f_d \|x_i - x_k^*\|^2}, \quad k = 1, \cdots, c-1, \quad (11)$$

其中 $D_k^* = \max\{D_i^{(k-1)}, i = 1, \cdots, n\}$，对应的样本点 x_k^* 取为第 k 个初始聚类中心位置．由 (9)，(10)，(11) 式决定的中心初始化方法，其原理与势函数方法相似，但运算量却比势函数方法小得多．

由于势函数法和密度函数法都是以样本之间的距离为基础构造的，故比较适合于对样本空间中呈团状分布的样本集合进行初始化．而对诸如线状和椭球壳状等的数据样本集，则要修改相应的范数形式方可使用．

3 实验结果分析

为了检验势函数和密度函数法对聚类中心初始化方法的有效性，以及所选的邻域半径的合理性，本文选取了 12 种不同类型的数据集合进行了实验，其中有两个集合为著名的 IRIS 和 BRITISH[5] 数据集合，这两个 4 维数据集合在聚类研究中经常被用来检验聚类效果，其余 10 个数据集合在空间中的分布见图 1．以上 12 组数据的描述及标准分类情况见表 1．

检验初始化方法是否有效，可以通过检验由以上方法得到的 C 个初始中心，观察其所对应的 C 个数据样本是否分别落在 C 个标准分类样本子集 (吸引域) 中．若每个标准样本子

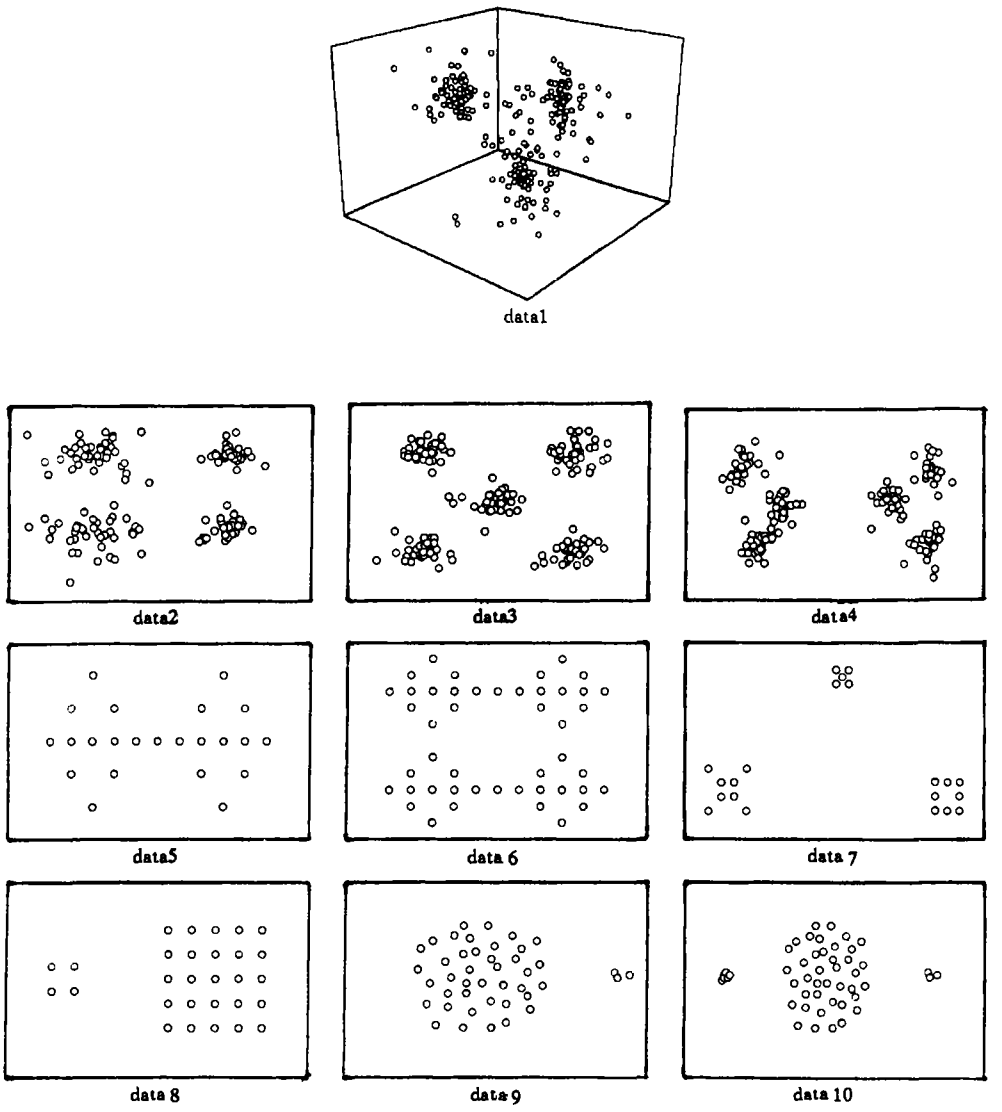

图 1　10 组数据集合在样本空间的分布图

集都有一个初始中心，则一般经聚类后可以得到希望的聚类结果。反之则认为初始化结果不理想。

表 2 是分别令 $r_\alpha = r_\beta = r_f, r_\alpha = r_\beta = r_m, r_d = r_f/2, r_d = r_m/2$，使用密度函数法和势函数法得到的实验结果。从结果分析，除了 data8, data9, data10 外，使用上述两种方法对 9 组样本数据都得到了比较理想的结果。值得一提的是 BRITISH 数据集合，其类之间的可分性较差，在初始化中心选择不合适的情况下，极易进入局部极小点，从而使聚类后的类中心离标准中心偏差较大，同时使样本点误分率增大。从表 2 结果可看出，取 $r_\alpha = r_\beta = r_f$ 及 $r_d = r_f/2, r_d = r_m/2$ 的情况下，对 BRITISH 数据，势函数法及密度函数法都得到了理想的初始化中心位置，而在 $r_\alpha = r_\beta = r_m$ 时，对 BRITISH 数据，由势函数法得到的初始中心位

置不理想 (第 1 类中没初始化中心点，而在第 3 类中有两个初始化中心点)。

表 1 12 组实验数据的描述和标准分类情况

	样本数量	样本维数	标准类数	标准数据分类						附注
				第 1 类	第 2 类	第 3 类	第 4 类	第 5 类	第 6 类	
IRIS	150	4	3	1–50	51–100	101–150				
BRITISH	50	4	4	1–10	11–20	21–36	37–50			
data1	240	3	3	1–80	81–160	161–240				
data2	160	2	4	1–40	41–80	81–120	121–160			
data3	200	2	5	1–40	41–80	81–120	121–160	161–200		
data4	180	2	6	1–30	31–60	61–90	91–120	121–150	151–180	
data5	23	2	2	1–12	12–23					12 为桥点
data6	46	2	4	1–12	12–23	24–35	35–46			12, 35 为桥点
data7	21	2	3	1-8	9–13	14–21				
data8	29	2	2	1–25	26–29					
data9	43	2	2	1–40	41–43					
data10	49	2	3	1–6	7–46	47–49				

表 2 经初始化得到的聚类中心所对应的数据样本编号表 (a)

	势函数法		密度函数法	
	$r_\alpha = r_\beta = r_f$	$r_\alpha = r_\beta = r_m$	$r_d = r_f/2$	$r_d = r_m/2$
IRIS	127 8 83	8 79 148	64 8 113	64 8 113
BRITISH	25 44 16 6	25 43 17 33	25 16 44 6	25 44 16 6
data1	120 194 64	120 194 64	106 215 64	120 215 64
data2	135 83 67 36	135 83 67 36	134 113 77 36	134 83 67 36
data3	110 44 139 20 187	110 44 139 20 187	20 82 123 80 187	20 82 123 44 187
data4	78 106 6 151 137 46	78 106 30 162 121 48	78 106 151 8 137 46	78 106 162 30 121 48
data5	9 15	6 18	9 18	15 6
data6	11 36 26 21	9 38 29 18	35 12 3 44	11 36 29 18
data7	15 6 11	15 6 11	15 6 11	15 6 11
data8	13 2	13 7	13 4	13 3
data9	17 34	15 20	17 34	15 32
data10	27 23 45	24 40 3	27 23 5	24 40 5

分析 data8, data9, data10 我们发现，这三组数据有一个共同的特点，即类之间样本数量相差很大，对这种类型的样本集合进行聚类时，传统的 FCM 方法不论初始中心位置如何选取，一般都会将数据量大的那一类中的部分样本划分到其他类中，要解决这种数据的聚类问题，需要采用部分加权 FCM 算法 [6]。对于这种数据，由于大类中数据点较多，故其相关的邻域半径较平均值要大，需选用较大的相关邻域半径进行初始化。

表 3 是分别令 $r_\alpha = r_\beta = 4r_f$, $r_\alpha = r_\beta = 4r_m$, $r_d = 2r_f$, $r_d = 2r_m$，用本文的两种方法进行的实验结果。表明，除在 $r_\alpha = r_\beta = 4r_m$, $r_d = 2r_m$ 时， data9 初始化的中心结果不理想外，其余均得到了理想的初始化中心。

表 3 经初始化得到的聚类中心所对应的数据样本编号表 (b)

	势函数法		密度函数法	
	$r_\alpha = r_\beta = 4r_f$	$r_\alpha = r_\beta = 4r_m$	$r_d = 2r_f$	$r_d = 2r_m$
data8	13 26	13 26	13 26	13 27
data9	21 41	21 38	21 41	21 38
data10	24 48 1	27 6 47	24 48 1	27 6 47

4 结　论

聚类算法，特别是 FCM 聚类算法，尽管得到了广泛的应用，但还存在着有待研究的问题．其中算法的初始化问题是一般非线性迭代优化算法中面临的一个共同的问题．合理的初始化，可以使算法很快达到需要的最优解．不合适的初始化，可能使优化过程收敛很慢，并且极易终止在不希望的局部最优解中．本文针对这一问题，对势函数法初始化的几个参数的选取进行了研究，同时提出了一种密度函数初始化方法．密度函数初始化方法与势函数算法相比，在初始化性能相当的情况下，其运算量大大减少．本文的初始化方法对于样本在空间中成团状分布的情况比较合适，若样本集合中各类样本数目相差悬殊，这可能使数目大的那一类样本类内的相关半径远大于其他类的类内样本相关半径，这时需要调整初始化函数的参数，以使初始化模型更符合数据集合本身．对于数据样本在空间中分布为非团状的情况 (如线状、椭球壳、超椭球壳等)，需要修正初始化模型．关于这一问题将进一步进行研究．

参 考 文 献

[1] Bezdek J C. Pattern Recognition with Fuzzy Objective Function Algorithms. New York: Plenum, 1981, 43–93.

[2] Bezdek J C, Hathaway R, Sabin M, Tucker W. Convergence theory for fuzzy C-means: Counterexample and repairs. IEEE Trans. on SMC., 1987, SMC-17(5): 873–877.

[3] Yager R R, Filev D P. Approximate clustering via the mountain method. IEEE Trans. on SMC., 1994, SMC-24(8): 1279–1284.

[4] Chiu S L. Fuzzy model identification based on cluster estimation. Journal of Intelligent and Fuzzy Systems, 1994, 2(3): 267–278.

[5] Moser C A, Scott W. British Towns. Edinburgh: Oliver and Boyd, 1961.

[6] 裴继红. 基于模糊信息处理的图像分割方法研究: [博士论文]. 西安电子科技大学, 1998.

A NEW INITIALIZATION METHOD OF CLUSTER CENTERS

Pei Jihong　　Fan Jiulun　　Xie Weixin

(*School of Electronic Engineering, , Xidian University, Xi'an* 710071)

*(*President Office, Shen-Zhen Univ., Shenzhen* 518060)

Abstract　The problems of parameter selections for potential function used to initialize cluster centers are discussed, and two methods are given for determining these parameters. Then a new density function to initialize cluster centers is also given which is computational effective. Finally, a set of compared experiments is presented to show the effectiveness of the proposed methods.

Key words　Clustering, Initialization, Potential function, Density function, Nonlinear optimal

裴继红：　男，1966 年生，博士，主要研究方向有模糊信息处理，计算机视觉，模式识别，图像理解，自然语言理解．

范九伦：　男，1963 年生，副教授，主要研究方向有模糊集理论，模糊信息处理，模式识别．

谢维信：　男，1941 年生，教授、博士生导师，主要研究方向有模糊集理论，模糊信息处理，智能信息处理，信号处理，模式识别．