# 一种改进的模糊C均值聚类算法

李 雷,罗红旗,丁亚丽
(南京邮电大学 自动化学院,江苏 南京 210003)

**摘 要**:针对经典的C均值聚类算法以及模糊C均值聚类算法所存在的两个方面的问题:一是算法对初始聚类中心的过分依赖性,通常的聚类算法往往对于不同的初始聚类中心会得到不同的聚类结果;二是算法需要预先知道实际的聚类数目,而在实际应用中,聚类数目却是未知的。基于此提出了模糊C均值聚类算法的一种改进算法,即在标准的模糊C均值聚类算法的基础上,给目标函数加入了一个惩罚项,使得上述问题得以解决。并通过仿真实验证实了新算法的可行性和有效性。

**关键词**:聚类分析;模糊;C均值
**中图分类号**:TP301.6 **文献标识码**:A **文章编号**:1673-629X(2009)12-0071-03

# A Novel FCM Clustering Algorithm

LI Lei, LUO Hong-qi, DING Ya-li
(Institute of Automatic Control Engineering of Nanjing Univ. of Posts and Telecommunications, Nanjing 210003, China)

**Abstract**: There are two issues in the application of FCM clustering algorithm: one is that the FCM algorithm is too sensitive to the initial cluster centers, people can get different clustering result from different original clustering center, and the other is that the number of the clusters C needs to be determined in advance as an input to the algorithm, but C always does not be known. Based on this, a novel algorithm of FCM is proposed in this paper. Based on the FCM, a penalty term is added into the objective function and the above-mentioned issues can be resolved. The simulation demonstrates the feasibility and validity of the proposed method.

**Key words**: cluster analysis; fuzzy; C-means

## 0 引 言

将物理或抽象对象的集合分成相似的对象类的过程称为聚类[1~4]。聚类是一种非监督模式识别问题,它是指按照某种相似性的度量,使相似的样本对象归为相同的类,不相似的样本对象归为不同的类。聚类分析已经被广泛地应用于许多领域,包括市场研究、模式识别、数据分析和图像处理等。

常用的聚类算法有经典C均值算法和它的改进版本模糊C均值聚类算法。然而它们都存在两个问题[5~7],首先必须预先知道聚类数目C的值,在实际应用中,C的值却是未知的。另外,C均值聚类算法对初始点很敏感,往往对于不同的初始聚类中心会得到不同的聚类结果。基于此,提出了模糊C均值聚类算法的一种改进算法,在标准的模糊C均值聚类算法的基础上,给目标函数加入了一个惩罚因子,从而解决了聚类程序对初始聚类中心过于敏感的问题。通过试验证实了新的算法在聚类结果一致性,以及初始聚类数目的确定上,都有很好的效果。

## 1 算法介绍

### 1.1 C均值算法介绍

C均值算法是一种基于样本间相似性度量的间接聚类方法,属于非监督学习方法,此算法以 $C$ 为参数,把 $N$ 个对象分为 $C$ 个簇,使簇内具有较高的相似度,而簇间的相似度较低。相似度的计算根据一个簇中对象的平均值(被看作簇的重心)来进行。此算法首先随机选择 $C$ 个对象,每个对象代表一个聚类的质心。对于其余的每一个对象,根据该对象与各聚类质心之间的距离,把它分配到与之最相似的聚类中。然后,计算每个聚类的新质心。重复上述过程,直到准则函数会

**收稿日期**:2009-03-10;**修回日期**:2009-06-20
**基金项目**:国家自然科学基金项目(10371106, 10471114);江苏省高校自然科学基金项目(04KJB110097, 08KJB520003);南京邮电大学攀登计划(NY207064)
**作者简介**:李 雷(1958-),男,安徽砀山人,教授,研究方向为智能信号处理、非线性分析与计算智能。

聚。C 均值算法是一种较典型的逐点修改迭代的动态聚类算法，其要点是以误差平方和为准则函数。逐点修改类中心：一个样本按某一原则，归属于某一组类后，就要重新计算这个组类的均值，并且以新的均值作为凝聚中心点进行下一次聚类；逐批修改类中心：在全部样本按某一组的类中心分类之后，再计算修改各类的均值，作为下一次分类的凝聚中心点。

C 均值聚类的核心思想是：算法把 $n$ 个向量 $\boldsymbol{x}_j(j=1,2,\cdots,n)$ 分为 $c$ 个组 $G_i(i=1,2,\cdots,c)$，并求每组的聚类中心，使得非相似性（或距离）指标的价值函数（或目标函数）达到最小。

### 1.2 模糊 C－均值(FCM)聚类算法

聚类算法分为硬聚类算法和软聚类算法[8~10]。硬聚类算法将每个数据对象归到一个类，但是数据对象往往具有大量性和多样性的特点，经常被归到几个类中，归属于每个类的程度也不相同。结合这一点就出现了 C 均值的改进算法——FCM 聚类算法。

FCM 是由 Ruspini 和 Bezdek 于 1981 年提出的，目前被广泛应用。模糊 C 均值因算法设计简单，解决问题范围广，易于计算机实现，所以被应用于很多领域。它是一种非常有效的模糊聚类算法，使用每个样本隶属于某个聚类的隶属度，即使对于很难分类的变量，FCM 也能够得到比较满意的聚类效果。

假如给定了数据对象 $X=\{x_1,x_2,\cdots,x_n)\subset R^d$ 集合为一有限数据对象集合，元素 $x_i$ 为 $d$ 维向量。FCM 算法的基本思想是找到一个模糊划分矩阵 $(\boldsymbol{u}_{ij})_{c\times n}$ 以及 $c$ 个类中心点 $V=\{v_1,v_2,\cdots,v_c\}$，使得

$$J=\min\sum_{i-1}^{c}\sum_{j-1}^{n}(\boldsymbol{u}_{ij})^m d^2(x_j,v_i) \tag{1}$$

满足：$\sum_{i-1}^{c}\boldsymbol{u}_{ij}=1$，$\sum_{j-1}^{n}\boldsymbol{u}_{ij}\leqslant n$，$0\leqslant\boldsymbol{u}_{ij}\leqslant 1$，$i=1,2,\cdots,c$，$j=1,2,\cdots,n$。

其中 $m\in[1,+\infty)$ 称为模糊加权系数。$u_{ij}$ 称为隶属度，$x_j$ 表示向量隶属于中心点的程度。$d(x_j,v_i)$ 是目标数据 $x_j$ 与 $v_i$ 的距离。

### 1.3 改进的模糊 C 均值聚类算法

假设数据集合 $X=\{X_1,X_2,\cdots,X_n\}$ 包含 $n$ 个对象，其中每个对象表示 $[x_{i1},x_{i2},\cdots,x_{im}]$，$m$ 表示属性的数目。

为了使新的模糊 C 均值算法将 $X$ 聚为 $C$ 类，需要将下面的目标函数最小化：

$$P(\boldsymbol{U},\boldsymbol{Z})=\sum_{j=1}^{c}\sum_{i=1}^{n}\boldsymbol{u}_{ij}D_{i,j}+\lambda\sum_{j=1}^{c}\sum_{i=1}^{n}\boldsymbol{u}_{i,j}\log\boldsymbol{u}_{i,j}$$

s.t.

$$\sum_{i=1}^{c}\boldsymbol{u}_{i,j}=1,\boldsymbol{u}_{i,j}\in(0,1],1\leqslant i<n \tag{2}$$

式中 $\sum_{j=1}^{c}\sum_{i=1}^{n}\boldsymbol{u}_{ij}D_{i,j}$ 是标准的模糊 C 均值聚类算法的目标函数，$\lambda\sum_{j=1}^{c}\sum_{i=1}^{n}\boldsymbol{u}_{i,j}\log\boldsymbol{u}_{i,j}$ 就是我们引入的惩罚因子，用来最大化聚类对象与类中心的差异性。

其中 $\boldsymbol{U}=[\boldsymbol{u}_{i,j}]$ 表示一个 $n\times c$ 的矩阵，$\boldsymbol{u}_{i,j}$ 表示第 $i$ 个对象 $x_i$ 对于第 $j$ 个类 $z_j$ 的隶属度，$\boldsymbol{Z}=[\boldsymbol{z}_1,\boldsymbol{z}_2,\cdots,\boldsymbol{z}_c]^{\mathrm{T}}$ 表示一个包含了所有聚类中心的 $k\times m$ 矩阵。$D_{i,j}$ 表示了第 $i$ 个对象对于第 $j$ 个聚类中心的差异度，在此，像许多聚类算法中一样，使用标准的欧式距离的平方来度量这个差异性：

$$D_{i,j}=\sum_{l=1}^{m}(z_{j,l}-x_{i,l})^2$$

在(2)式中，能够通过扩展标准的 C 均值聚类程序来最小化 $P$。对于 $P$ 最小化的常用方法是对 $\boldsymbol{U}$ 和 $\boldsymbol{Z}$ 使用局部最优化方法。在这种方法中，首先固定 $\boldsymbol{U}$ 的值，根据 $\boldsymbol{Z}$ 来缩小 $P$，然后固定 $\boldsymbol{Z}$ 的值，根据 $\boldsymbol{U}$ 来缩小 $P$。

固定 $\boldsymbol{U}$ 的值，根据下式更新 $\boldsymbol{Z}$ 的值：

$$z_{jl}=\frac{\sum_{i=1}^{n}\boldsymbol{u}_{i,j}x_{i,l}}{\sum_{i=1}^{n}\boldsymbol{u}_{i,j}},1\leqslant j\leqslant k,1\leqslant l\leqslant m \tag{3}$$

可以发现，(3)式不依赖于参数 $\lambda$ 的值。

固定 $\boldsymbol{Z}$ 的值，根据下式更新 $\boldsymbol{U}$ 的值：

$$\widetilde{P}(\boldsymbol{U},\boldsymbol{\alpha})=\sum_{j=1}^{c}\sum_{i=1}^{n}(\boldsymbol{u}_{i,j}D_{i,j}+\lambda\boldsymbol{u}_{i,j}\log\boldsymbol{u}_{i,j})+\alpha_i\sum_{i=1}^{n}(\sum_{j=1}^{c}\boldsymbol{u}_{i,j}-1)$$

其中 $\boldsymbol{\alpha}=[\alpha_1,\cdots,\alpha_n]$ 表示含有拉格朗日乘子的向量。如果要使 $\widetilde{P}(\boldsymbol{U},\boldsymbol{\alpha})$ 取得最小值 $(\hat{\boldsymbol{U}},\hat{\boldsymbol{\alpha}})$，必须使两个变量集合的梯度都为零。因此：

$$\frac{\partial\widetilde{P}(\hat{\boldsymbol{U}},\hat{\boldsymbol{\alpha}})}{\partial\hat{\boldsymbol{u}}_{i,j}}=D_{i,j}+\lambda(1+\log\boldsymbol{u}_{i,j})+\hat{\alpha}_i=0 \tag{4}$$

$1\leqslant j\leqslant k,1\leqslant i\leqslant n$

$$\frac{\partial\widetilde{P}(\hat{\boldsymbol{U}},\hat{\boldsymbol{\alpha}})}{\partial\hat{\boldsymbol{\alpha}}_i}=\sum_{i=1}^{c}\hat{\boldsymbol{u}}_{i,j}-1=0 \tag{5}$$

由(4)式，可以得到：

$$\boldsymbol{u}_{i,j}=\exp(\frac{-D_{i,j}}{\lambda})\exp(-1)\exp(\frac{-\hat{\boldsymbol{\alpha}}_i}{\lambda}) \tag{6}$$

将(6)式代入(5)式中，可得：

$$\sum_{j=1}^{c}\boldsymbol{u}_{i,j}=\sum_{j=1}^{c}\exp(\frac{-D_{i,j}}{\lambda})\exp(-1)\exp(\frac{-\hat{\boldsymbol{\alpha}}_i}{\lambda})$$

$$=\exp(-1)\exp(\frac{-\hat{\boldsymbol{\alpha}}_i}{\lambda})\sum_{j=1}^{c}\exp(\frac{-D_{i,j}}{\lambda})=1 \tag{7}$$

其中：

$$\hat{\boldsymbol{u}}_{i,j}=\frac{\exp(\frac{-D_{i,j}}{\lambda})}{\sum_{l=1}^{c}\exp(\frac{-D_{l,j}}{\lambda})}$$

通过此式来更新 $\boldsymbol{U}$ 的值。

改进的模糊 C 均值聚类算法：

Step1:给出惩罚因子 $\lambda$,随即选取 $\boldsymbol{Z}^{(0)}=\{Z_1,Z_2,\cdots,Z_c\}$作为初始聚类中心,给出 $\boldsymbol{U}^{(0)}$的值,并根据(7)式计算 $P(\boldsymbol{U}^{(0)},\boldsymbol{Z}^{(0)})$。令 $t=0$;

Step2:令 $\hat{\boldsymbol{Z}}=\boldsymbol{Z}^t$,通过 $P(\boldsymbol{U},\hat{Z})$得到 $\boldsymbol{U}^{t+1}$。如果 $P(\boldsymbol{U}^{t+1},\hat{\boldsymbol{Z}})=P(\boldsymbol{U}^t,\hat{\boldsymbol{Z}})$,输出$(\boldsymbol{U}^t,\hat{\boldsymbol{Z}})$,停止;否则,转 Step3;

Step3:令 $\boldsymbol{U}^t=\boldsymbol{U}^{t+1}$,通过 $P(\hat{\boldsymbol{U}},\boldsymbol{Z})$得到 $\boldsymbol{Z}^{t+1}$。如果 $P(\hat{\boldsymbol{U}},\boldsymbol{Z}^{t+1})=P(\hat{\boldsymbol{U}},\boldsymbol{Z}^t)$,输出$(\hat{\boldsymbol{U}},\boldsymbol{Z}^t)$,停止;否则,令 $t=t+1$,转 Step2。

## 2　仿真实验

为了测试文中提出的改进的模糊聚类算法的性能,进行了如下的测试实验。

实验使用的样本集合是二维空间中的一个包含 1500 个点的数据集。它有 4 个中心点,分别是(1,5),(1,1),(2,6),(5,5)。

首先假定 $C=6$,并且选择同一类中的6个点作为初始聚类中心(有意这样选择,来测试算法的性能)。接下来逐渐改变 $\lambda$ 的值。

在图 1 中给出了新算法对于测试数据集,使用不同的 $\lambda$ 值所得到的不同的聚类结果。可以发现当 $\lambda$ 很小时,聚类数目等于初始聚类中心的数目。随着 $\lambda$ 值增大,聚类数目减少了,这是由于其中的一些初始聚类中心移动到了相同的位置上。随着 $\lambda$ 增大到某个水平值的时候,聚类数目就和实际类的数目相同了,这就是找到了正确的 $\lambda$ 值。然而,随着 $\lambda$ 进一步增加,聚类数目变得比实际的类数目少了。最终,当 $\lambda$ 增大到某一个值时,聚类数目变成了 1。

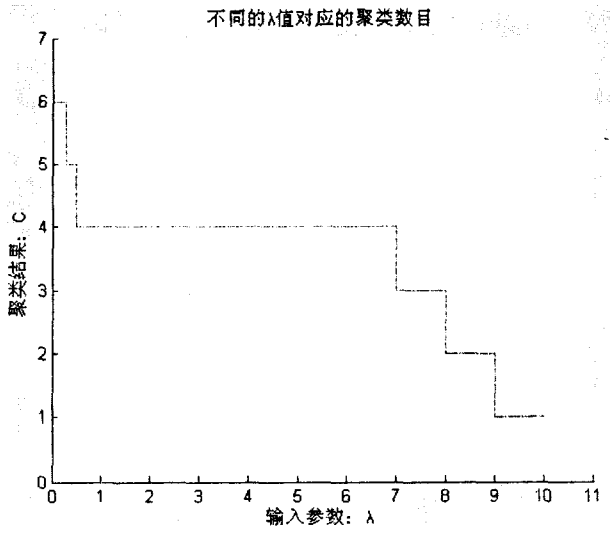


图 1　不同的 $\lambda$ 值对应的聚类数目图

例如,当 $\lambda=2$ 时,在经过 12 次迭代之后,程序停止了,6 个初始聚类中心移动到了 4 个位置,和 4 个实际的聚类中心非常接近。表 1 记录了识别出的聚类中心和实际的聚类中心位置的对比。

表 1　识别出的聚类中心和实际聚类中心位置对比表

| | 类Ⅰ | 类Ⅱ | 类Ⅲ | 类Ⅳ |
|---|---|---|---|---|
| 实际聚类中心 | (1,5) | (1,1) | (2,6) | (5,5) |
| 终止聚类中心 | (0.9912,4.8803) | (1.0009,0.9843) | (2.0374,5.9948) | (5.0136,4.9875) |

## 3　结束语

经典的 C 均值聚类算法以及模糊 C 均值聚类算法存在两个方面的问题:一是算法对初始聚类中心的过分依赖,二是需要预选知道实际的聚类数目。这就大大局限了算法的应用领域和效果。基于此,提出了一种基于模糊 C 均值聚类算法的改进算法,在标准的模糊 C 均值聚类算法的基础上,给目标函数加入了一个惩罚项,使得上述问题得以解决。仿真实验证实了算法的可行性,以及比经典算法具有更加广泛的实用性。

**参考文献:**

[1] Li Mark Junjie, Ng Michael K, Cheung Yiu - ming, et al. Agglomerative Fuzzy K - Means Clustering Algorithm with Selection of Number of Clusters[J]. IEEE Trans. Knowledge and Data Engineering, 2008, 20:1519 - 1534.

[2] 罗军生,李永忠. 基于模糊 C - 均值聚类算法的入侵检测[J]. 计算机技术与发展,2008,18(1):178 - 180.

[3] 王　伟,高　亮. 一种基于模糊聚类的离散化方法[J]. 计算机技术与发展,2008,18(3):53 - 55.

[4] 吴　瑛,王秋生. 模糊 C 均值聚类算法在 web 使用挖掘上的应用研究[J]. 计算机技术与发展,2008,18(6):32 - 35.

[5] Ruspini E R. A New Approach to Clustering[J]. Information Control, 1969, 19:22 - 32.

[6] Bezdek J C. A Convergence Theorem for the Fuzzy ISODATA Clustering Algorithms[J]. IEEE Trans. Pattern Analysis and Machine Intelligence, 1980, 2:1 - 8.

[7] 彭秋生,魏文红. 基于核方法的并行模糊聚类算法[J]. 计算机工程与设计,2008,29(8):1881 - 1883.

[8] 孙晓霞,刘晓霞. 模糊 C 均值( FCM)聚类算法的实现[J]. 计算机应用与软件, 2008,25(3):48 - 50.

[9] Miyamoto S, Mukaidono M. Fuzzy c - means as a Regularization and Maximum Entropy Approach[C]//Proc. Seventh Int'l Fuzzy Systems Assoc. World Congress (IFSA '97). [s. l.][s. n.], 1997:86 - 92.

[10] Han Jiawei, Kamber M. Data Mining: Concepts and Techniques[M]. Los Altos, CA: Morgan Kaufmann Publishers, 2001.