Doi:10.3969/j.issn.1672-0105.2015.01.010

一种改进的模糊C-均值聚类算法*

郭华峰,洪年松,范渊
(浙江工贸职业技术学院 信息传媒学院，浙江 温州 325003)

摘 要：相对于抑制式模糊C-均值聚类算法，半抑制式模糊C-均值聚类算法引入抑制门限，在保持收敛速度变化不大的前提下改善了聚类的效果。然而，该算法也存在参数选择较多的问题。针对该问题，我们引入面向隶属度修正的模糊聚类参数选择方法，提出了改进的半抑制式模糊C-均值聚类算法。实验表明，该算法具有较好的可操作性，在具有较快收敛速度的同时，也维持了较好的收敛效果。

关键词：模糊聚类；半抑制式；参数选择；可操作性
中图分类号：TP301.6 **文献标识码**： A **文章编号**：1672-0105（2015）01-0043-04

An Improved Fuzzy C-Means Clustering Algorithm

GUO Hua-feng, HONG Nian-song, FAN Yuan
(College of Information and Communications, Zhejiang Industry & Trade Vocational College, Wenzhou, 325003, China)

Abstract: Comparison of suppressed fuzzy C- means clustering algorithm, half suppressed fuzzy C- means clustering algorithm introduces suppression threshold parameter, which improves the clustering effect while keeping with little change in convergence speed. However, the algorithm has the problem of parameter selection. To solve the problem, an parameter selection method for membership correction fuzzy clustering is introduced, and the improved half suppressed fuzzy C- means clustering algorithm is put forward. Experiments show that, the new algorithm has a better operability, faster convergence speed, and a good convergence effect.

Key Words: fuzzy clustering; half suppressed; parameter selection; operability

0 引言

自从Jiulun Fan提出抑制式模糊C-均值聚类（简称S-FCM）算法以来[1]，其抑制式的思想就得到了广泛和深入的研究[2-4]，范九伦在综述文献[5]中较为详细的阐述了这一点。在众多研究文献中，文献[6]引入抑制门限，提出了半抑制式模糊C-均值聚类算法（简称HSFCM），改善了S-FCM算法的聚类效果。然而，该算法由于多添加了一个抑制门限的参数，所以其可操作性变得较差。针对这个问题，我们引入面向隶属度修正的模糊聚类参数选择方法，提出了改进的半抑制式模糊C-均值聚类算法。仿真实验表明，改进后的算法具有较好的可操作性，其参数选择方法具有较好的自适应效果，对于低维数据集和高维数据集都是有效的。

1 半抑制式模糊C-均值聚类算法

2003年，针对模糊C-均值聚类算法收敛速度慢的问题，Jiulun Fan引入抑制因子α（$0 \leqslant \alpha \leqslant 1$），提出了抑制式模糊C-均值聚类算法。其算法步骤如下：

Step 1 取模糊度 $1<m<\infty$，分类数 $2 \leqslant c \leqslant n$，迭代终止门限 $\varepsilon>0$，迭代次数 $l=0$，抑制因子 $0 \leqslant \alpha \leqslant 1$，给定聚类中心点的起始值 $Z^{(0)}$。

Step 2 根据(1)式计算隶属值 μ_{ij}。

收稿日期：2015-01-25
基金项目：浙江省高等学校访问学者教师专业发展项目（FX2014175）；温州市科技计划项目（G20130031）；浙江工贸职业技术学院教师科技创新活动计划项目（X140203）
作者简介：郭华峰，硕士，浙江工贸职业技术学院讲师，主要研究方向：图像处理、模式识别；洪年松，硕士，浙江工贸职业技术学院讲师，主要研究方向：图像处理；范渊，硕士，浙江工贸职业技术学院讲师，主要研究方向：图像处理。

$$\mu_{ij}=\left(\sum_{k=1}^{c}\frac{\left\|x_i-Z_j\right\|^{2/(m-1)}}{\left\|x_i-Z_k\right\|^{2/(m-1)}}\right)^{-1}\quad,\quad j=1,2,\cdots,c\quad,\quad i=1,2,\cdots,n \tag{1}$$

Step 3 使用以下方法更新 μ_{ij}，i=1,...,n，j=1,...,c：如果 $\mu_{is}=\max_{1\leqslant j\leqslant c}\mu_{ij}$，那么 $\mu_{is}=1-\alpha+\alpha\mu_{is}$，$\mu_{is'}=\alpha\mu_{is'}(s'\neq s)$

Step 4 根据(2)式计算中心点 $Z^{(k+1)}$。

$$Z_j=\frac{\sum_i\mu_{ij}^m x_i}{\sum_i\mu_{ij}^m}\quad,\quad j=1,2,\cdots,c \tag{2}$$

Step 5 若 $\left\|Z^{(k+1)}-Z^{(k)}\right\|<\varepsilon$，则停止，否则转到Step 2。

S-FCM算法由于对除最大隶属度之外的其它隶属度进行了抑制，加快了最大隶属度的增长，从而提高了算法的收敛速度。然而S-FCM算法具有聚类效果不佳的缺点，文献[6]引入了抑制门限β（$0\leqslant\beta\leqslant1$），提出了半抑制式模糊C-均值聚类算法。该算法步骤如下：

Step 1 取模糊度 $1<m<\infty$，分类数 $2\leqslant c\leqslant n$，迭代终止门限 $\varepsilon>0$，迭代次数 $l=0$，抑制因子 $0\leqslant\alpha\leqslant1$，抑制门限β（$0\leqslant\beta\leqslant1$），给定聚类中心点的起始值 $Z^{(0)}$。

Step 2 根据(1)式计算隶属值 μ_{ij}。

Step 3 使用以下方法更新 μ_{ij}，i=1,...,n，j=1,...,c：若 $\mu_{is}=\max_{1\leqslant j\leqslant c}\mu_{ij}$，如果 $\mu_{is}>\beta$，那么 $\mu_{is}=1-\alpha+\alpha\mu_{is}$，$\mu_{is'}=\alpha\mu_{is'}(s'\neq s)$，否则不作更新。

Step 4 根据(2)式计算中心点 $Z^{(k+1)}$。

Step 5 若 $\left\|Z^{(k+1)}-Z^{(k)}\right\|<\varepsilon$，则停止，否则转到Step 2。

相比较S-FCM算法，HSFCM算法在Step3中对隶属度的抑制更新添加了一个条件，只有最大隶属度大于抑制门限，才进行抑制操作，这样就阻止了过小最大隶属度的不合理增长，改善了聚类的效果。然而，HSFCM算法由于在抑制因子参数的基础上又加入了抑制门限参数，使得在算法的初始化阶段step1需要设定两个参数的值，这降低了算法的可操作性。此时，如何提高HSFCM算法的可操作性就成了必须探讨的问题。

2 半抑制式模糊C-均值聚类算法的改进

相对于S-FCM算法，HSFCM算法难以操作是因为有抑制因子α和抑制门限β两个参数的值需要确定，所以可以从降低待选择的参数个数方面着手。借鉴文献[7]面向隶属度修正类模糊聚类算法提出的参数选择方法，对抑制因子α进行参数选择，我们得到了α的自适应取值公式：

$$\alpha=\exp\left(-\frac{\min_{j\neq k}\left\|Z_j-Z_k\right\|^2}{\lambda}\right) \tag{3}$$

其中 $\lambda=\frac{\sum_{i=1}^{n}\left\|x_i-\bar{x}\right\|^2}{n}$，$\bar{x}=\frac{\sum_{i=1}^{n}x_i}{n}$，由于该取值公式为每次迭代自动计算，所以无需为抑制因子设置初始值，这样就减少了HSFCM算法的参数选择个数，提高了算法的可操作性。将（3）式代入HSFCM算法的抑制因子中，可以得到一种改进的HSFCM（Improved HSFCM，简称I-HSFCM）算法，新的算法步骤如下：

Step 1 取模糊度 $1<m<\infty$，分类数 $2\leqslant c\leqslant n$，迭代终止门限 $\varepsilon>0$，迭代次数 $l=0$，抑制门限β（$0\leqslant\beta\leqslant1$），给定聚类中心点的起始值 $Z^{(0)}$。

Step 2 根据(1)式计算隶属值 μ_{ij}，根据(3)式计算抑制因子α。

Step 3 使用以下方法更新 μ_{ij}，i=1,...,n，j=1,...,c：若 $\mu_{is}=\max_{1\leqslant j\leqslant c}\mu_{ij}$，如果 $\mu_{is}>\beta$，那么 $\mu_{is}=1-\alpha+\alpha\mu_{is}$，$\mu_{is'}=\alpha\mu_{is'}(s'\neq s)$，否则不作更新。

Step 4 根据(2)式计算中心点 $Z^{(k+1)}$。

Step 5 若 $\left\|Z^{(k+1)}-Z^{(k)}\right\|<\varepsilon$，则停止，否则转到Step 2。

观察算法步骤可知，相对于HSFCM算法，I-HSFCM算法不需要在算法初始化阶段设置抑制因子的值，只是在Step2中加入了对抑制因子的自适应计算，这减少了HSFCM的参数选择，提高了算法的可操作性。为了验证I-HSFCM算法的有效性，进行以下实验。

3 仿真实验

实验一，随机生成4堆正态分布的数据点，每堆50个数据，共200个数据点，如图1所示。

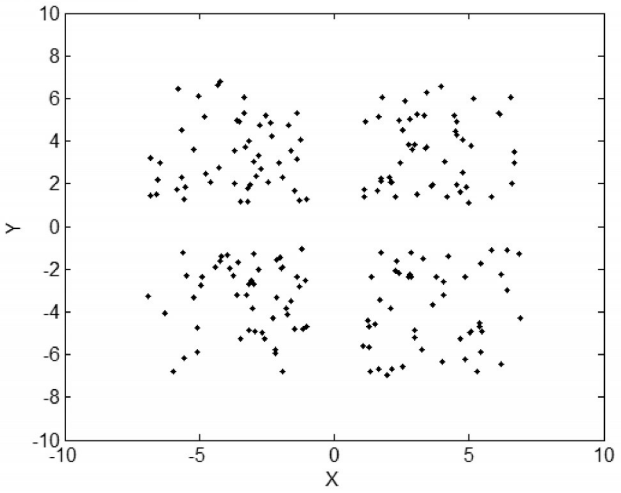

图1　随机生成的正态分布数据点

在图1所示的数据集中使用新提出的I-HSFCM算法和抑制因子α分别为0、0.2、0.4、0.6、0.8和1的HSFCM算法，取抑制门限β=0.5，同时设置各种算法的模糊度m=2，分类数c=4，其他的初始条件也设置相同，执行上述算法20次，取其平均，得到如表1所示的结果。

表1 I-HSFCM算法和HSFCM算法对图1数据的聚类结果

比较项	HSFCM (α)						I-HSFCM
	0	0.2	0.4	0.6	0.8	1	
迭代数	14	16	17	19	22	28	16
mse	2.1708	2.1710	2.1702	2.1692	2.1674	2.1652	2.1704

从表1可以发现，相对于不同抑制因子的HSFCM算法，I-HSFCM算法在迭代数和mse方面取得了较好的平衡，迭代数处于较少的水平，这表示收敛速度得到了较好的提升，同时mse也处于较小的水平，这说明收敛效果也得到了较好的维持。从表1的结果可以看出，I-HSFCM算法对于抑制因子的参数选择取得了较好的自适应效果。同时我们也得到了I-HSFCM算法在不同抑制门限β的情况下对图1数据的聚类结果，如表2所示。

表2　不同抑制门限的I-HSFCM算法对图1数据的聚类结果

比较项	I-HSFCM (β)					
	0	0.2	0.4	0.6	0.8	1.0
迭代数	16	16	16	17	26	28
mse	2.1689	2.1689	2.1689	2.1761	2.1694	2.1652

由于β=1时I-HSFCM算法等价于FCM算法，所以表2的数据表明，I-HSFCM算法对于抑制因子的自适应选择并没有降低收敛速度的提升，相对于FCM算法，I-HSFCM算法保持了S-FCM算法和HS-FCM算法在收敛速度提升上的优势，且收敛效果不差，这说明I-HSFCM算法是有效的。以上是算法在二维数据集的表现，下面进行高维数据集的实验。

实验二，为了更好的检验I-HSFCM算法的性能，采用模糊聚类中经典的高维数据集Iris和Wine[8]。与实验一相同，使用I-HSFCM算法和抑制因子α分别为0、0.2、0.4、0.6、0.8和1的HSFCM算法，取抑制门限β=0.5，同时设置各种算法的模糊度m=2，分类数c=3，其他的初始条件也设置相同，执行上述算法20次，取其平均，得到如表3所示的结果。

表3的结果表明，在高维数据集中I-HSFCM算法也取得了收敛速度和聚类效果的平衡，其总体表现处于较优水平。另外我们也发现，在维数越高的Wine数据集中，I-HSFCM算法对于收敛速度的提升更明显。这说明I-HSFCM算法在高维数据集中也同样有效。

表3　I-HSFCM算法和HSFCM算法对Iris、Wine数据集的聚类结果

数据集	比较项	HSFCM (α)						I-HSFCM
		0	0.2	0.4	0.6	0.8	1.0	
Iris	迭代数	12	14	16	18	21	23	17
	准确率	89.33%	89.33%	89.33%	89.33%	89.33%	89.33%	89.33%
Wine	迭代数	14	16	22	42	59	68	19
	准确率	72.47%	72.47%	72.47%	69.66%	69.66%	68.54%	72.47%

4 结论

针对HSFCM算法由于初始化参数过多而可操

作性差的问题，引入面向隶属度修正模糊聚类的参数选择方法，得到了改进的I-HSFCM算法。I-HSFCM算法使用迭代公式对抑制因子进行自适应选择，减少了HSFCM算法的参数选择个数，提高了算法的可操作性。仿真实验表明，该算法在收敛速度和收敛效果方面取得了较好的平衡，对低维数据集和高维数据集，都具有较好的自适应效果，这说明算法的改进是有效的。

参考文献：

[1] Jiu-Lun Fan, Wen-Zhi Zhen, Wei-Xin Xie. Suppressed fuzzy c-means clustering algorithm[J]. Pattern Recognition Letters, 2003, 24: 1607-1612.

[2] F. Zhao, J. L.Fan, H. Liu. Optimal-selection-based suppressed fuzzy c-means clustering algorithm with self-tuning non local spatial information for image segmentation[J]. Expert Systems with Applications, 2014, 41(9): 4083-4093.

[3] W. L. Hung, D. H. Chen, M. S. Yang. Suppressed fuzzy-soft learning vector quantization for MRI segmentation[J]. Artificial intelligence in medicine, 2011, 52(1): 33-43.

[4] L.Szilágyi. Lessons to learn from a mistaken optimization[J]. Pattern Recognition Letters, 2014, 36: 29-35.

[5] 范九伦.抑制式模糊C-均值聚类研究综述[J].西安邮电大学学报, 2014,19(3):1-5.

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

[7] 郭华峰,陈德华,陆慧娟.面向隶属度修正模糊聚类的参数选择方法[J].计算机系统应用, 2015,24(1): 166-170.

[8] C.L. Blake, C.J. Merz. UCI repository of machine learning databases. [http://archive.ics.uci.edu/ml/]. Irvine, CA: University of California, Department of Information and Computer Science, 1998.

（责任编辑：潘修强）

（上接第9页）

的同学选择更贴近温州实际，31.78%的同学选择拓宽社会资源。

3、学生参与创业的比例有较大提高。已有初步创业想法的比例由以前的29.05%上升到43.33%，参加创业精英班学习的同学进行创业的比例由10.81%上升到21.76%。

综上所述，通过设计与实施阶段递进式产教融合推进大学生创业教育，取得了较好的效果，但课程内容的衔接如何更加科学合理、项目对接如何更加有效等，还需要进一步研究完善。

参考文献：

[1]教育部.关于印发《普通本科学校创业教育教学基本要求（试行）》的通知(教高厅〔2012〕4号),2012.

（责任编辑：邱开金）