DOI:10.3969/j.issn.1673-4785.201503033
网络出版地址:http://www.cnki.net/kcms/detail/23.1538.tp.20150709.1649.001.html

基于改进的遗传算法的模糊聚类算法

张永库[1],尹灵雪[2],孙劲光[1]
(1.辽宁工程技术大学 电子与信息工程学院,辽宁 葫芦岛 125105;2. 辽宁工程技术大学 研究生学院,辽宁 葫芦岛 125105)

摘 要:针对传统的模糊 C 均值聚类(fuzzy C-means clustering)算法容易陷入局部最优解,并且对初始值敏感的缺陷,提出一种基于改进的遗传算法的模糊聚类算法。该算法针对遗传算法的早熟问题提出一种改进的遗传算法,并将其应用于 FCM 算法,来寻找全局最优的聚类中心。实验表明,该算法与基于传统遗传算法的 FCM 算法相比,具有更强的寻优能力,更优的聚类效果。

关键词:模糊 C 均值算法;聚类分析;遗传算法;动态分析;模糊聚类;初始值;避免早熟;全局最优;局部最优
中图分类号: TP18 **文献标志码**:A **文章编号**:1673-4785(2015)04-0627-09

中文引用格式:张永库,尹灵雪,孙劲光. 基于改进的遗传算法的模糊聚类算法[J]. 智能系统学报, 2015, 10(4): 627-635.
英文引用格式:ZHANG Yongku, YIN Lingxue, SUN Jinguang. Fuzzy clustering algorithm based on improved genetic algorithm[J]. CAAI Transactions on Intelligent Systems, 2015, 10(4): 627-635.

Fuzzy clustering algorithm based on the improved genetic algorithm

ZHANG Yongku[1], YIN Lingxue[2], SUN Jinguang[1]
(1. College of Electronics and Information Engineering, Liaoning Technical University, Liaoning 125105, China; 2. Institute of Graduate, Liaoning Technical University, Liaoning 125105, China)

Abstract: The traditional fuzzy C-means(FCM) clustering algorithm is prone to fall into the solution of local optimum and is sensitive to initial value. Aiming at these drawbacks, a fuzzy C-means based on the improved genetic algorithm is presented. The improved genetic algorithm is employed to optimise the FCM algorithm, finding the cluster center of the global optimum. Finally, the experimental results show that compared with the traditional FCM, the proposed algorithm has stronger optimisation ability and better clustering effect

Keywords: fuzzy C-means clustering; cluster analysis; genetic algorithm; dynamic analysis; fuzzy clustering; initial values; premature contraction avoidance; global optimum; local optimum

作为一种无监督的学习方法,聚类分析被视为机器学习研究以及数据挖掘应用中的一个主要内容。它仅根据在数据中发现的描述对象及其关系的信息,将数据对象分组,其目标是,组内的对象相互之间是相似的,而不同组中的对象是不同的[1]。由于现实生活中许多问题是在类属方面存在模糊性的,那么对其进行划分界限明确的聚类分析显然是不恰当的。L.A.Zedeh 提出了一种用模糊集理论来解决的聚类问题,即模糊聚类分析。在大量的模糊聚类算法中,应用的最为广泛的算法便是基于目标函数的模糊 C 均值算法(FCM)。FCM 作为聚类分析中的一个重要研究领域,目前针对其研究的应用已经非常广泛。但该算法易陷入局部最优解和对初始值敏感[2]。文献[3]提出将遗传算法(genetic algorithm)应用于 FCM,利用 GA 的全局搜索性能去确定最佳聚类数并寻找到全局最优聚类中心,即 GA-

收稿日期:2015-03-18. **网络出版日期**:2015-07-02.
基金项目:国家自然科学基金资助项目(61172144);国家科技支撑计划资助项目(2013BAH12F02);辽宁省教育厅科学研究一般资助项目(L201432).
通信作者:尹灵雪. E-mail:ylx19910708@163.com.

FCM 算法[3]。近些年来,GA-FCM 算法得到了十分广泛的应用[4-6],众多学者针对 GA-FCM 提出了改进的算法,LIU Su-hua[7] 等将模拟退火算法引入了 GA 中,以此来改进 GA 的早熟问题,进而获得较好的聚类效果,与此同时改进 GA 的交叉算子和变异算子。Feryel Souami 等提出的文献[8]使用了一种新的适应度函数,从而避免了存储和计算隶属度矩阵在效率上的浪费[8]。

这些算法用不同的方法来对 GA-FCM 进行了改进并且取得了良好的效果。为了进一步优化 GA-FCM 所达到的聚类效果,改进 GA 所存在的早熟问题,本文通过对 GA 的选择算子进行改进,并非每次循环都进行选择,而是通过目前种群中个体的多样性动态决定选择操作是否发生,进而提出了一个新的 Dynamic GA-FCM 算法,改进了 GA 的早熟问题,通过人工数据以及经典数据集的仿真实验,表明本文所提出的改进方法确实具有更佳的聚类性能。

1 FCM 算法

模糊 C 均值聚类(FCM)最早是由 JAMES C. BEZDEK 等人在参考文献[9]中提出来的,FCM 算法基于模糊划分,用隶属度来确定每一个数据点属于某一个聚类的程度,最终实现被划分到同一簇的对象之间相似度最大,而不同簇之间的相似度最小[1]。

它把向量 $\boldsymbol{x}_i(i = 1,2,\cdots,n)$ 分为 c 组 $\boldsymbol{V} = \{\boldsymbol{V}_1, \boldsymbol{V}_2,\cdots,\boldsymbol{V}_c\}$,并求出每一组的聚类中心 $\boldsymbol{A} = \{\boldsymbol{A}_1,\boldsymbol{A}_2,\cdots,\boldsymbol{A}_c\}$,每一个向量 $\boldsymbol{x}_i$ 用值在 0~1 的隶属度来表示其属于各个组的程度。一个数据集的隶属度之和恒等于 1。即隶属度矩阵 $\boldsymbol{U}$ 必须满足如下条件:

$$\sum_{i=1}^{c} u_{ij} = 1, \forall j = 1,2,\cdots,n \tag{1}$$

当 $u_{ij} = 1$ 时表示第 j 个对象完全属于第 i 个类,$u_{ij} = 0$ 时表示第 j 个对象完全不属于第 i 个类。

FCM 的目标函数为

$$J(\boldsymbol{U},\boldsymbol{A}_1,\boldsymbol{A}_2,\cdots,\boldsymbol{A}_c) = \sum_{i=1}^{c} J_i = \sum_{i=1}^{c}\sum_{j}^{n} u_{ij}^{\lambda} d_{ij}^2 \tag{2}$$

$d_{ij} = \| \boldsymbol{A}_i - \boldsymbol{x}_j \|$ 为第 i 个聚类中心和第 j 个数据点之间的欧几里德距离;其中 $\lambda \in [1, \infty)$,它是一个加权指数。在式(1)约束条件下对式(2)应用拉格朗日乘法,求导得:

$$\boldsymbol{A}_i = \frac{\sum_{j=1}^{n} u_{ij}^{\lambda} x_j}{\sum_{j=1}^{n} u_{ij}^{\lambda}} \tag{3}$$

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} \left[\frac{d_{ij}}{d_{kj}}\right]^{2/(\lambda-1)}} \tag{4}$$

为了使目标函数 $\boldsymbol{J}$ 在已经给定的约束条件下取得极小值,FCM 算法基本步骤如下:

1)随机生成 c 个聚类中心 $\{\boldsymbol{A}_1,\boldsymbol{A}_2,\cdots,\boldsymbol{A}_c\}$。

2)根据式(4),计算求得隶属矩阵 $\boldsymbol{U}$,使其符合式(1)中的约束条件。

3)根据式(2)计算目标函数值。若它小于某个确定的阈值,或者它相对上一次目标函数值的改变量小于某一个给定阈值,则算法终止。

4)用式(3)计算得到新的聚类中心。返回 2)。

2 遗传算法

John H. Holland 在 20 世纪 60 年代提出遗传算法,该算法是一种基于遗传学中的自然选择和适者生存机制而产生的简单、强健、有效的优化技术[10]。GA 中所求问题的可行解使用个体即染色体来表示,并且每一个体以 STRING 类型进行编码,该算法以遗传学为基础,每一个体都按照提前设计好的适应度函数来计算该个体的适应度值,然后按照优胜劣汰原则并且通过全局并行搜索来不断地获得更加优秀的个体,进而获得更加优秀的种群[11]。编码的方法、对种群的初始化、设计适应度函数、遗传算子以及各个参数的设置,以上 5 部分是遗传算法的主要部分[12]。

而上述提到的遗传算子,其中主要包括选择算子、交叉算子、和变异算子。以下为基本遗传算法的步骤:

1)随机产生染色体种群,每一个染色体都代表一个所求解问题的解决方案。

2) 计算种群中每一个染色体的适应度值。

3)重复以下步骤,直到完成一个新的种群的产生。

①根据每个个体的适应度值从种群中选择 2 个父代染色体。适应度值越高,被选中作为父代染色体的概率越大。

②给定一个交叉概率 P_c,若 2 个父代染色体交叉则形成新的子代染色体。若未发生交叉,则其子代染色体与父代染色体相同。

4) 用新产生的种群代替原来的种群。

5)给定一个变异概率 P_m,种群中每个个体以该概率发生变异,产生新的子代。

6)用新产生的种群代替原来的种群。

7)如果新种群满足了终止条件,则跳出循环,并且返回种群中的最佳解决方案,得到所求问题的最优解。

8)回到 2)。

3　Dynamic GA-FCM

GA-FCM 算法虽然在一定程度上改善了模糊 C 均值算法对初始聚类中心敏感的问题,但遗传算法的早熟问题却严重影响了聚类算法的准确性[13]。在传统的 GA-FCM 算法中,由于早熟问题而产生的局部最优聚类中心,可能会降低该算法聚类的质量[14]。

基于以上问题,本文将传统遗传算法进行改进,目的是降低其发生早熟现象的概率,从而提高聚类的质量。早熟问题发生的最主要原因就是选择的速度过快,而产生新个体的速度过慢[15]。在遗传算法中,新个体的产生是通过交叉算子和变异算子来实现,因此可以通过增加交叉和变异算子实现的次数来加快产生新个体的速度,但同时,若种群中个体的多样性过高,即产生新个体的速度过快,又会导致种群中个体所保存的解决方案的信息将会被丢失,并且很难实现最终的收敛,或收敛的速度过慢。为了能够平衡种群中个体的多样性和稳定性,本文将遗传算法中的选择算子进行改进,使其根据当前种群中个体的多样性,来自适应的调整选择的速度。进而保证种群中个体的多样性保持在适当的水平,不会过高,也不会过低。由于传统的 GA-FCM 每次计算适应度函数值时都需要计算隶属度矩阵,这大大增加了实现算法所用的时间,在本算法中,采用了 Feryel Souami 等提出的新的适应度函数公式[8],计算适应度时不再计算隶属度矩阵,缩短了算法所用时间[20]。

3.1　染色体编码和初始种群产生

染色体编码有很多种方式,本文采用了浮点数编码。初始种群的产生采用了随机生成,方法为:在所给出的参与聚类分析的 n 个样本点中,随机的抽取 c 个样本点,将其作为 c 个聚类的聚类中心,并通过染色体表示出来,其中 c 为聚类数。即 1 条染色体可以使用由 c 个基因位组成的浮点码串 $\{\boldsymbol{A}_1,\boldsymbol{A}_2,\boldsymbol{A}_3,\cdots,\boldsymbol{A}_c\}$ 表示,重复进行 m 次(m 为种群大小),得到初始种群。如图 1 所示,为一个由 4 个二维聚类中心组成的染色体,其中 A_{i1}、A_{i2}($\forall i = 1,2,\cdots,c$)皆为浮点数。

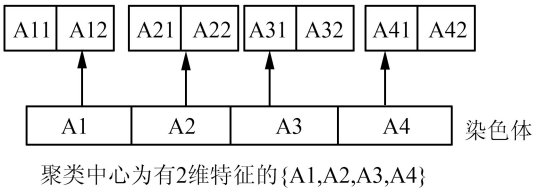


图 1　染色体的表示方式

Fig.1　Representation of a chromosome

3.2　适应度函数

对于所求解问题,解决方案的好坏通过每一个个体的适应度值来评价。而适应度值则是根据事先定义好的适应度函数计算得到的。

传统 GA-FCM 算法的适应度函数需要使用隶属度矩阵,而每次循环中,隶属度矩阵的更新花费大量的时间,在很大程度上降低了算法的效率,为了解决这一问题,Hathaway 和 Bezdek 基于式(2)提出了一个新的 FCM 目标函数的形式,Feryel Souami 就是基于该目标函数以及 Frigui 所提出的竞争凝聚(competitive agglomeration)算法,提出了一个新的适应度函数[19],该适应度函数定义如下:

$$F = \frac{1}{R * f^2(t)} \tag{5}$$

$$f(t) = 1 - a * \exp(-t/b) \tag{6}$$

式中:t 为种群解决方案的代数,a、b 为 2 个常量。

$$R = \sum_{j=1}^{n} \left(\sum_{i=1}^{c} (d(x_j, A_i))^{1/(1-l)} \right)^{1-l} \tag{7}$$

其中,d 可以是欧几里得距离,也可以是其他距离。在本文中,使用了欧几里得距离,l 为模糊指数。

3.3　选择操作

选择操作就是从种群中选择适应度值较高的个体,将其保留到下一代种群,适应度值越高的个体,被保留到下一代的概率就越高,反之则越小,选择操作使种群中个体的适应度值不断的接近于最优解。本文使用轮盘赌选择法。该方法是一种回放式随机采样法,种群中每一个个体被选择的概率为

$$p(\mathrm{ind}_i) = F(\mathrm{ind}_i) / \sum_{k=1}^{m} F(\mathrm{ind}_k) \tag{8}$$

式中:$F(\mathrm{ind}_i)$ 为第 i 个个体的适应度值,$\sum_{k=1}^{m} F(\mathrm{ind}_k)$ 为种群中所有个体,即 m 个个体的适应度值之和。

3.4　交叉操作

本文算法采用了离散重组的一点交叉,假设有染色体对 $\boldsymbol{x}_a(t) = (\boldsymbol{x}_{a1},\boldsymbol{x}_{a2},\cdots,\boldsymbol{x}_{ai},\cdots,\boldsymbol{x}_{ac})$,$\boldsymbol{x}_b(t) = (\boldsymbol{x}_{b1},\boldsymbol{x}_{b2},\cdots,\boldsymbol{x}_{bi},\cdots,\boldsymbol{x}_{bc})$。其中 $i \in (1,2,\cdots,c)$,t 为遗传算法迭代次数。每个染色体的任意 2 个相邻基因位之间设立一个交叉点,从左到右依次为 1,2,…,$c-1$,总共 $c-1$ 个不同的交叉点。每一对染色体以 P_c 概率来进行交叉,在 $c-1$ 个交叉点中随机的选择一点,交换 2 个染色体自该交叉点以后的所有基因。

3.5　变异操作

对每一个个体的每一个基因位,产生随机数 p,当 $p < P_m$ 时,对此基因位进行随机变异操作,在参与

聚类分析的样本集中随机的抽取一个对象,代替此基因位的基因,作为新聚类中心,生成下一代种群。

变异概率 P_m 一般很小,通常在 0.001~0.1,如果变异概率过大,就会破坏很多优良个体,可能无法得到最优解,而如果变异概率过小,个体则无法变异到更优的解,导致算法收敛速度变慢,只能达到局部最优解。

3.6 保持种群多样性

在聚类分析中,传统的遗传算法当种群中各个体的目标函数值趋向一致或者趋向局部最优时,对于交叉算子以及变异算子来说,不易生成新的最优个体的结构,因而搜索范围将局限于局部最优的区域,只能取得局部最优的结果,即早熟。针对这一问题,在本算法中,当种群中的各个体其目标函数值趋向一致时,根据种群的多样性度量值,动态决定选择操作是否发生,而非每次循环都进行选择,只有当种群的多样性度量值大于所期望的最低多样性度量值时,才进行选择操作。

本算法采用标准偏差法(standard deviation)来度量种群的多样性,标准偏差法是 Kenny Q. Zhu 在文献[16]中所介绍的。根据该方法,种群的多样性度量值为

$$\mathrm{stddev}(P)=\sqrt{\frac{\sum_{i=1}^{m}(f_i-\bar{f})^2}{m-1}} \tag{9}$$

式中:m 为种群大小,f_i 为第 i 个个体的适应度值,$\bar{f}$ 为种群中所有个体适应度值的平均值。

而最低多样性度量值,本算法根据 Kim Nguyen 等在文献[17]所提出的模型,将其定义为

$$\mathrm{Std}_{\min}(b)=\mathrm{Std}_{\min}(0)\times\exp(-b) \tag{10}$$

式中:$\mathrm{Std}_{\min}(0)$ 为种群初始多样性度量值,b 为种群当前的迭代次数。

3.7 Dynamic GA-FCM 算法描述

Dynamic GA-FCM 算法的思想是先随机选择 m 个个体,组成大小为 m 的种群,然后计算种群的多样性度量值,若种群的多样性小于给定阈值,则不进行选择操作,对种群中的个体不断的循环进行交叉操作和变异操作,增加种群中个体的多样性,直至种群多样性大于给定阈值。当种群多样性度量值大于阈值时,则按照原遗传算法的流程,顺序进行选择、交叉、以及变异操作。交叉算子和变异算子的作用是产生新的个体,每次交叉变异结束,用新个体代替原个体保留在种群中。通过该种思想,使种群中个体的多样性动态控制选择算子实现的速度。由于选择算子的作用主要是实现种群的收敛,故通过本文的动态选择可以动态的调整种群收敛速度。将上述过程不断循环,最终求得全局最优解。

算法的步骤为:

1)参数初始化,种群大小 m,交叉概率 P_c,变异概率 P_m,聚类数 c;

2)初始化由 m 个个体组成的初始种群,每个个体由 c 个聚类中心组成;

3) 对于每一个个体,用式(5)~(7)计算该个体的适应度值 $F(\mathrm{ind}_i)$,其中 $i=1,2,\cdots,m$;

4) 根据适应度值 $F(\mathrm{ind}_i)$ 用式(9)、(10)计算出种群中个体的多样性度量值 Stddev(P)是否大于最低多样性度量值 $\mathrm{Std}_{\min}$,若是,则执行 5),否则执行 6);

5)根据适应度值 $F(\mathrm{ind}_i)$,用式(8)计算出选择概率,执行选择操作;

6)对种群中各个体执行交叉和变异操作产生新个体;

7)将新个体代替原个体保留在种群中;

8)若在一定迭代次数 L 内种群的性能没有改进或达到了最大迭代次数 it,跳出循环,否则执行步骤 3)。

3.8 算法复杂性分析

算法的时间复杂度可以用来度量算法运行的时间,表示算法计算效率的高低,其大小反映了算法性能的优劣,不考虑硬件及环境因素,假设每一次执行时硬件条件和环境条件是相同的,设有 n 个待聚类对象,聚类中心个数为 c,种群大小为 m,算法的迭代次数为 L,则本文算法的时间复杂度为 $O(Lncm)$,以下为对时间复杂度的说明。

1)以算法的一次迭代为例,从待聚类的 n 个对象中随机选取 m 个个体作为初始种群,其中每个个体由 c 个对象组成,代表 c 个聚类中心,其时间复杂度为 $O(cm)$。

2)求解种群中每个个体适应度时,需要遍历 n 个待聚类对象与该个体所代表的 c 个聚类中心的距离矩阵 $\boldsymbol{D}_{c\times n}$,其时间复杂度为 $O(nc)$,由于需要计算种群中 m 个个体的适应度,那么该步骤总共消耗的时间为 $O(ncm)$。

3)计算种群中个体的多样性度量值 Stddev(P)需要对种群中所有个体的适应度进行遍历,时间复杂度为 $O(m)$。

4)选择操作需要计算种群中每个个体的选择概率并根据选择概率重新选择 m 个个体,时间复杂度为 $O(2m)$。

5)交叉操作过程中,需要对种群中 m 个个体进行遍历,时间复杂度为 $O(m)$。

6)变异操作同样要遍历种群中 m 个个体,时间复杂度为 $O(m)$。

7)在每一次迭代完成后需要判断是否已经达到终止条件,此操作的时间复杂度为 $O(1)$。

通过以上的分析可以得到 1)~7)的时间复杂度为 $O(5m+cm+ncm+1)$。整体算法的迭代次数为 L,则本文算法的整体运行时间为 $O((5m+cm+ncm+1)L)$。当数据规模逐渐增大,即 n 趋于无穷大时,本文算法的时间复杂度为 $O(Lncm)$。

4　实验结果与分析

为检验 Dynamic GA-FCM 算法的有效性以及可行性,将 Dynamic GA-FCM 算法、传统 GA-FCM 算法、K-means 算法以及 Helo′lna Alves Arnaldo 等人在文献[18]所提出的改进 FCM 算法进行性能比较,分别选用人工数据集、UCI 数据集以及数字数据集进行测试。Dynamic GA-FCM 和传统 GA-FCM 算法中模糊指数 l = 2,种群个体总数 m = 100,种群最大迭代次数 it=100,变异概率 P_m = 0.05,交叉概率 P_c = 0.8。

4.1　人工数据实验

本节将人工数据分成 2 组,以测试 Dynamic GA-FCM 算法的性能,表一给出了人工数据集的性质。

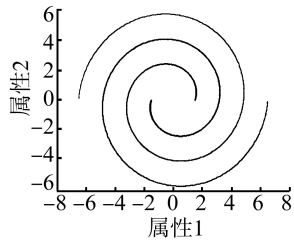
表 1　人工数据集的数据特征

Table1　Artificial data sets characteristics

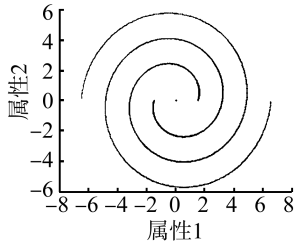
数据集	对象数/个	维数	类个数
Long1	1 000	2	2
Spiral	1 000	2	2
Lineblobs	266	2	3
Square1	1 000	2	4

4.1.1　人工数据集 I

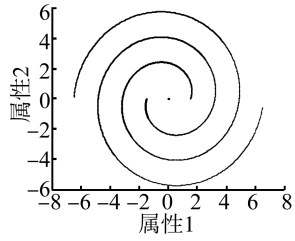
人工数据集 I 是 3 个具有复杂的流形分布的数据集,分别为数据集 Long1,Spiral 以及 Lineblobs。以下分别展示了上述 3 种算法对人工数据集 I 的聚类结果,使数据分布情况及聚类效果更为直观的显现。对于每个数据集,独立运行了 50 次。表 2 列出了各个算法在求解以上 3 个聚类问题时得到的聚类的正确率平均值。从表 2 中的统计数据和如图 2~4 所示的聚类结果可以清楚的看到,对于 Long1,Spiral 和 Lineblobs 这 3 个具有明显的流形分布特点的数据集,本文所提出的算法在总体上较之其他的 3 种算法,聚类效果更优。



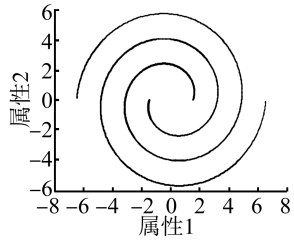
(a) 数据集正确聚类结果



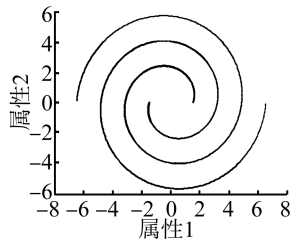
(b) 传统 GA-FCM 算法



(c) k-means 算法



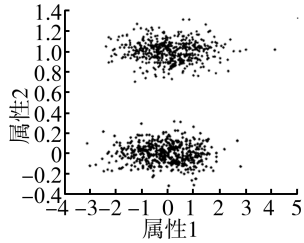
(d) 文献[13]算法



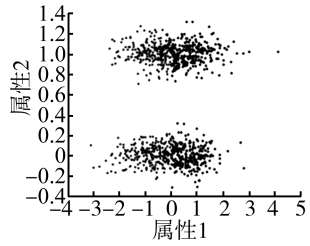
(e) 本文算法

图 2　不同算法对 Spiral 的聚类结果

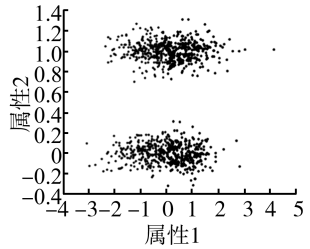
Fig.2　Clustering results on Spiral



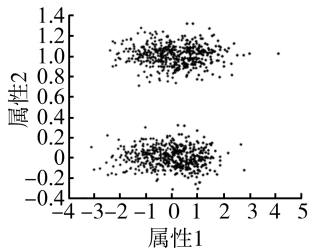
(a) 数据集正确聚类结果



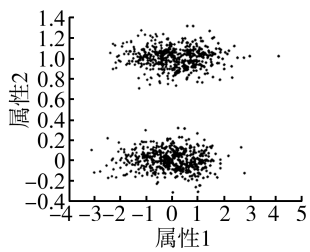
(b) 传统 GA-FCM 算法



(c) k-means 算法



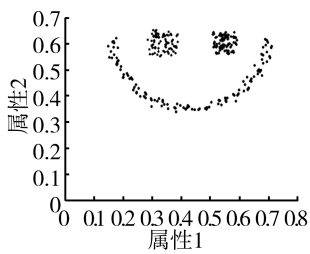
(d) 文献[13]算法



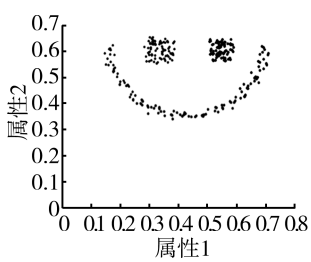
(e) 本文算法

图 3 不同算法对 Long1 的聚类结果

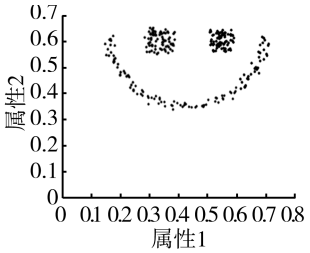
Fig.3 Clustering results on Long1



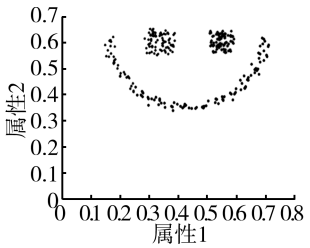
(a) 数据集正确聚类结果



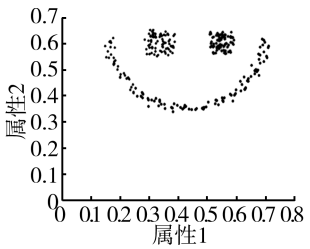
(b) 传统 GA-FCM 算法



(c) k-means 算法



(d) 文献[13]算法



(e) 本文算法

图 4 不同算法对 Lineblobs 的聚类结果

Fig.4 Clustering results on Lineblobs

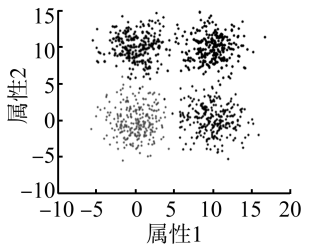
表 2 人工数据集 I 的聚类结果

Table 2 Clustering results on artificial data sets I

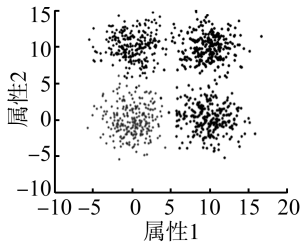
数据集	平均准确率			
	GA-FCM	k-means	文献[13]	本文算法
Spiral	0.592 0	0.589 8	0.600 1	0.681 3
Long1	0.520 1	0.515 4	0.547 3	0.652 2
Lineblobs	0.741 0	0.743 5	0.743 1	0.853 2

4.1.2 人工数据集 II

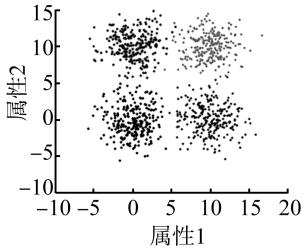
人工数据集 II 选取具有球形分布的数据集 Square1,以下分别给出 4 种算法对于 Square1 的聚类结果。同样,对于该数据集,独立地运行 50 次,在表 3 中列出了各算法求解 Square1 数据集时所得到的聚类正确率平均值。从表 3 中的统计数据以及图 5 所示的聚类结果可以看出,对于 Square1 这个球形分布的数据集,上述 4 种算法的聚类效果都比较好,且比较接近。



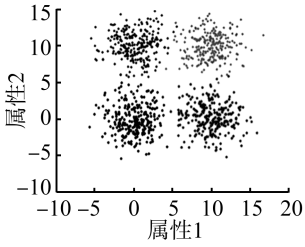
(a) 数据集正确聚类结果



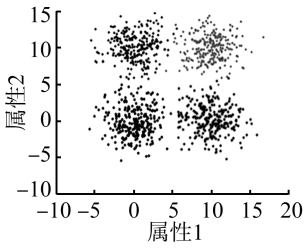
(b) 传统 GA-FCM 算法



(c) k-means 算法



(d) 文献[13]算法



(e) 本文算法

图 4　不同算法对 Square1 的聚类结果

Fig.4　Clustering results on Square1

表 3　人工数据集 II 的聚类结果

Table 3　Clustering results on artificial data sets II

数据集	平均准确率			
	GA-FCM	k-means	文献[13]	本文算法
MNIST	0.990 0	0.990 0	0.990 0	0.990 0

4.2　UCI 数据实验

此外,选择了 3 个 UCI 数据集分别的对 4 种算法进行了测试,以便更加全面的考察算法的性能。数据集来自 http://archive.ics.uci.edu/ml/,分别为 Iris 数据集、Glass 数据集以及 Teaching Assistant Evaluation 数据集,表 4 给出了 UCI 数据集的性质。

对每个数据集独立的运行 50 次后,表 5 分别列出了各个算法在求解这 3 个聚类问题时所得到的聚类正确率平均值。从实验结果中可以看到,对于 Iris,TAE,Glass 这 3 个 UCI 数据集,本文算法较之另外 3 种算法更优,说明本文的算法对于现实世界的数据聚类的问题有着很好的性能。

表 4　UCI 数据集的数据特征

Table 4　UCI data sets characteristics

数据集	对象数/个	维数	类个数
TAE	151	6	3
Glass	214	10	7
Iris	150	4	3

表 5　在 UCI 数据集下的聚类结果

Table 5　Clustering results on UCI data sets

数据集	平均准确率			
	GA-FCM	k-means	文献[13]	本文算法
Iris	0.810 2	0.807 9	0.893 5	0.903 9
TAE	0.701 6	0.699 7	0.762 6	0.802 4
Glass	0.732 5	0.722 5	0.803 0	0.817 2

4.3　数字数据集实验

为了进一步评估算法的性能,本文又在著名的 USPS 和 MNIST 2 个数字数据集上基于归一化互信息(normalized mutual information,NMI),进行了 4 种算法的对比实验。NMI 是一个外部评价标准,用其来评价在某个数据集上的聚类结果与这一数据集的真实划分的相近程度,NMI 越大,说明聚类性能越好。

4.3.1　USPS 数据集

USPS 数据集总共有 9 298 个 16 × 16 维的灰度图像样本,其中 2 007 个为测试样本,其余为训练样本。本实验选取所有测试样本作为聚类分析数据集,分别执行以下 4 组实验:

USPS-08:包含数字 0 、8 的灰度图像测试样本;

USPS-358:包含数字 3、5 、8 的灰度图像测试样本;

USPS-1234:包含数字 1、2、3、4 的灰度图像测试样本;

USPS-024679:包含数字 0、2、4、6、7、9 的灰度图像测试样本。

其中 USPS-08 和 USPS-358 这 2 组数据较难识别,USPS-1234 和 USPS-024679 则相对容易一些。

对每组数据独立的运行 50 次后,表 6 给出了 NMI 标准下 4 种算法的实验结果。

表 6　4 组 USPS 数据集在 NMI 标准下的实验结果

Table 6　Four clustering results on USPS at the NMI standard

数据集	GA-FCM	k-means	文献[13]	本文算法
USPS-08	0. 676 4	0.612 1	0.681 4	0.732 1
USPS-358	0.525 3	0.482 1	0.521 8	0.559 4
USPS-1234	0.752 3	0.724 7	0.760 7	0.800 7
USPS-024679	0.701 3	0.704 2	0.709 8	0.745 6

实验结果表明,本文算法的 NMI 值比比另外 3 种算法的对应值大,这说明本文算法在 USPS 数据集上的聚类划分结果,相比另外 3 种算法,有所提高。

4.3.2 MNIST 数据集

MNIST 数据集中每个字符标准化为 16×16 的灰度图像,总共有 70 000 个样例。对于 0~9 每一个数字都选取 200 个字符用于实验,独立运行 50 次后, 表 7 为 4 种算法的实验结果

从实验结果可以看出,本文算法在 NMI 上优于其他 3 种算法。

表 7 MNIST 数据集在 NMI 标准下的实验结果

Table 7 Clustering results on MNIST

数据集	GA-FCM	k-means	文献[13]	本文算法
MNIST	0.542 3	0. 524 1	0.604 2	0.668 7

5 结束语

本文为了克服传统的基于 GA 的 FCM 算法中存在的早熟问题,在 GA 中通过种群中个体的多样性来动态的控制 GA 选择算子实现的速度,即动态控制了收敛速度,进而设计了 Dynamic GA-FCM 算法。通过与传统 GA-FCM 算法、K-means 算法以及 Helo′lna Alves Arnaldo 等在文献[18]所提出的改进 FCM 算法进行对比实验,实验结果表明,本文提出的 Dynamic GA-FCM 算法具有更好的聚类效果。

参考文献:

[1]TAN Pangning, STEINBACH M, KUMAR V. 数据挖掘导论[M]. 北京: 人民邮电出版社, 2006: 298-320.

[2]GAO Yunguang, WANG Shicheng, LIU Shunbo. Automatic clustering based on GA-FCM for pattern recognition[C]// Computational Intelligence and Design. Changsha, China, 2009: 146-149.

[3]VIJAYACHITRA S, TAMILARASI A, KASTHURI N. Multiple input single output (MISO) process optimization using [C]//International Conference on Education Technology and Computer. Singapore, 2009: 248-252.

[4]LIU Suhua, HOU Huifang. A combination of mixture genetic algorithm and fuzzy c-means clustering[C]//IEEE International Symposium on IT in Medicine & Education. Ji′nan, China, 2009: 254-258.

[5]BELAHBIB F Z B, SOUAMI F. Genetic algorithm clustering for color image quantization[C]//2011 3rd European Workshop on Visual Information Processing. Paris, French, 2011: 83-87.

[6]BEZDEK J C, EHRLICH R, FULL W. FCM: the fuzzy c-means clustering algorithm [J]. Computers and Geosciences, 1984, 10(2/3): 191-203.

[7]LIU Zhide, CHEN Jiabin, SONG Chunlei. A new RBF neural network with GA-based fuzzy c-means clustering[C]// Chinese Control and Decision Conference. Guilin, China, 2009: 208-211.

[8]HOLLAND J H. Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence[M]. Cambridge: MIT Press, 1992: 95-110.

[9]WANG Jianxin. Reducing the overlap among hierarchical clusters with a GA-based approach[C]//2009 1st International Conference on Information Science and Engineering. Nanjing, China, 2009: 924-927.

[10]MENENDEZ H D, BARRERO D F, CAMACHO D. A multi-objective genetic graph-based clustering algorithm with memory optimization[C]//2013 IEEE Congress on Evolutionary Computation. Cancun, Mexico, 2013: 3174-3181.

[11]RAZIZADEH N, BADAMCHIZAEH M A, GHASEMPOUR M S G. A new GA based method for improving hybrid clustering[C]//2013 21st Iranian Conference on Electrical Engineering. Mashhad, Iran, 2013: 1-6.

[12]NGUYEN D D, NGO L T. Multiple kernel interval type-2 fuzzy c-means clustering[C]//IEEE International Conference on Fuzzy Systems (FUZZ). Hyderabad, India, 2013: 1-8.

[13]ARNALDO H A, BEDREGAL B R C. A new way to obtain the initial centroid clusters in fuzzy c-means algorithm [C]//2013 2nd Workshop-School on Theoretical Computer Science (WEIT). Rio Grande, Brazil, 2013: 139-144.

[14]NGUYEN D H M, WONG K P. Controlling diversity of evolutionary algorithms[C]//Proceedings of the 2nd International Conference on Machine Learning and Cybernetics. Guilin, China, 2003: 775-780.

[15]ZHU K Q. Population diversity in genetic algorithm for vehicle routing problem with time windows[C]//European Conference on Machine Learning. Pisa, Italy, 2004: 537-547.

[16]CHENG S S, CHAO Y H, WANG H M, et al. A prototypes-embedded genetic K-means algorithm[C]//18th International Conference on Pattern Recognition. Hong Kong, China, 2006: 724-727.

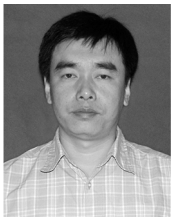
[17]REZAEE M R, LELIEVELDT B P F, REIBER J H C. A new cluster validity indexes for the fuzzy c-mean[J]. Pattern Recognition Letters, 1998, 19(3-4): 237-246.

[18]BANDYOPADHYAY S, MAULIK U. An evolutionary technique based on K-means algorithm for optimal clustering in RN[J]. Information Sciences, 2002, 146(1-4): 221-237.

[19] SABAU A S. Variable density based genetic clustering [C]//2012 14th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC). Timisoara, Romania, 2012: 200-206.

[20] XIAO Huiming. Application in performance assessment of the clinical department in hospital based on fuzzy cluster and genetic algorithm[C]//2014 International Conference on Computational Intelligence and Communication Networks. Bhopal, India, 2014: 1057-1061.

作者简介:



张永库,男,1972 年生,副教授,主要研究方向为图形图像和多媒体、数据处理和数据挖掘,先后主持和参加“唐钢集团庙沟铁矿管理信息系统”、“基于点模型的布尔运算和混合建模技术”、“义煤集团医疗保险管理系统”、“滑坡灾害远程智能监测系统”等 10 余项课题,获市科技进步一等奖 1 项、市科技进步二等奖 4 项、辽宁省教学成果一等奖 1 项。

尹灵雪,女,1991 年生,硕士研究生,主要研究方向为数据处理和数据挖掘。

孙劲光,女,1962 年生,教授,博士,主要研究方向为图形图像和多媒体、数据处理和数据挖掘。

[责任编辑:郑可为]

2015 年第 13 届文档分析与识别国际会议
2015 13th International Conference on Document Analysis and Recognition (ICDAR)

ICDAR is the premier international forum for researchers and practitioners in the document analysis community for identifying, encouraging and exchanging ideas on the state-of-the-art technology in document analysis, understanding, retrieval, and performance evaluation. The term document in the context of ICDAR encompasses a broad range of documents from historical forms such as palm leaves and papyrus to traditional documents and modern multimedia documents.

Areas:

1) Character and symbol recognition
2) Printed/Handwritten text recognition
3) Graphics analysis and recognition
4) Document analysis
5) Document understanding
6) Historical documents and digital libraries
7) Document based forensics
8) Camera and video based scene text analysis

Website: http://2015.icdar.org/