文章编号:1007-6735(2012)04-0351-04

一种改进的模糊C-均值聚类算法

曹 易, 张 宁
(上海理工大学 管理学院, 上海 200093)

摘要: 由于现有模糊C-均值聚类算法固有的局限性,本文提出了一种改进的模糊C-均值聚类算法.首先用概率密度函数来确定初始聚类中心点和聚类数,其次用竞争学习思想提出使对手增加抑制因子来修改隶属度得到加快收敛速度的效果,最后提出用一个类内差异与类间差异兼备的新的有效性指标来作为迭代条件的目标函数.通过实验获取参数的最优取值范围,通过与经典模糊C-均值聚类算法的比较,证明了该改进算法不仅加快了收敛速度,而且在聚类结果的质量上有一定程度的提高.

关键词: 模糊C-均值聚类; 概率密度; 隶属度; 有效性指标
中图分类号: TP 301.6; N 94 **文献标志码**: A

Improved Fuzzy C-means Clustering Algorithm

CAO Yi, ZHANG Ning
(*Business School, University of Shanghai for Science and Technology, Shanghai 200093, China*)

Abstract: The fuzzy C-means algorithm was improved to break through the existing performance limitations. The function of probability consistency was used to determine the original clustering center and the clustering number. For each object an inhibitory factor was added to its opponent to accelerate the convergence. A new validity index which takes a balance between intra-clustering and inter-clustering variation was proposed to act as the aim function. Experiments show that the improved algorithm behaves comparatively higher performance in convergence speed and clustering quality.

Key words: *fuzzy C-means clustering*; *probability consistency*; *subjection value*; *validity index*

聚类是根据对象之间的相似性来将他们聚集成不同类别的方法.评价一个聚类质量的好坏,总体是该聚类结果中同一类内部的对象尽可能相似,不同类之间的对象尽可能相异.到目前为止,数据挖掘中常用的聚类算法有层次聚类、划分聚类、基于网格聚类、基于密度聚类及模糊聚类等[1].

传统的聚类是一种硬性划分,具有"非此即彼"性,但是,现实生活中很多事物是"亦此亦彼",很难将它们严格地划分到一个具体的类中.模糊C-均值聚类算法(FCM)是应用最广泛的聚类算法之一[2],它具有算法简单、收敛速度快、能处理大规模数据等优点,因此,该算法已经有效地应用在数据挖掘、模式识别及决策支持等领域,具有很大的理论以及实践价值.但是,FCM算法同时也存在着很大的

收稿日期: 2011-06-20
基金项目: 国家自然科学基金资助项目(70971089);上海市重点学科建设资助项目(S30501)
作者简介: 曹 易(1987-),男,硕士研究生.研究方向:复杂网络.E-mail:caoyi8689494@126.com
张 宁(联系人),女,教授.研究方向:系统分析与集成、复杂网络.E-mail:zhangning@usst.edu.cn

局限性[3]：聚类数与聚类初始中心的选择极大地影响着聚类效果，并且该算法采用梯度法求解极值，所求解往往是局部最优. 为此，文献[4]用信息熵来计算最佳聚类数目，Yager 和 Filev[5] 提出了一种称为爬山法的初始聚类中心方法.

由于一般模糊 C-均值算法的上述缺点，本文提出了一种改进的 FCM 算法. 首先用概率密度的思想得到最佳聚类数和初始聚类中心，其次通过对拥有次大隶属度的中心点加入一个抑制因子来加速算法收敛，最后用一个兼顾类内距与类间距的新的目标函数来替代原有的目标函数. 经实验证实，该算法在聚类结果质量与算法速度上都有了一定程度的改进.

1 普通模糊 C-均值的算法

设 $\boldsymbol{X}=\{x_1,x_2,\cdots,x_n\}$ 是待聚类的对象的全体(论域)，$\boldsymbol{X}$ 中每个对象(样本) $x_k(k=1,2,\cdots,n)$ 可以用有限个参数值来描述，每个参数刻画 x_{kj} 的某个特征. 所以，对象 x_k 就可以用一个向量 $\boldsymbol{P}(x_k)=(x_{k1},x_{k2},\cdots,x_{ks})$ 来表示，$\boldsymbol{P}(x_k)$ 为 x_k 的特征向量[6]. u_{ik} 表示 $\boldsymbol{X}$ 中第 k 个对象对第 i 类的隶属度函数，$v_i(i=1,2,\cdots,c)$ 表示聚类中心，则第 k 个对象到第 i 个聚类中心 v_i 的欧式距离为

$$d_{ik}^2=\|x_k-v_i\|^2=\sum_{j=1}^{s}(x_{kj}-v_{ij})^2 \quad (1)$$

目标函数定义为

$$\min J_m(U,V,c)=\sum_{i=1}^{c}\sum_{k=1}^{n}u_{ik}^m d_{ik}^2=\sum_{i=1}^{c}\sum_{k=1}^{n}u_{ik}^m\|x_k-v_i\|^2,\quad 1.5\leqslant m\leqslant 2.5 \quad (2)$$

其中，$\sum_{i=1}^{c}u_{ik}=1, 1\leqslant k\leqslant n; 0\leqslant u_{ik}\leqslant 1, 1\leqslant i\leqslant c; 1\leqslant k\leqslant n$；$m$ 为模糊因子[7-8].

随着隶属函数 u_{ik} 和中心点 v_i 不断更新，若目标函数 $J_m(U,V,c)$ 达到了满意的稳定程度，就终止迭代算法.

2 改进的模糊 C-均值的算法及其实现

2.1 聚类数和初始聚类中心选取的改进

上述算法具有简单、收敛速度快、能处理大规模数据等优点，但是，聚类数和聚类初始中心的选择极大地影响着聚类效果，并且该算法采用梯度法求解极值，所求解往往是局部最优.

目前，在 FCM 算法中，聚类数和初始中心点的选择对算法的复杂度以及聚类效果的影响相当大，因此，选择一个适合的中心点是至关重要的. 本文利用一种概率密度函数来选择聚类数和初始中心[9,10]. 定义对象 x_i 处的密度函数为

$$D_i^{(0)}=\sum_{j=1}^{n}\frac{1}{1+f_d\|x_i-x_j\|^2},\quad f_d=4/r_d^2 \quad (3)$$

其中，r_d 为邻域半径，其数值与数据的分布特性有关.

本文取 r_d 为 n 个对象的平均距离，即

$$r_d=\sqrt{\frac{\sum_{j=1}^{n}\sum_{i=1}^{n}\|x_i-x_j\|^2}{n(n-1)}}$$

显然，x_i 周围分布越密集，r_d 值越小，密度函数 $D_i^{(0)}$ 值越大. 令 $D_1^*=\max\{D_i^{(0)}, i=1,2,\cdots,n\}$，其满足条件的点取为第一个初始聚类中心，设为 x_1^*. 第 k 个聚类中心点为

$$D_k^*=\max\{D_i^{(k-1)}, i=1,2,\cdots,n\} \quad (4)$$

第 k 次迭代时的聚类中心的密度函数为

$$D_i^{(k)}=D_i^{(k-1)}-D_k^*\frac{1}{1+f_d\|_i-x_k^*\|^2},\quad k=1,2,\cdots,n \quad (5)$$

从式(5)可以看出，距离 x_k^* 越近，对应的 D 值越小，则其成为下一个聚类中心的可能性就越小. 给定一个参数 $\delta\in[0,1]$，根据式(3)～(5)，当 $D_c^*<\delta D_1^*$，停止计算，此时的 c 就是所求的聚类数，x_1^*，$x_2^*,\cdots,x_c^*$ 就是所求的初始聚类中心点.

2.2 隶属度的改进

由 FCM 算法可知，聚类实际上就是一个隶属矩阵 $\boldsymbol{u}$ 和聚类中心 $\boldsymbol{v}$ 交替优化过程. 可以修正隶属矩阵 $\boldsymbol{u}$ 来计算下一次迭代的聚类中心 $\boldsymbol{v}$，使计算结果更合理，提高算法的收敛速度. 隶属度越大，样本点对类中心的吸引力就越大，类中心的下一次迭代值受隶属度的影响就越大[11]. 本文根据竞争学习算法，给出了一种修正隶属矩阵 $\boldsymbol{u}$ 的算法. 本文称距离样本点最近的类中心为赢者，距离次近的为赢者对手，通过减弱对手的吸引力来加快赢者的收敛速度. 加入一个抑制因子 $\alpha\in[0,1]$，抑制次近样本点的吸引力，来加快算法收敛速度. 具体描述为：对于对象 x_j，假如它对第 t 类的隶属度最大，为 u_{tj}；对第 s 类的隶属度次大，为 u_{sj}. 给定抑制因子 α，根据式(6)修改隶属度为

$$u'_{tj}=u_{tj}+(1-\alpha)u_{sj},\quad u'_{sj}=\alpha u_{sj} \quad (6)$$

其余对象的隶属度不变.

2.3 目标函数选取的改进

聚类结果应该是类内尽可能紧凑，类间尽可能疏远. 但是，传统的 FCM 算法的目标函数只考虑了

类内距离,没有重视类间距离. 本文根据 Xie-Beni 提出的聚类有效性指标[12],给出一种兼顾类内和类间距离的有效性指标,将它作为新的目标函数.

类内差异 $W(\boldsymbol{u},\boldsymbol{v},c)$和类间差异 $B(\boldsymbol{u},\boldsymbol{v},c)$分别为

$$W(\boldsymbol{u},\boldsymbol{v},c) = \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}^{m} d_{ik}^{2} = \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}^{m} \| x_k - v_i \|^2, \quad 1.5 \leqslant m \leqslant 2.5 \tag{7}$$

$$B(\boldsymbol{u},\boldsymbol{v},c) = n \min_{j\neq k} \| v_j - v_k \|^2 \tag{8}$$

将 $W(\boldsymbol{u},\boldsymbol{v},c)$和 $B(\boldsymbol{u},\boldsymbol{v},c)$的商作为新的目标函数 $J_m(\boldsymbol{u},\boldsymbol{v},c)$,即

$$J_m(\boldsymbol{u},\boldsymbol{v},c)=\frac{W(\boldsymbol{u},\boldsymbol{v},c)}{B(\boldsymbol{u},\boldsymbol{v},c)}=\frac{\sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}^{m} \| x_k - v_i \|^2}{n \min_{j\neq k} \| v_j - v_k \|^2} \tag{9}$$

2.4　模糊 C-均值算法改进的具体实现

综上所述,现给出该算法的具体步骤.

Step 1　给定待聚类对象集 X,参数 δ,模糊因子 m,抑制因子 α,迭代参数 ε.

Step 2　根据式(3)～(5)求出初始聚类数 c 和聚类中心 $\boldsymbol{v}$.

Step 3　计算隶属矩阵 u_{ik},再根据式(6)修改 $\boldsymbol{u}$.

$$u_{ik} = \frac{1}{\sum_{j=1}^{c}\left(\frac{d_{ik}}{d_{jk}}\right)^{\frac{2}{m-1}}} \tag{10}$$

Step 4　更新聚类中心 v_i.

$$v_i = \frac{\sum_{k=1}^{n} u_{ik}^{m} x_k}{\sum_{k=1}^{n} u_{ik}}, \quad i = 1,2,\cdots,c \tag{11}$$

Step 5　根据式(9)计算 $J_m(\boldsymbol{u},\boldsymbol{v},c)$,若式(12)成立,终止计算;否则,$l = l + 1$,转向 Step 3.

$$\frac{| J_m(\boldsymbol{u},\boldsymbol{v},c)^{(l+1)} - J_m(\boldsymbol{u},\boldsymbol{v},c)^{(l)} |}{J_m(\boldsymbol{u},\boldsymbol{v},c)^{(l+1)}} \leqslant \varepsilon \tag{12}$$

3　实验结果及分析

通过实验来测试改进算法的效率和聚类质量,并与普通的模糊 C-均值算法进行比较. 本次实验平台操作系统为 Windows XP,CPU 为双核 E7500 2.9 GHz,内存 2 GB. 数据采用某高校的 Web 访问日志,共有 2 993 个 IP 用户,访问的网页被综合成了教育、娱乐、搜索等 35 个类别,每个类别认为是用户的一个属性值,大小取该用户对该类别的访问频率,得到了 2 993×35 的用户类别矩阵. 实验取模糊因子 m 值为 2,最大可能迭代次数为 200,通过改变参数 δ,α 和 ε 的值来测试算法的性能,得到参数的最佳取值范围. 聚类结果的有效性指标 p[13] 用式(13)来评价,值越小,则聚类效果越好;反之亦然. N 为算法迭代次数.

$$p = \frac{\frac{1}{n}\sum_{i=1}^{c}\sum_{k=1}^{n} u_{ij}^{m} \| v_j - x_i \|^2 + \frac{1}{c}\sum_{i=1}^{c} \| v_i - \bar{v} \|^2}{\min_{i\neq j} \| v_i - v_j \|^2} \tag{13}$$

经调整实验控制参数得出结果如图 1～4 所示.

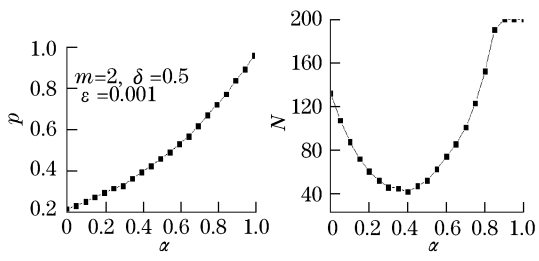

图 1　聚类有效性 p 和迭代次数 N 与 α 的关系

Fig. 1　Relationship of clustering validity p, iteration number N and α

从图 1～4 可以看出:

a. 当 $m = 2, \delta = 0.5, \varepsilon = 0.001$ 时,随着参数 α 从 0 变化到 1 时,有效性指标 p 与迭代次数 N 的变化趋势如图 1,综合考虑该算法的聚类质量以及迭代次数,取 $\alpha = 0.3$ 较为合理.

b. 在图 2 中,当 $m = 2, \alpha = 0.3, \varepsilon = 0.001$ 时,随着参数 δ 从 0 变化到 1 时,有效性指标 p,迭代次数 N 以及聚类数 c 变化趋势如图 2(见下页),同样综合考虑该算法,取 $\delta = 0.5$ 较为合理,此时聚类数 $c = 43$.

c. 当 $m = 2, \alpha = 0.3, \delta = 0.5$ 时,随着参数 ε 从 0.000 5～0.001 4 之间变化时,有效性指标 p 与迭代次数 N 的变化趋势如图 3(见下页),同样综合考虑该算法,取 $\varepsilon = 0.001$ 较为合理.

d. 取 $m = 2, \alpha = 0.3, \delta = 0.5, \varepsilon = 0.001$时,用本文的改进 FCM 算法与经典的 FCM 算法进行比较,从图 4(见下页)中可以看出,当聚类数目相同时,与经典 FCM 算法相比,本文算法在有效性指标 p 与迭代次数 N 上均有一定程度的提高.

综上所述,本文提出的改进 FCM 算法中,通过调节参数 α,δ,ε 的大小,其中本文的数据中 $\alpha = 0.3$、$\delta = 0.5$、$\varepsilon = 0.001$,较原有的 FCM 算法在聚类质量和算法速度有一定程度的提高.

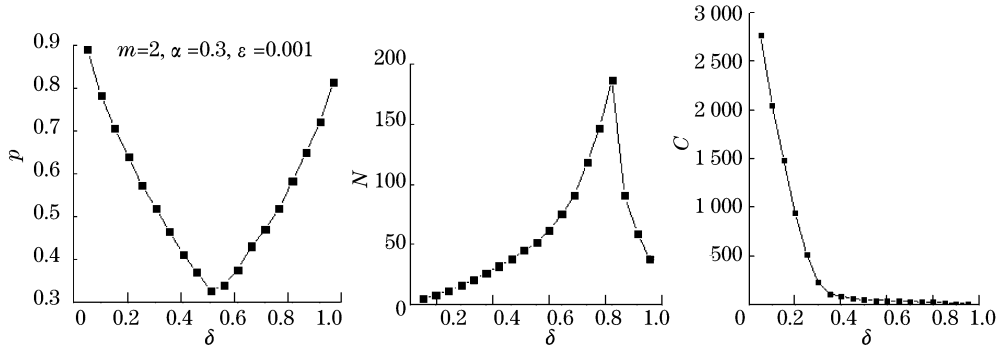

图2 聚类有效性 p,迭代次数 N 及聚类数 c 与 δ 的关系

Fig.2 Relationship of clustering validity p, iteration number N, cluster number c and α

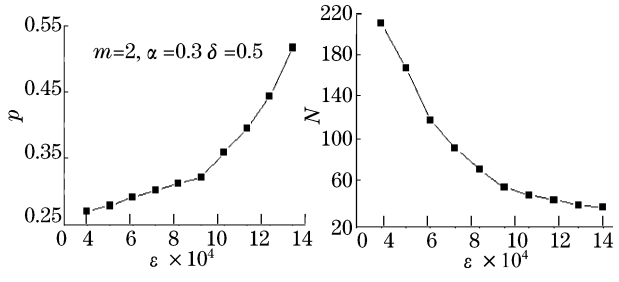

图3 聚类有效性 p 和迭代次数 N 与 ε 的关系

Fig.3 Relationship of clustering validity p, iteration number N and ε

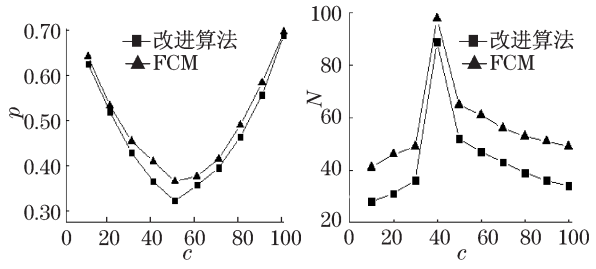

图4 改进的FCM算法与经典FCM算法比较

Fig.4 Comparison of the improved FCM and classical FCM algorithm

4 结 论

通过分析经典的FCM算法中的局限性,例如聚类结果对聚类数和初始聚类中心的敏感性,以及目标函数选取只考虑类内部距离而忽略了类间距离,提出了一种改进的FCM算法.经实验证明,与经典算法相比,改进算法不论是在聚类质量上还是在算法复杂度上,都有一定程度的提高.用概率密度函数找到最佳的聚类数以及初始聚类中心点;利用竞争学习算法中的抑制对手来修改隶属矩阵,从而达到加快算法的收敛速度;用一个类内距离与类间距离兼顾的新目标函数替换原有目标函数.实验证明,本文算法在参数设置合理的情况下,聚类质量和算法速度在原有FCM算法上有一定程度的提高.

参考文献:

[1] Mitra S, Pal S K, Mitra P. Data mining in soft computing framework: a survey[J]. IEEE Transactions on Neural Networks, 2002, 13(1): 3-14.

[2] 贺玲,吴玲达,蔡益朝. 数据挖掘中的聚类算法综述. 计算机应用研究,2007,1(1):16-19.

[3] 齐淼,张化祥. 改进的模糊C-均值聚类算法研究. 计算机工程与应用,2009,45(20):133-135.

[4] 沈红斌,杨杰,王士同,等. 基于信息理论的合作聚类算法研究[J]. 计算机学报,2005,28(8):1287-1294.

[5] Yager R R, Filev D P. Approximate clustering via the mountain method[J]. IEEE Transactions on SMC, 1994, 24(8): 1279-1284.

[6] 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.

[7] 朱文婕,吴楠,胡学钢. 一个改进的模糊聚类有效性指标[J]. 计算机工程与应用 2011,47(5):206-209.

[8] 高新波,裴继红,谢维信. 模糊C-均值聚类算法中的加权指数m的研究[J]. 电子学报,2000,28(4):80-83.

[9] 饶泓,扶名福,谢明详. 基于模糊聚类的神经网络故障诊断方法[J]. 微计算机信息,2007,1(1):196-197.

[10] 李春生,王耀南. 聚类中心初始化的新方法[J]. 控制理论与应用,2010,27(10):1435-1440.

[11] 张曙红,孙建勋,诸克军. 基于遗传优化的采样模糊C-均值聚类算法[J]. 系统工程理论与实践,2004,5(1):121-125.

[12] Xie X L, Beni G. A validity measure for fuzzy clustering[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1991, 13(8): 841-847.

[13] Kwon S H. Cluster Validity index of fuzzy clustering [J]. Electronics Letters, 1998, 34(22): 2176-2177.