基于减法聚类改进的模糊 c-均值算法的模糊聚类研究*

于 迪[1]，李义杰[2]

(1. 辽宁工程技术大学 研究生学院，辽宁 葫芦岛 125105；
2. 辽宁工程技术大学 软件学院，辽宁 葫芦岛 125105)

摘 要：针对模糊 c-均值(FCM)聚类算法受初始聚类中心影响，易陷入局部最优，以及算法对孤立点数据敏感的问题，提出了解决方案：采用快速减法聚类算法初始化聚类中心，为每个样本点赋予一个定量的权值，用来区分不同的样本点对最终的聚类结果的不同作用，为提高聚类速度采用修正隶属度矩阵的方法，并将算法与传统的 FCM 相比。实验结果表明，该算法较好地解决了初值问题，与随机初始化方法相比，迭代次数少、收敛速度快、具有较好的聚类结果。

关键词：模糊 c-均值；减法聚类；权值

中图分类号：TP18 **文献标识码**：A **文章编号**：1674-7720(2010)16-0014-03

Research on fuzzy clustering based on subtractive clustering and improved fuzzy c-means algorithm

YU Di[1], LI Yi Jie[2]

(1. Institute of Graduate, Liaoning Technical University, Huludao 125105, China;
2. College of Software Engineering, Liaoning Technical University, Huludao 125105, China)

Abstract: For the fuzzy c-means clustering algorithm is affected by the initial cluster center, easy to fall into local optimum and algorithm is sensitive to outlier data. This paper give the solution: subtractive clustering is utilized to initialize the cluster centers, every sample holds a quantificational weight in order to differentiate the different effects of different samples for knowledge discovery, use the approach of ameliorating degree of membership to improve the clustering speed and compare the algorithm with the traditional algorithm. Experiment results show that the algorithm can solve the initial problem, and it has fewer iterations and faster convergence than the random initialization. The algorithm has better clustering results.

Key words: fuzzy c-means(FCM); subtractive clustering; weights

模糊聚类作为无监督机器学习的主要技术之一，广泛应用于数据挖掘、矢量量化、图像分割、模式识别、医学诊断等领域。引入模糊数学方法，通过建立数据样本类属的不确定描述，将相似性质的事物分开并加以分类，能比较客观地反映现实世界。

模糊 c-均值(FCM)算法是模糊聚类的基本方法之一，它是一种聚类不定归属的方法。它通过引入隶属度函数来表示每个样本点属于各个类别的程度，从而决定样本点的类属，对数据进行软划分。

FCM 算法就是通过搜索目标函数的最小点，反复修改聚类中心矩阵和隶属度矩阵的分类过程。目前算法的收敛性已得到证明[1]，但它是一种局部搜索算法，对初值的选取十分敏感，如果初值选取不当，它容易收敛到局部极小点。且 FCM 对孤立点数据、样本分布不均衡也很敏感。鉴于此，提出基于减法聚类的改进的模糊 c-均值聚类，使得算法的收敛速度和准确性都得以改善。

1 模糊 c-均值算法分析

设样本空间为 $X=\{x_1,x_2,\cdots,x_n\}$，其中每个元素包含 s 个属性。模糊聚类就是将 x 划分为 c 类，c 个聚类中心为 $v=\{v_1,v_2,\cdots,v_c\}$。u_{ij} 是样本空间 X 中的第 j 个元素对第 i 个类中心的隶属度。$d_{ij}=\|v_i-x_j\|$ 是第 i 个聚类中心与第 j 个数据点之间的欧几里德距离，在 FCM 聚类算法中，隶属度矩阵和聚类中心分别为 $U=\{u_{ij}\}$ 和 $V=\{v_i\}$，FCM

* 基金项目：辽宁省教育厅基金项目(2009A350)

算法的目标函数为：

$$J(U,V)=\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^{m}\parallel v_i-x_j\parallel^2=\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^{m}d_{ij}^2 \tag{1}$$

其中，m 为模糊指数，$m\in[1,\infty]$，表示控制分类矩阵 U 的模糊程度，m 越大，分类的模糊程度越高，不做特殊要求时 m 一般取 2。FCM 算法就是求解使式(1)在满足条件 $u_{ij}\in[0,1]$，$\sum_{i=1}^{c}u_{ij}=1$ 和 $0<\sum_{i=1}^{c}u_{ij}<n$ 的情况下得到 J 的最小值。在求 J 的条件极值时，由拉格朗日乘数法求得隶属度为：

$$u_{ij}=\frac{1}{\sum_{k=1}^{c}\left(\frac{d_{ij}^2}{d_{kj}^2}\right)1/(m-1)} \tag{2}$$

聚类中心为 $v_i=\dfrac{\sum_{j=1}^{n}u_{ij}^{m}x_j}{\sum_{j=1}^{n}u_{ij}^{m}}$ (3)

2 基于减法聚类的改进的模糊 c-均值算法

2.1 初始聚类中心的选择

减法聚类是一种爬山法，它把所有的样本点作为聚类中心的候选点，其基本思想是计算每个样本点的密度指标，如果该样本点周围的点多，则密度指标就大，就选取密度指标最大的样本点作为聚类中心。减法聚类是一种快速独立的近似的聚类方法，用它计算，计算量由样本数目决定且与样本点的数目成简单的线性关系，而已与所考虑问题的维数无关。

M 维空间的 n 个样本点 $x_i(i=1,2,\cdots,n)$ 全部都为聚类中心的候选点，定义样本点 x_i 处密度指标为：

$$D_i=\sum_{j}^{n}\exp\left[-\frac{\parallel x_i-x_j\parallel^2}{(0.5r_a)^2}\right] \tag{4}$$

减法聚类的过程如下：

(1)用式(4)计算每个样本点 x_i 的密度指标，选择具有最高密度指标的数据点 x_{c1} 作为第一个聚类中心，D_{c1} 为其密度指标。其中 r_a 是一个正数，定义了该点的领域半径，半径以外的数据点对该点的密度指标贡献非常小，这里取：

$$r_a=\frac{1}{2}\min_{k}\{\max_{i}\{\parallel x_i-x_k\parallel\}\}。$$

(2) 令 x_{ci} 为第 i 次选出的聚类中心，D_{ci} 为其密度指标，则其他样本点的密度指标可用式(5)修正。选出密度指标最高的数据点 $x_{ci}+1$ 作为新的聚类中心。其中 r_b 是一个正数，定义了一个密度指标函数显著减小的领域，这里取 $r_b=1.2r_a$。

$$D_i=D_i-D_{ci}\exp\left[-\frac{\parallel x_i-x_{c_i}\parallel^2}{(0.5r_b)^2}\right]\quad i\neq c_i \tag{5}$$

(3) $\dfrac{D_{ci}+1}{D_{c1}}<\delta$ (6)

判断式(6)是否成立，若不成立，则转到步骤(2)；若成立则退出。预先给定参数 δ、r_a、r_b。δ 决定了最终产生的初始聚类中心数目，δ 越小，产生的聚类数越多；反之则聚类数越少。r_a，r_b 越大，产生的类数就越少，反之，则产生的类数就越多。

2.2 改进的 FCM 算法

(1) 为样本加权

样本空间为 $X=\{x_1,x_2,\cdots,x_n\}$，每个样本点对于分类结果来说贡献是不同的，例如样本空间中，孤立点就是对分类不重要的样本点，FCM 算法对于这一点不敏感。因此为了区分各个样本点的不同之处，给每个样本点赋予一个权值 w_i[4]。

$$w_i=\frac{\sum_{i=1}^{n}m}{m} \tag{7}$$

$$m=\frac{1}{n}\sum_{j=1}^{n}d(x_i,x_j) \tag{8}$$

则计算聚类中心的公式变为：

$$V_i=\frac{\sum_{i=1}^{N}w_i u_{ij}^{m}x_i}{\sum_{i=1}^{N}w_i u_{ij}^{m}} \tag{9}$$

其中 $d(x_i,x_j)$ 表示两个样本点 x_i 与 x_j 之间的欧式距离，$d(x_i,x_j)$ 的值越接近 0 则表示 x_i 与 x_j 之间越相似或越接近，则权重 w_i 越大；反之，x_i，x_j 差异性越大或越远，则权重 w_i 越小。如果样本点周围的点越多，则它的权重越大，因此可以用权重 w_i 表示第 i 个样本 x_i 对分类的影响程度。由于算法中噪声和孤立点的权重比较小，这样就能消除它们的影响。为样本加权后目标函数为：

$$J(U,V)=\sum_{i=1}^{c}\sum_{j=1}^{n}w_i u_{ij}^{m}d_{ij}^2 \tag{10}$$

(2) 修正隶属度矩阵

FCM 算法的思想是：迭代调整隶属矩阵和聚类中心使目标函数值最小，为保证 FCM 算法每次的迭代都朝着全局最优的方向逼近，其关键就在于保证确定 V 的下一次迭代值，加快收敛于全局最优点的速度。在此采用修正隶属矩阵来计算下一次迭代的聚类中心，使得到的 V 更靠近聚类中心，更合理，从而提高 FCM 算法的收敛速度。因此修正隶属度矩阵[5]可以提高聚类速度，使聚类效果更好。

样本离聚类中心距离越远属于该聚类中心的程度越小，反之越大，样本对类中心的影响即称为样本对类中心施加的吸引力，在这里设定了一个抑制因子，由它来控制对离样本点次最近的类中心的抑制作用。

当 $\alpha=1$ 时，算法退化为 FCM 算法，对离样本点次最近的类中心没有任何抑制作用。

当 $\alpha=0$ 时，算法完全抑制了样本对离它次最近类中心的吸引力，对离样本最近类中心的吸引力的增强力度

最大。

当 $1<\alpha<0$ 时，算法对离样本次最近类中心的吸引力有一定的抑制作用，对离样本最近类中心的吸引力有一定的增加作用。

修正隶属度矩阵的过程如下：

(1) 初始化类中心为 $V(0)$。迭代次数 $L=0$ 给定模糊指数 m，$m\in(1,\infty)$ 置吸引力抑制因子 α（即样本点对离它最近的类的吸引力），$\alpha\in[0,1]$。

(2) 计算出 $\boldsymbol{U}(L)$:

$I_j=\{i|1\le i\le C,d_{ij}=0\},\bar{I}_j=\{1,2,\cdots,C\}-I_j$

当 $I_j=\phi$ 时；

$$u_{ij}=\frac{1}{\sum_{k=1}^{c}\left(\frac{d_{ij}^2}{d_{kj}^2}\right)^{1/(m-1)}} \tag{11}$$

当 $I_j\neq\phi$ 时，$\forall\, i\in\bar{I}_j$，$u_{ij}=0$，

且 $\sum_{i\in I_j}u_{ij}=1$ (12)

(3) 修正隶属度矩阵 $\boldsymbol{U}(L)$：假设样本 x_i 对第 q 类的隶属度最大，值为 u_{qi}；它对第 s 类的隶属度次最大，值为 u_{si}。

对其进行修正后，样本 x_i 对第 q 类的隶属度为：

$$u_{qi}=u_{qi}+(1-\alpha)u_{si} \tag{13}$$

对第 s 类的隶属度为：

$$u_{qi}=\alpha u_{si} \tag{14}$$

除此之外各类的隶属度不变。

(4) 用修正后的 $U(L)$ 计算下一次的迭代中心 $V(L+1)$（加权后的 V_i）。

$$V_i=\frac{\sum_{i=1}^{N}w_i u_{ij}^{m}x_i}{\sum_{i=1}^{N}w_i u_{ij}^{m}} \tag{15}$$

(5) 判断是否终止迭代。终止而退出，否则，$L=L+1$，返回步骤(2)，继续迭代。

经过对隶属度矩阵的修正可知：改进后的算法，样本点增大了对离它最近的类中心的吸引力强度；样本点减小了对离它次最近的类中心的吸引力强度，从而减弱了离样本次最近类中心对离样本最近的类中心收敛速度的延缓作用。对其余类中心的吸引力强度不变，从而提升了 FCM 算法的收敛速度。

2.3 基于减法聚类改进的模糊 c-均值算法过程

为保证改进的 FCM 聚类结果为全局最优解，采用减法聚类的聚类中心作为改进的 FCM 聚类的初始聚类中心。算法步骤如下：

(1) 设定聚类参数：领域的半径 r_a、r_b，比例参数 δ，FCM 聚类数 c，模糊指数 m 和最小误差 ε，迭代次数 L，吸引力抑制因子 α。

(2) 应用式(4)计算所有样本点的密度指标，将密度指标最高的一个作为第一个聚类中心点 x_{c1}。

(3) 依据公式(5)利用减法步骤(2)中的 x_{c1} 进一步计算余下的 $n-1$ 个数据点的密度指标，找出最高的作为第二个聚类中心 x_{c2}，依此类推，找到 p 个聚类中心，从中选取前 c 个作为 FCM 的初始聚类中心 $v(0)$。

减法聚类中心中，密度指标越大的聚类中心出现得越早，越有可能成为改进的 FCM 初始聚类中心。所以，当聚类数为 c 时，取减法聚类产生的前 c 个聚类中心作为改进的 FCM 的初始中心，无须再重新初始化，从而提高了聚类的效率。

(4) 求式(10)的最小值

(5) 按式(11)和式(12)计算出隶属度 $\boldsymbol{U}(L)$

(6) 依据式(13)和式(14)修正隶属度矩阵 $\boldsymbol{U}(L)$。

(7) 依据式(15)，用修正后的 $U(L)$ 计算下一次的迭代中心 $V(L+1)$。

(8) 判断是否满足终止迭代条件。对给定的阈值，$\|U(L+1)-U(L)\|<\varepsilon$ 如果终止而退出，否则，$L=L+1$，返回步骤(5)，继续迭代。

3 仿真与结果分析

为验证基于减法聚类的改进的 FCM 算法的效果，利用 Iris 植物样本数据进行仿真实验，将结果与传统 FCM 进行对比。Iris 数据集是公认的最适用于数据挖掘的数据集，它有四个属性、三种植物种类(setosa、versicolor、virginica)，每个种类含有 50 个样本。Iris 的实际中心分别为(6.588、2.974、5.552、2.026)、(5.006、3.418、1.464、0.244)、(5.936、2.77、4.26、1.326)。分别用传统的 FCM 和基于减法聚类的改进的 FCM 对 Iris 数据集进行聚类分析。实验中，设定允许最小误差 ε 均为 10^{-3}，模糊指数 $m=2$，$r_a=0.5$，$r_b=0.6$，$\alpha=0$，Iris 数据集的聚类结果如图 1、图 2 所示。Iris 数据集的比较如表 1 所示。

从图 1、图 2 与表 1 中可以看出，传统 FCM 与本文中的算法相比迭代次数少、搜索速度更快、聚类平均准确率更高。

基于减法聚类的改进的 FCM 算法很好地解决了 FCM 算法对初始值敏感及易陷入局部最优的问题，同时

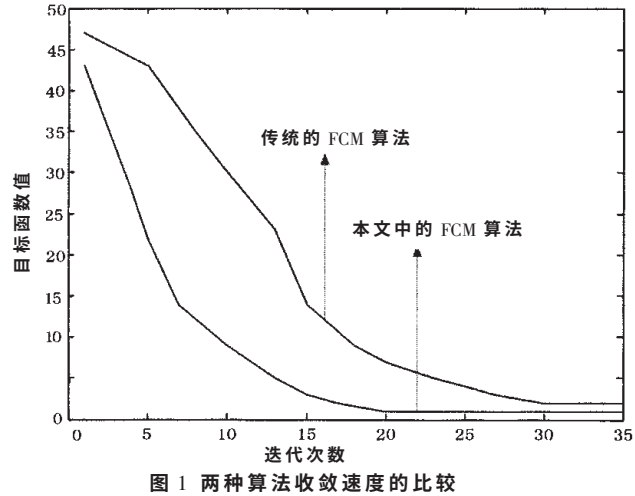

图 1 两种算法收敛速度的比较

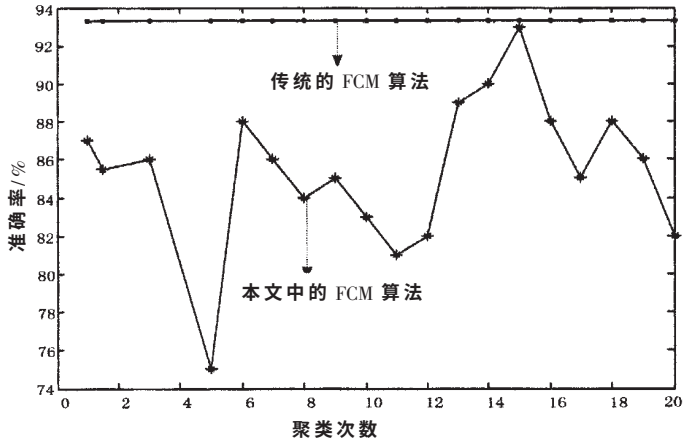

图 2 迭代 20 次两种算法准确率上的比较

表 1 Iris 数据集的性能比较

算法	迭代次数	聚类中心				平均准确率/%
FCM	30	6.628 6	3.013 2	5.256 7	1.924 5	87.1
		5.100 3	3.371 2	1.642 1	0.342 1	
		5.578 2	2.876 5	3.762 3	1.092 6	
基于减法聚类的改进的 FCM	20	6.598 0	3.002 1	0.530 6	2.031 8	93.34
		5.003 7	3.424 5	1.459 8	0.245 7	
		5.875 3	2.732 2	4.224 2	1.327 8	

也改善了 FCM 对孤立点敏感的问题，提高了聚类的速度，具有很高的实用价值。

参考文献

[1] GAMES R A, CHAN A H. A fast algorithm for determining the linear complexity of a pseudorandom sequence with period 2n[J].IEEE Trans Inf Theory ,1983,IT-29(1):144-146.

[2] HAND D, MANNILA H, SMYTH P. Principles of data mining [M].Cambridge MA:MITPress,2001.

[3] PAL N R, CHAKRABORTY D. Mountain and subtractive clustering method; Improvements and Generalization. International Journal of Intelligent Systems , 2000,15 (4):329-341.

[4] 齐淼，张化祥.改进的模糊 c-均值聚类算法研究[J].计算机工程与应用,2009,45(20).

[5] 闫兆振.自适应模糊 c-均值聚类算法研究[D]. 济南：山东科技大学,2006.

(收稿日期：2010-03-18)

作者简介：

于迪，女，1983 年生，硕士研究生，主要研究方向：数据库理论与研究。

李义杰，男，1954 年生，教授，硕士生导师，主要研究方向：数据库理论与研究。