面向隶属度修正模糊聚类的参数选择方法①

郭华峰[1], 陈德华[1], 陆慧娟[2]

[1](浙江工贸职业技术学院 信息传媒学院, 温州 325003)
[2](中国计量学院 信息工程学院, 杭州 310018)

摘 要: 隶属度修正是模糊 C-均值聚类算法改进的一个重要方向, 该类改进算法引入模糊阈值修正隶属度, 极大的加快了算法的收敛. 然而其模糊阈值的自适应取值一直是一个较难解决的问题. 针对这个问题, 从数据对聚类中心的物理吸引和相似关系等角度提出了一种针对隶属度修正类 FCM 算法的模糊阈值参数选择方法, 并从该参数选择公式的单调性、收敛性和鲁棒性等角度理论验证了该方法的有效性. 仿真实验表明, 该参数选择方法有效并具有较好的自适应效果, 在加入离群点时也有着较强的鲁棒性, 对于隶属度修正类 FCM 算法的参数选择有着较高的应用价值.

关键词: 模糊聚类; 隶属度修正; 参数选择; 相似关系; 自适应

Parameter Selection Method for Membership Correction Fuzzy Clustering

GUO Hua-Feng[1], CHEN De-Hua[1], LU Hui-Juan[2]

[1](College of Information and Communications, Zhejiang Industry & Trade Vocational College, Wenzhou 325003, China)
[2](College of Information Engineering, China Jiliang University, Hangzhou 310018, China)

Abstract: Membership correction is an important direction in the improvement of fuzzy c-means clustering algorithm. This type of improved algorithms introduce fuzzy threshold to correct membership value, which greatly speed up the algorithm convergence. However, the adaptive value of fuzzy threshold is always a difficult problem. To solve the problem, a method is presented to select the parameter of fuzzy threshold based on similarity relation and physical attraction between data and clustering centers. The monotonicity, convergence and robustness of the parameter selection formula are discussed to verify the effectiveness of this method. Simulation shows that the parameter selection method is effective, adaptive and robust, which has high application value to parameter selection of membership modified FCM algorithms.

Key words: fuzzy clustering; membership correction; parameter selection; similarity relation; adapting

自从 Dunn 和 Bezdek 提出模糊 C-均值聚类算法(FCM)算法[1,2], 对其各方面的改进就成为研究的热点, 主要集中在收敛性研究、目标函数修改、隶属度约束条件研究、多数据类型聚类和聚类有效性等几个方面[3-8].

其中很多学者把目光放在通过对隶属度的修正提升算法的收敛速度上. 文献[9]从竞争学习的角度提出了对手抑制式模糊 C-均值算法, 该算法引入模糊阈值抑制了第二大隶属度, 提升了最大的隶属度, 加快了算法的收敛. 文献[10]在文献[9]的基础上, 同时抑制了除最大隶属度之外的所有隶属度, 提高了算法的鲁棒性. 文献[11]基于分离强度指标为文献[10]的参数选择问题提供了解决方案. 文献[12]则对文献[11]的参数选择过程进行了探讨, 给出了另外一种方法. 区别于前述研究对隶属度的抑制, 文献[13]引入参数 α 作为隶属阈值, 直接判断数据的隶属关系, 从另一个角度修正隶属度, 提高了算法的收敛速度.

① 基金项目:国家自然科学基金(61272315,60842009);浙江省科技厅国际合作项目(2012C24030);浙江省温州市科技计划(G20130031)
收稿时间:2014-06-11;收到修改稿时间:2014-06-30

然而与文献[10]提出的 S-FCM 算法存在的问题相同, 文献[13]提出的截断阈值模糊 C-均值聚类(FCMα)算法也存在着模糊阈值参数选择的问题. 文献[13]给出了隶属阈值 α 的取值区间, 但并没有给出其自适应取值的方法. 针对这个问题, 从数据对聚类中心的物理吸引和相似关系等角度提出一种自适应的参数选择方法, 接着对该方法的性能做了理论上的探讨, 最后使用仿真实验对该参数选择方法的有效性和性能做了验证.

1 截断阈值模糊C均值聚类算法

模糊 C-均值聚类算法于 1974 年由 Dunn 最先提出, 由 Bezdek 推广到更普遍的形式, 并建立了模糊 C-均值聚类理论. FCM 算法的应用非常广泛, 可应用于信息检索与分类、生物学、气候学、心理学、医学和商业等多个领域. 其算法步骤如下:

1)算法初始化: 确定聚类类别数 c 和模糊参数 b, 设置迭代停止阈值 $\varepsilon>0$, 迭代次数 $l=0$, 并初始化各个聚类中心 m_j.

2)根据式(1)计算各个隶属度, 更新 $U^{(l)}\to U^{(l+1)}$.

$$\mu_{ij}=\frac{(1/\|x_i-m_j\|^2)^{1/(b-1)}}{\sum_{k=1}^{c}(1/\|x_i-m_k\|^2)^{1/(b-1)}},$$

$$i=1,2,...,n,\ j=1,2,...,c \tag{1}$$

3)根据式(2)计算各个聚类中心点 m_j.

$$m_j=\frac{\sum_{i=1}^{n}[\mu_j(x_i)]^b x_i}{\sum_{i=1}^{n}[\mu_j(x_i)]^b},\ j=1,2,...,c \tag{2}$$

4)查看迭代停止条件, 若 $\|U^{(l)}-U^{(l+1)}\|<\varepsilon$, 停止迭代, 否则 $l=l+1$, 转到步骤(2).

截断阈值模糊 C-均值聚类算法(Alpha-Cut Implemented Fuzzy Clustering Algorithms, 简称 FCMα 算法)由 M.S. Yang 在文献[13]中提出, 其实质是在传统 FCM 算法基础上加入了对隶属度的修正. FCMα 算法引入阈值参数 $\alpha(0.5\le\alpha\le1)$, 在 FCM 算法的步骤(2)和步骤(3)之间跟每个数据点的最大隶属度进行比较, 若最大隶属度大于 α, 那么最大隶属度等于 1, 其它隶属度都等于 0, 否则各隶属度不变化. FCMα 算法步骤如下:

1)算法初始化: 确定聚类类别数 c 和模糊参数 b, 设置迭代停止阈值 $\varepsilon>0$, 迭代次数 $l=0$, 并初始化各个聚类中心 m_j, 给定隶属阈值参数 $\alpha(0.5\le\alpha\le1)$.

2)根据式(1)计算各个隶属度, 更新 $U^{(l)}\to U^{(l+1)}$.

3)如果 $U_{ik}^{(l+1)}=\max_{1\le j\le c}U_{jk}^{(l+1)}>\alpha$, 那么 $U_{ik}^{(l+1)}=1$, $U_{i'k}^{(l+1)}=0$ ($i'\ne i$).

4)根据式(2)计算各个聚类中心点 m_j.

5)查看迭代停止条件, 若 $\|U^{(l)}-U^{(l+1)}\|<\varepsilon$, 停止迭代, 否则 $l=l+1$, 转到步骤(2).

FCMα 算法通过参数 α 对算法中隶属度的快速放大, 加快了算法的收敛速度, 也提高了算法的稳定性. 然而参数 α 的自适应取值方法并未给定, 这个问题制约了 FCMα 算法的进一步应用.

为了解决这个问题, 从数据对聚类中心的物理吸引和相似关系等角度进行推导.

2 FCMα算法的参数选择推导

在式(2)中, 类中心点 m_j 的计算由数据点 x_i 和其隶属度 μ_{ij} 共同决定, 其中隶属度大的数据点影响越大. 把这种影响比作作用力, 则 μ_{ij} 可以理解为 x_i 施加于 m_j 的作用力, μ_{ij} 越大, x_i 对 m_j 的作用力越大. 在式(1)中, μ_{ij} 的大小取决于 x_i 到 m_j 的距离, 距离越近, μ_{ij} 越大, 则 μ_{ij} 可以进一步理解为 x_i 对 m_j 的吸引力. 所有的数据点对所有中心点都有吸引力, FCM 算法的最终收敛其实是寻找一种物理状态, 让所有数据点对所有中心点的吸引趋于平衡.

由以上理论来考虑图 1, 类 2 中的数据点对类 1 的中心点 m_1 有吸引力, 这个吸引力将使 m_1 偏离类 1. 类 1 中的数据点则把吸引力施加于 m_1, 使 m_1 留在类 1, 同时对类 2 的中心点 m_2 产生影响. 中心点可以视为所在类数据点的平均, 则类 2 对 m_1 的吸引可以视为 m_2 对 m_1 的吸引. 考虑类 1 中的数据点 x_1, x_1 和 m_2 对 m_1 都有着吸引力, 其中 m_2 使 m_1 偏离类 1, x_1 使 m_1 留在类 1. 根据作用力反作用力的原理, x_1 隶属于 m_1 也可以视为 m_1 隶属于 x_1(所在的类). 根据最大隶属度原则, m_1 隶属于 x_1 还是 m_2, 取决于各自隶属度的大小, 则可以得出一个结论, x_1 要隶属于 m_1, 其隶属度要大于 m_1 对 m_2 的隶属度, 即大于 m_2 对 m_1 的吸引力.

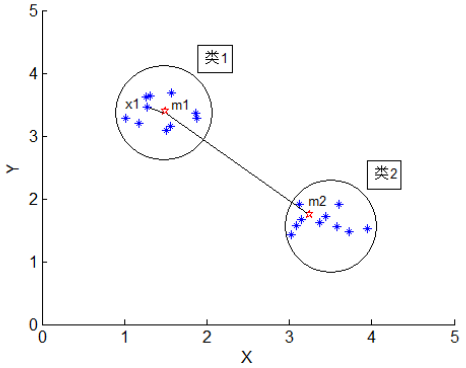

图 1　数据对聚类中心的吸引

推广到分多类的情况，可以得到同样的结论，即 x_1 要隶属于 m_1，其隶属度要大于其他中心点对 m_1 的作用力. 所以 FCMα 算法的参数选择问题可以转化为获取中心点之间的作用力值. 两点之间的相互作用可以用相似关系来阐述，其作用值可以用相似度值来代替. 关于相似关系和相似度，L.A. Zadeh 在文献[14]中进行了详细的阐述，给出了相似关系的定义和各项性质的探讨. 在对其第三节相似关系的传递性探讨过程中，文献[14]给出了一个相似度量的参考公式：$\mu_S(x,y)=e^{-\beta|x-y|}$，其中 $\beta>0$. $\mu_S(x,y)$ 代表 x 与 y 的相似程度，应用于中心点 m_j 与 m_k，则 m_j 与 m_k 的相似度量表示为 $\mu_S(m_j,m_k)=e^{-\beta\|m_j-m_k\|^2}$. 在实际的 FCMα 算法中，类中心点可能会超过 2 个，为确保数据点隶属于相应中心点，则其隶属度需大于最大的中心点相似度. 在 $\mu_S(m_j,m_k)=e^{-\beta\|m_j-m_k\|^2}$ 代表的相似度值中，由于类中心点之间越近，相似度越大，所以 α 的取值公式推导为 $\alpha=e^{-\beta\min\limits_{j\neq k}\|m_j-m_k\|^2}$. 为了调适 α 的最终取值，设置 β 为样本方差的倒数，则最终公式为：

$$\alpha=\exp(-\frac{\min\limits_{j\neq k}\|m_j-m_k\|^2}{\beta}) \tag{3}$$

其中，$\beta=\frac{\sum_{j=1}^{n}\|x_j-\bar{x}\|^2}{n}$，$\bar{x}=\frac{\sum_{j=1}^{n}x_j}{n}$，

把式(3)代入 FCMα 算法的每次迭代中，可以得到新的算法，这里命名为 Mα-FCM(Modified FCMα)算法. Mα-FCM 算法使用了自动迭代的参数 α，提高了算法的自适应性.

式(3)作为推导出来的参数选择公式，有其性能上的优势. 为了检验其有效性，从公式的单调性、收敛性和鲁棒性等角度来验证.

3　参数选择公式性质研究

3.1　单调性

性质 1. 当迭代次数 l 递增时，参数选择公式(3)总体递减.

令 $t=\min\limits_{j\neq k}\|m_j-m_k\|^2$，则 $\alpha=\exp(-\frac{t}{\beta})$. 显然，$\alpha$ 是 t 的单调递减函数，也即中心点之间距离越小，需要的参数 α 的值越大，中心点之间的距离越大，需要的参数 α 的值越小. 在 FCM 和 FCMα 算法中，算法随着迭代的进行逐渐收敛，中心点之间的距离也将变大并趋于稳定. 所以 t 随着迭代次数 l 的逐渐变大而变大，从而总体上，α 也是迭代次数 l 的递减函数. 也即随着迭代的进行，α 逐渐变小直到趋于稳定.

3.2　收敛性

性质 2. 参数选择公式(3)是收敛的.

J.C. Bezdek 在文献[15]中给出了 FCM 算法的收敛性证明. 在此基础上，文献[13]引入动力系统理论，给出了 FCMα 算法的收敛性证明，也即对于任意固定阈值 $\alpha(0.5\le\alpha\le1)$，FCMα 算法是收敛的. Mα-FCM 算法与 FCMα 算法步骤相同，不同于 FCMα 算法的只是阈值参数 α 的动态化，且总体递减，这不影响文献[13]引理 1 的证明，所以很容易得到这样的结论，即 Mα-FCM 也是收敛的.

随着 Mα-FCM 的收敛，$\min\limits_{j\neq k}\|m_j-m_k\|^2$ 也将趋于稳定，所以很容易得出一个结论：存在 $\alpha_0>0$，当迭代次数 l 趋向于无穷大时，$\lim\limits_{l\to\infty}\alpha(l)=\alpha_0$，也即参数选择公式(3)也是收敛的.

3.3　鲁棒性

性质 3. 参数选择公式(3)具有一定的鲁棒性.

正态分布由于其概率分布具有集中性，横轴区间(μ-2.58σ，μ+2.58σ)内的面积为 99.730020%，所以具有较强的抗噪性. 因为公式 $\alpha=\exp(-\frac{\min\limits_{j\neq k}\|m_j-m_k\|^2}{\beta})$ 符合正态分布的概率密度函数 $f(x)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$，所以参数选择公式(3)具有较强的鲁棒性.

以上对单调性、收敛性和鲁棒性的探讨部分验证了公式(3)的有效性. 为了更好的检验 Mα-FCM 算法的

效果和式(3)的各项性质, 进行以下实验.

4 仿真实验

通过三组实验检验算法的效果.

1)给定 4 群范围在 0 到 5 之间高斯分布的二维数据, 每群 150 个数据, 如图 2 所示.

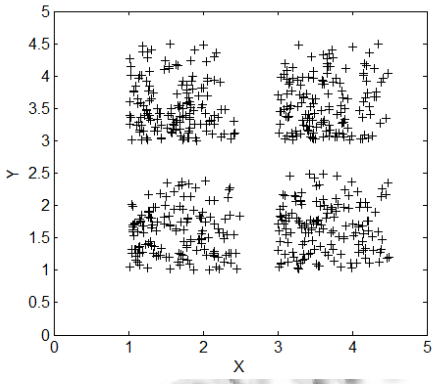

图 2 高斯分布的二维数据

选定模糊系数 b=2, 设定相同的初始条件和迭代结束条件, 分别对 Mα-FCM 算法、MS-FCM 算法和 FCMα 算法在 α 等于 0.5、0.7、0.9 和 1.0 的情况下做二十次实验, 取其平均值, 可以得到如表 1 所示的结果.

表 1 各算法性能比较

	MS-FCM	α (FCMα)				Mα-FCM
		0.5	0.7	0.9	1.0	
迭代数	15	14	17	21	22	13
MSE	0.5172	0.5174	0.5185	0.5165	0.5164	0.5171

表 1 提供的数据表明, Mα-FCM 算法解决了 FCMα 算法的参数选择问题, 取得了较好的自适应效果, 与 FCMα 算法在 α 等于 0.5、0.7、0.9 和 1.0 时的结果相比, 迭代次数和 MSE 等统计处于较优的状态.

(2)在图 2 所示的数据集中加入一个离群点(1, 28), 再次对 Mα-FCM 算法、MS-FCM 算法和 FCMα 算法做二十次实验, 取其平均值, 可以得到如表 2 所示的结果.

表 2 加离群点下的各算法性能比较

	MS-FCM	α (FCMα)				Mα-FCM
		0.5	0.7	0.9	1.0	
迭代数	14	13	16	19	19	12
MSE	0.5621	0.5577	0.5584	0.5568	0.5568	0.5574

表 2 提供的数据表明, 在加入离群点的情况下, Mα-FCM 算法也能很好的适应, 收敛速度和收敛效果都达到了较优的水平, 且优于 MS-FCM 算法的表现, 具有较强的鲁棒性.

实验中, 我们也得到了 Mα-FCM 算法中参数 α 的迭代取值轨迹, 如图 3 所示.

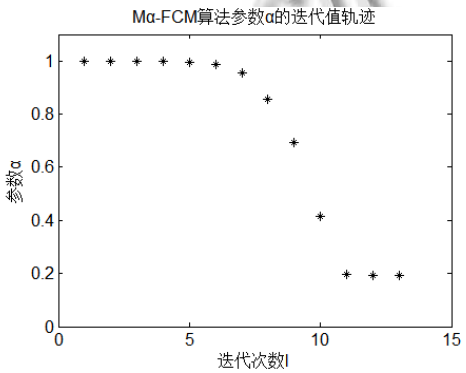

图 3 Mα-FCM 算法参数 α 的迭代值轨迹

3)采用 Iris 和 Wine 数据集作为测试数据集[16], 设定相同的初始条件和迭代结束条件, 分别对 Mα-FCM 算法和 FCMα 算法在 α 等于 0.5、0.7、0.9 和 1.0 的情况下做二十次实验, 取其平均值, 可以得到如表 3 所示的结果.

表 3 高维数据集下的各算法性能比较

数据集	比较项	α (FCMα)				Mα-FCM
		0.5	0.7	0.9	1.0	
Iris	迭代数	12	21	19	22	12
	正确率(%)	89.33	89.33	89.33	89.33	89.33
Wine	迭代数	15	20	45	54	15
	正确率(%)	70.22	67.42	69.66	68.54	70.22

从表 3 的数据可以看出, 在使用 Iris 和 Wine 等高维数据集的情况下, Mα-FCM 算法也可以有很好的表现. 通过比对表 3 和表 1、表 2 的数据, 我们发现, 相对于低维数据集, Mα-FCM 算法在高维数据集中的表现更优异, 迭代数和正确率都达到了最优的效果. 这进一步验证了 Mα-FCM 算法的可靠性, 也说明本文提出的参数选择方法是有效的.

5 结论

本文从数据点与聚类中心点相互作用力的角度分析了 FCMα 算法的参数选择问题, 得到了自适应参数的参数选择方法, 据此提出了 Mα-FCM 算法, 并通过对参数选择公式相关性质的探讨部分验证了该选择方法的有效性. 最后的仿真实验表明, 无论是对低维的数据集还是高维的数据集, 该参数选择方法都具有较好的自适应效果, Mα-FCM 算法的收敛速度和收敛效果都达到了一个较优的水平, 当加入离群点时, Mα-FCM 算法也取得了较优的结果, 具有较强的鲁棒性.

仿真实验也验证了参数选择公式(3)性质的有效性. 图 3 的结果表明式(3)得到的 α 值总体上是递减的, 最终也将收敛于一固定的值. 表 2 的数据也间接验证了式(3)的鲁棒性. 由于式(3)与聚类中心点之间的最小距离有关, 具有收敛和鲁棒等特点, 其一般可用于具有迭代过程优化的 FCM 改进算法中, 如 S-FCM 算法和 FCMα 算法等, 对于文献[9, 10]、文献[13]和文献[17, 18]所代表的隶属度修正类算法的参数选择有着较高的参考价值.

在参数推导过程中, 把隶属度解释成了数据点对聚类中心点的作用力, 得到了FCM算法的收敛其实就是达到了数据点和中心点作用力平衡状态的结论. 后续的研究工作也可以尝试从物理学的作用力平衡角度对 FCM 算法进行阐述.

参考文献

1 Dunn JC. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. Journal of Cybernetics, 1973, 3(3): 32–57.

2 Bezdek JC. Pattern recognition with fuzzy objective function algorithms. Plenum Press, New York, 1981.

3 Huang HC, Chuang YY, Chen CS. Multiple kernel fuzzy clustering. IEEE Tran. on Fuzzy Systems, 2012, 20(1): 120–134.

4 Pal NR, Sarkar K. What and when can we gain from the kernel versions of c-means algorithm? IEEE Tran. on Fuzzy Systems, 2014, 22(2): 363–379.

5 Havens TC, Bezdek JC, Leckie C, Hall LO, Palaniswami M. Fuzzy C-means algorithms for very large data. IEEE Trans. on Fuzzy Systems, 2012, 20(6): 1130–1146.

6 Wu JJ, Xiong H, Liu C, Chen J. A generalization of distance functions for fuzzy-means clustering with centroids of arithmetic means. IEEE Tran. on Fuzzy Systems, 2012, 20(3): 557–571.

7 Zhao ZX, Cheng LZ, Cheng GQ. Neighbourhood weighted fuzzy c-means clustering algorithm for image segmentation. IET Image Processing, 2014, 8(3): 150–161.

8 O. Linda, M. Manic. General type-2 fuzzy c-means algorithm for uncertain fuzzy clustering. IEEE Tran. on Fuzzy Systems, 2012, 20(5): 883–897.

9 魏立梅,谢维信.对手抑制式模糊 C-均值算法.电子学报, 2000,28(7):63–66.

10 Fan JL, Zhen WZ, Xie WX. Suppressed fuzzy c-means clustering algorithm. Pattern Recognition Letters, 2003, 24: 1607–1612.

11 Hung WL, Yang MS, Chen DH. Parameter selection for suppressed fuzzy c-means with an application to MRI Segmentation. Pattern Recognition Letters, 2006, 27: 424–438.

12 Saad MF, Alimi AM. Improved modified suppressed fuzzy C-means. 2010 2nd International Conference on IEEE Image Processing Theory Tools and Applications (IPTA). 2010. 313–318.

13 Yang MS, Wu KL, Hsieh JN, Yu J. Alpha-cut implemented fuzzy clustering algorithms and switching regressions. IEEE Trans. on Systems, Man, and Cybernetics, 2008, 38(3): 588–603.

14 Zadeh LA. Similarity relations and fuzzy orderings. Information Sciences, 1971, 3(2): 177–200.

15 Bezdek JC. A convergence theorem for fuzzy ISODATA clustering algorithm. IEEE Tran. on Pattern Analysis and Machine Intelligence, 1980, 1(2): 1–8.

16 Blake CL, Merz CJ. UCI repository of machine learning databases. [Technical Report]. Irvine, CA: University of California, Department of Information and Computer Science, 1998.

17 黄建军,谢维信.半抑制式模糊 C-均值聚类算法.中国体视学与图像分析,2004,10(2):109–113.

18 赵凤,范九伦.优选抑制式非局部空间模糊 C-均值图像分割方法.计算机应用研究,2012,29(7):2737–2746.