模糊 C-均值聚类算法的改进

王小姣，徐夫田，单国杰
(山东师范大学 信息科学与工程学院，山东 济南 250014)

摘　要：针对传统的模糊 C-均值算法 FCM 受初始聚类中心影响而易于收敛到局部极小值的问题，提出了具体的改进方法。初始聚类中心不再随机获取而是通过改进的算法有目的地进行选取，同时采用冗余聚类中心的方法先将大簇分割成多个小类，再按一定条件将相邻的小类合并。实验结果表明，改进后的 FCM 算法减小了对初始聚类中心的依赖，聚类结果更加精确。

关键词：聚类；模糊 C-均值；初始聚类中心

中图分类号：TP391　　**文献标识码**：A　　**文章编号**：1674-7720(2010)12-0042-03

Improvement of fuzzy C-means clustering algorithm

WANG Xiao Jiao, XU Fu Tian, SHAN Guo Jie
(School of Information Science & Engineering, Shandong Normal University, Jinan 250014, China)

Abstract: The traditional Fuzzy C-means algorithm has the shortage that be sensitive to the initial cluster centers, easily converge to a local minimum result. To solve the above problem, this paper presents an improved algorithm. Through the improved algorithm, the initial cluster centers are purposefully selected but not randomly selected. At the same time we can express a big cluster used several small clusters, then merger adjacent clusters that satisfy certain conditions. Experiment result demonstrates that the improved FCM algorithm can decrease the dependence on the initial cluster centers and get more accurate clustering results.

Key words: clustering; fuzzy C-means; initial cluster centers

模糊聚类算法现已广泛应用于数据挖掘、模式识别、图像分割等领域，具有巨大的实用价值[1]。在众多的模糊聚类算法中，模糊 C-均值 FCM(Fuzzy C-Means)算法应用最为广泛且比较成功。模糊 C-均值算法是在传统的 C-均值算法的基础上结合模糊集合理论而得到的一种柔性的模糊划分算法，它有别于传统 C-均值算法“非此即彼”的硬划分。FCM 算法中待划分的数据样本点以不同的隶属度归属于每一类，通过优化目标函数可得到每个数据样本点对所有类中心的隶属度，从而决定样本点的归属，以达到自动对数据样本分类的目的[2]。同时，FCM 算法有一些自身的缺点：(1)聚类的类数不能自动确定，需要随机给出初始中心个数，而且此过程没有准则可遵循；(2)由于 FCM 算法本质是用梯度下降方法寻求最优解，因此很容易陷于局部最优值，同时受初始值的影响较大；(3)算法中距离衡量尺度一般选取的是欧氏距离，适合于识别团状或超球体簇类，而对于其他不规则簇的识别存在很大缺陷；(4)多数情况下对噪声数据较敏感。

针对 FCM 算法存在的不足，研究人员对算法进行了许多改进。对于问题(2)中存在的缺点，人们把进化计算的思想引入 FCM，以期达到全局最优的目的，主要的方法有模拟退火算法[3]、遗传算法[4]等。针对噪音数据敏感问题，参考文献[5]中提出了属性加权 FCM 算法，该算法根据属性重要性的不同为样本数据加上不同的权重值，取得了一定的效果。

本文提出一种改进的 FCM 算法，将随机改为在全局范围内有目的地选取初始聚类中心，同时采用冗余聚类中心的方法将一个大类用多个中心点来表示，而后再合并适当的小类。此算法可以避免随机求取聚类初始中心时算法收敛到局部最优的情况，降低对初始聚类中心的依赖性。

1 模糊 C-均值算法基本思想

模糊 C-均值算法[6,7]FCM 于 1974 年由 Dunn 提出并由 Bezdek 加以推广，是目前被广泛采用的聚类算法之一。该算法基本思想如下所述：

给定样本数据集合 $X=\{x_1,x_2,\cdots,x_n\}$，其中每个样本包含 p 个属性。FCM 算法是把 n 个数据元素 x_i $(i=1,2,\cdots,n)$ 划分成 $c(2\leqslant c\leqslant n)$ 个模糊簇，并求每个簇的类中

心 $v_i(i=1,2,\cdots,c)$，使目标函数达到最小。FCM 聚类的目标函数为：

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

函数满足条件 $\sum_{i=1}^{c}u_{ij}=1, u_{ij}\in[0,1](i=1,2,\cdots,c; j=1,2,\cdots,n)$

$\boldsymbol{U}=\{u_{ij}\}$ 表示各数据样本点对应于各个聚类中心的隶属度矩阵，u_{ij} 为第 j 个数据点在第 i 类中的隶属度。$\boldsymbol{V}=\{v_i\}(i=1,2,\cdots,c)$ 表示类中心矩阵。$d_{ij}=\|x_j-v_i\|$ 为样本点与类中心之间的欧氏距离。m 为模糊加权指数 $(1\leqslant m\leqslant\infty)$，$m$ 值越大，分类的模糊程度越高，通常情况下取 $m=2$。目标函数 $J(U,V)$ 为每个数据样本点到各聚类中心的加权距离的平方和。

FCM 的实质就是一个将目标函数 $J(U,V)$ 最小化的迭代收敛过程。为使目标函数 $J(U,V)$ 的值达到最小，隶属度和类中心用(2)和式(3)式来更新：

$$u_{ij}=\left(\sum_{k=1}^{c}\left(\frac{\|x_j-v_i\|}{\|x_j-v_k\|}\right)^{\frac{2}{m-1}}\right)^{-1} \tag{2}$$

$$v_i=\frac{\sum_{j=1}^{n}u_{ij}^{m}x_j}{\sum_{j=1}^{n}u_{ij}^{m}} \tag{3}$$

算法具体步骤如下：

(1)预先给出聚类数目 c，模糊参数 m，迭代终止条件 ξ 的值；

(2)随机初始化聚类中心 $v_i(k)(i=1,2,\cdots,c)$，k 为循环次数，此处 $k=1$；

(3)利用公式(2)计算 u_{ij} 得到隶属度矩阵 $\boldsymbol{U}$；

(4)用公式(3)修正类中心 $v_i(k+1)$；

(5)误差 $e=\sum_{i=1}^{c}\|v_i(k+1)-v_i(k)\|^2$，若 $e<\xi$ 则转步骤(6)；否则令 $k=k+1$ 转步骤(3)；

(6)算法结束，输出聚类结果 (U,V)。

算法结束时便得到了每个类的聚类中心以及每个特征向量对应于每个类的隶属度，从而可以对样本进行归类，若 $u_{ij}>u_{wj}(w、i=1,2,\cdots,c, i\neq w)$ 则将元素 x_j 归入第 i 类。

2 算法改进

2.1 聚类中心初始值的选取

FCM 对于初始聚类中心的依赖非常大。由于算法是沿使目标函数减小的方向进行逐次迭代的，而目标函数可能存在多个局部极小点，因此若初始化值落在一个局部极小点附近，就可能会使算法收敛到局部极小。本文提出一种改进措施，在选择初始聚类中心时不再是传统的随机设置，而是按照一定的规则在全局范围内有目的地选择。

给出待分类样本集合 $X=\{x_1,x_2,\cdots,x_n\}$，设定类间的最小距离阈值为 $\delta(\delta>0)$，可按以下步骤选择聚类中心：

(1)首先计算 n 个样本数据两两间的欧式距离，并生成距离矩阵 $\boldsymbol{D}$，选出矩阵中距离最近的两个数据样本，计算两个样本的中点并将其作为第 1 个聚类中心；

$$\boldsymbol{D}=\begin{bmatrix}0\\ d(2,1) & 0\\ d(3,1) & d(3,2) & 0\\ \cdots & \cdots & \cdots\\ d(n,1) & d(n,2) & \cdots & 0\end{bmatrix}$$

$$\begin{cases}d(i,j)\geqslant 0,\ d(i,j)=\|xi-xj\|\\ d(i,i)=0\\ d(i,j)=d(j,i)\end{cases}$$

(2)利用得到的距离矩阵 $\boldsymbol{D}$，找出与第 1 类中得到的两个数据样本的距离均大于所设定的阈值 δ 的所有数据样本，并在这些样本中找出距离最近的两个数据样本，取它们的中点作为第 2 个聚类中心；

(3)按相同方法，在剩余的样本中找出与前面已找到的样本的距离都大于阈值 δ 的所有样本，并在其中选出距离最短的两个样本，取它们的中点作为当前的聚类中心；

(4)重复步骤(3)，直到找出第 c 个聚类中心(c 为要初始化的聚类数)。

在本文提出的选取初始类中心的方法中，考虑到了对聚类中心的约束问题。为确保各聚类中心之间的分离性，可预先设定一个各中心间的最小距离阈值，即上文中的 δ，尽量使得初始聚类中心之间的距离大于所设定的阈值。这样在初始化时就可以在全局范围内对聚类中心进行求取，避免了 FCM 算法随机求取初始聚类中心时易收敛到局部极小值的情况。同时需注意的是，对于阈值 δ，样本点中有时可能不存在距离在 δ 之内的一对样本，此时就需要调整 δ 的取值。

2.2 小类合并思想

FCM 算法需要预先确定聚类数 c，在实际问题中对于一个给定的未知数据集，聚类数目是不能事先确定的。为克服这一缺陷，本文采用冗余聚类中心的方法，即先提供一个足够大的初始聚类个数 q，而最终的聚类结果可通过合并适当的小类确定。对于 q 的选择，并不是越大越好，q 值过大反而会影响算法的执行效率。

小类的合并：对于 FCM 算法生成的 q 个簇，如果两个簇中心点之间的欧式距离小于一个事先给定的值，且这两个簇各自包含的样本数据到另一个簇的中心点距离很接近，则此时可以将它们看做一个簇而加以合并。

从 FCM 算法得到的模糊矩阵 $\boldsymbol{U}$ 中找出各个类所包含的样本集合，$R=(r_{ik})$，其中 r_{ik} 表示属于第 i 类的 k 个样本点。

$$d'_{ij}=\|v_i-v_j\| \tag{4}$$

$$w_{ij}=\left|\frac{\sum_{j=1}^{q}\|R_{jk}-v_i\|}{\sum_{i=1}^{q}\|R_{ik}-v_j\|}\right| \tag{5}$$

循环计算 d'_{ij} 与 w_{ij} 的值，如果得到 $d'_{ij}<\varepsilon_1, w_{ij}<\varepsilon_2$，表

示符合合并条件，将第 i 类与第 j 类合并为一类。

2.3 改进后的算法流程

(1)预先给出足够大的初始聚类中心数 q、模糊参数 m，迭代终止条件 ξ 的值；

(2)按照本文 2.1 中选取初始聚类中心的方法找出 q 个初始类中心，记为 $v_i(k)(i=1,2,\cdots,q;k=1)$；

(3)用 FCM 算法聚类，并输出聚类结果(U,V)；

(4)在得到的聚类结果中按公式(4)和公式(5)计算，把满足条件的小类进行合并；

(5)算法结束，并输出合并后的最终聚类结果(U,V)。

3 实验结果分析

为了验证改进后的 FCM 算法的有效性，接下来采用一组人造的数据集合进行实验，数据集合如图 1 所示。具体方法是先按照传统 FCM 算法对数据集聚类，预先给出聚类数 c 为 4，并在集合中随机选出 4 个初始中心，得到的聚类结果如图 2 所示。然后用改进后的 FCM 算法对同一数据集聚类，此时给出冗余中心数 q 为 8，第一步得到冗余中心为 8 时的聚类结果，如图 3 所示。给出条件 $\varepsilon_1=0.5$，$\varepsilon_2=0.8$ 后，经过最终的类合并得到聚类结果如图 4 所示。

通过实验可以看出传统的 FCM 算法受初始值影响可能得不到精确的聚类结果。图 2 中虽然最终数据集合被划分为 4 类，但是显然第 2 类和第 4 类的划分范围是不合理的，因此得到的结果不理想。改进后的 FCM 算法

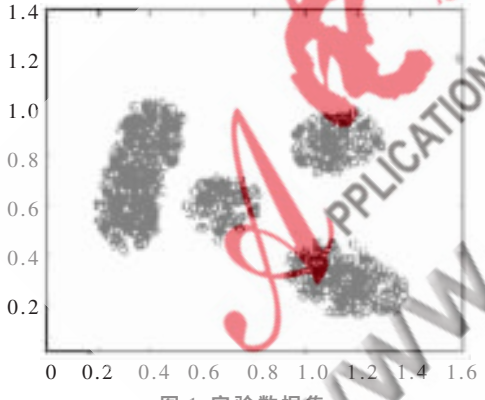

图 1 实验数据集

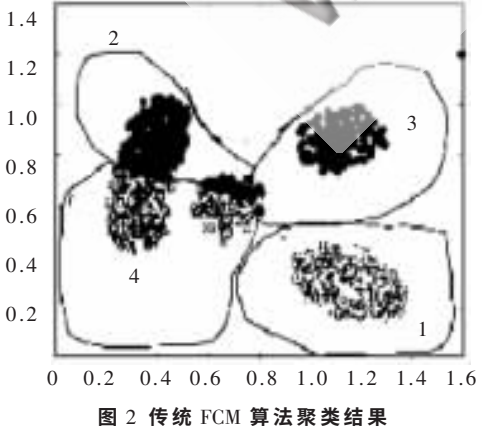

图 2 传统 FCM 算法聚类结果

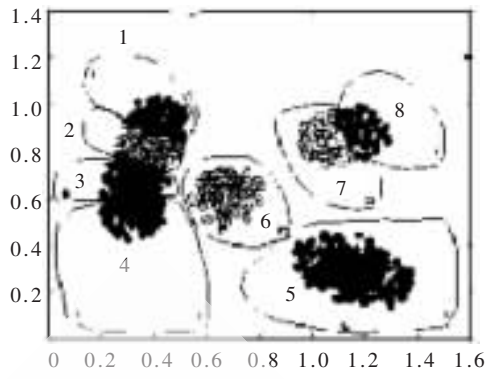

图 3 改进后的 FCM 算法聚类结果(冗余中心数 $q=8$)

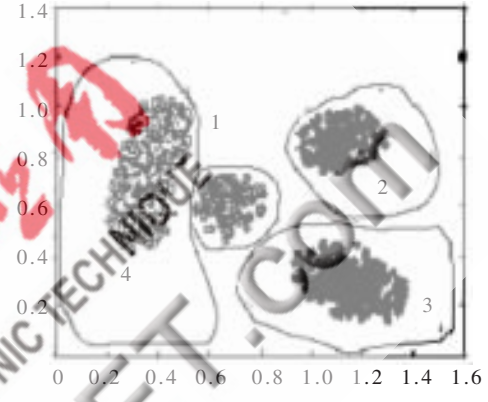

图 4 合并小类后的最终聚类结果

先输入一个足够大的聚类数 8，并在全局范围内有目的地选取初始类中心，用 FCM 算法计算得到的结果(图 3)属于有效聚类。经过合并后的最终聚类结果(图 4)能更准确地对数据集进行划分，降低了算法对初始中心的依赖而得到全局最优的结果。

本文通过一定的改进措施解决了 FCM 算法对初始值敏感易收敛到局部极小值的问题，并通过实验验证了算法的有效性。但由于 FCM 算法本身还存在一些缺陷，需要人们进一步地探索研究，接下来的工作可从算法的个别参数设置方面入手对其进行相应的优化。

参考文献

[1] 高永清，陈志红，黄鹤玲，等.基于 FCM 的无监督最优模糊聚类算法[J]. 信息技术，2009(07)：69-71.

[2] 张敏，于剑.基于划分的模糊聚类算法[J].软件学报，2004,15(6)：858-868.

[3] ROSE K, GUREWITZ E, FOX G C. Constrained clustering as optimization method [J]. IEEE Trans. on Pattern Analysis and Ma-chine Intelligence,1993,15(8):785-794.

[4] BUCKLES B P, PETRY F E, PRABBU D, et al. Fuzzy clustering with genetic search[C]//Proc IEEE Conf Evol Comput Proc ICEC. Piscataway：IEEE Press, 1994：46-50.

[5] WANG X Z, WANG Y D, WANG L J. Improving fuzzy C-means clustering based on feature-weight learning[J]. Pattern Recognition Letters,2004,25 (10)：1123-1132.

[6] 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,2004:56-67.

[7] 张新波.两阶段模糊C-均值聚类算法[J].电路与系统学报,2005,10(2):117-121.

(收稿日期:2010-04-07)

作者简介:

王小姣,女,1985年生,硕士研究生,主要研究方向:Web数据挖掘;

徐夫田,男,1965年生,研究员,主要研究方向:税务系统信息化;

单国杰,女,1985年生,硕士研究生,主要研究方向:信息管理与数据挖掘技术。