

文章编号:1001-9081(2011)05-01355-04

doi:10.3724/SP.J.1087.2011.01355

融合粒子群和混合蛙跳的模糊 C-均值算法

李真,罗可

(长沙理工大学 计算机与通信工程学院,长沙 410114)

(lz1357924680@sina.com)

摘要:针对模糊聚类算法中存在的对初始值敏感、易陷入局部最优等问题,提出了一种融合粒子群算法和混合蛙跳算法的模糊 C-均值聚类算法。通过设计了一种新颖的搜索粒度系数,充分利用粒子群算法收敛速度快、局部搜索能力强的优点与混合蛙跳算法全局寻优能力强、跳出局部最优能力好的特点,同时对 SFLA 中更新算法进行了改进。实验结果表明,该算法提高了模糊聚类算法的搜索能力和聚类效果,在全局寻优能力、跳出局部最优能力、收敛速度等方面具有优势。

关键词:混合蛙跳算法;粒子群算法;模糊 C-均值;目标函数

中图分类号:TP311.13 **文献标志码:**A

Improved FCM algorithm based on PSO and SFLA

LI Zhen, LUO Ke

(School of Computer and Communication Engineering, Changsha University of Sciences and Technology, Changsha Hunan 410014, China)

Abstract: The traditional fuzzy clustering algorithm is sensitive to the initial point and easy to fall into local optimum. In order to overcome these flaws, an improved Fuzzy C-Mean (FCM) algorithm which combines the Particle Swarm Optimization (PSO) algorithm and Shuffled Frog Leaping Algorithm (SFLA) was proposed. Through designing a new search granularity factor, it could take advantage of the fast convergence speed, strong local search ability of PSO and strong global search capability, ability to jump of local optimum of SFLA, making the integration of PSO and SFLA better. At the same time, the update algorithm of SFLA was improved. The experimental results show that this method improves the search capability and the clustering performance of fuzzy clustering algorithm, and it has the advantages in the global search ability escaping from local optimum capacity, and convergence speed.

Key words: Shuffled Frog Leaping Algorithm (SFLA); Particle Swarm Optimization (PSO) algorithm; Fuzzy C-Mean (FCM); objective function

模糊 C-均值(Fuzzy C-Mean, FCM)聚类算法由 Dunn 提出,后被 Bezdek 改进^[1],被广泛地应用于数据挖掘、图像分割、模式识别等方面^[2],但 FCM 算法存在一些缺点:如对初始值、噪声数据敏感,容易陷入局部最优等。近几年,已有一些基于智能优化的聚类算法被提出^[3-7],如文献[8-9]基于粒子群优化(Particle Swarm Optimization, PSO)算法提出的 PSO-FCM 算法^[10-11],该算法在一定程度上解决了 FCM 算法对初始化敏感的问题,但有时也会陷入局部最优。

PSO 算法具有收敛快、参数少、局部搜索能力强等优点,但跳出局部最优能力较弱。混合蛙跳算法(Shuffled Frog Leaping Algorithm, SFLA)^[12-13]全局寻优能力强,跳出局部最优能力好,但收敛速度较慢,局部搜索能力较差。本文在已有文献[8-9]的基础上,利用 PSO 算法与 SFLA 寻优能力强等优点,将两者结合提出了新的模糊聚类算法——SP-FCM 算法,通过实验证明,该算法能有效地提高模糊聚类算法的搜索能力与聚类效果。

1 知识预备

1.1 FCM 算法

设样本空间, c 为大于 1 的正整数。将 X 分为 c 类可以

收稿日期:2010-10-13;修回日期:2011-01-15。 基金项目:国家自然科学基金资助项目(10926189;10871031);湖南省自然科学-衡阳联合基金资助项目(10JJ8008);湖南省教育厅重点项目(10A015)。

作者简介:李真(1986-),男,湖南洪江人,硕士研究生,主要研究方向:数据挖掘、数据库; 罗可(1961-),男,湖南长沙人,教授,博士,主要研究方向:数据挖掘、数据库。

用一个模糊矩阵 $\mu = (\mu_{ij})$ 表示, μ_{ij} 表示第 i 个样本点属于第 j 个隶属度。

FCM 算法目标函数 $J(\mu, A)$ 定义如下:

$$J(\mu, A) = \sum_{i=1}^c \sum_{j=1}^n \mu_{ij}^m \|X_j - A_i\|^2 \quad (1)$$

其中: $m (m > 1)$ 是模糊指数, $A_i (i = 1, 2, \dots, c)$ 是聚类中心。

FCM 算法的目标就是将目标函数 $J(\mu, A)$ 最小化的迭代收敛过程。在迭代过程中 μ, A 的取值公式如下:

$$\mu_{ij} = \begin{cases} \left(\sum_{k=1}^c \frac{\|X_j - A_i\|^{2/(m-1)}}{\|X_j - A_k\|^{2/(m-1)}} \right)^{-1}, & \|X_j - A_k\| \neq 0 \\ 1, & \|X_j - A_k\| = 0, k = j \\ 0, & \|X_j - A_k\| = 0, k \neq j \end{cases} \quad (2)$$

$$A_i = \sum_{j=1}^n \mu_{ij}^m X_j / \sum_{j=1}^n \mu_{ij}^m \quad (3)$$

1.2 PSO 算法

在 PSO 算法中,每个粒子都认为是 d 维空间内的一个点。第 t 个粒子当前最好的位置表示为 P_t ,在 P_t 中最好的粒

子成员表示为 G_t , 第 t 个粒子位置变化的快慢用 V_t 表示。标准的粒子群优化算法的速度与位置更新公式如下:

$$\begin{aligned} V_t &= \omega V_t + r_1 \cdot \text{rand}() (\mathbf{P}_t - \mathbf{X}_t) + r_2 \cdot \text{rand}() (\mathbf{G}_t - \mathbf{X}_t) \\ (4) \end{aligned}$$

$$\mathbf{X}_t = \mathbf{X}_t + V_t \quad (5)$$

其中: r_1, r_2 为加速常数; $\text{rand}()$ 为区间 $[0,1]$ 上均匀分布的随机数; ω 为惯性权重系数, 用来调节搜索能力^[14]。目前普遍采用将 ω 设置为从 0.9 到 0.1 线性下降的方法:

$$\omega = \omega_{\max} - run \cdot \frac{\omega_{\max} - \omega_{\min}}{runMax} \quad (6)$$

1.3 SFLA

在 SFLA 中, 首先从已知可行域中随机初始化 Q 个解组成一个群体, 计算出每个解的适应值 $f(x_i)$, 将每个解按适应值按降序排列。然后将整个青蛙群体分为 q 个子群, 每个子群中包含 Q/q 个解, 在迭代过程中, 第 i 个解放入第 $i \bmod q$ 的子群中, 直到所有解分配完毕。在每一个子群中, 适应值最好最差的解分别记为 X_{best} 和 X_{worst} , 整个群体中适应值最好的解记为 X_{gbest} 。每次迭代中, 对各子群的 X_{worst} 进行更新操作, 更新成功后, 混合更新后的子群。其更新策略为:

$$\mathbf{D}_i = \text{rand}() \times (X_{\text{best}} - X_{\text{worst}}); i = 1, 2, \dots, p \quad (7)$$

$$\mathbf{X}'_{\text{worst}} = \mathbf{X}_{\text{worst}} + \mathbf{D}_i; -\mathbf{D}_{\max} \leq \mathbf{D}_i \leq \mathbf{D}_{\max} \quad (8)$$

其中: $\text{rand}() \in U(0,1)$, \mathbf{D}_{\max} 为最大移动步长。

2 算法基本思想与算法流程

2.1 算法基本思想

由于 FCM 算法的本质也是一种局部搜索的算法, 具有以下几个不足:

- 1) 对初始值和噪声数据敏感, 易陷入局部最优;
- 2) 如果初始簇中心远离最优点, 会大大降低算法的收敛速度;
- 3) 同时如果初始簇中心在局部最优点附近, 就会造成算法收敛于局部最优。

针对以上的不足, 本文利用 PSO、SFLA 等智能优化算法的寻优能力, 找出最优的初始聚类中心, 从而避免 FCM 算法对初始值的敏感、易陷入局部最优等问题。具体的理论依据如下:

1) PSO 算法具有较强的全局寻优能力、收敛速度快、参数少等优点。利用 PSO 算法能快速地寻找最优初始簇, 但也易陷入局部最优。

2) SFLA 具全局寻优能力强, 以及较强跳出局部最优能力等优点, 能加强最优初始聚类中心搜索能力。但该算法局部搜索能力差, 收敛速度慢。

为此, 通过设置一个搜索粒度系数将 SFLA 与 PSO 算法结合起来, 利用二者的优点, 寻找最优聚类中心, 来克服 FCM 算法的缺点。找出最优聚类中心之后在用 FCM 算法的基本算法进行聚类。

2.2 算法流程

第 1 步 初始化 Q 个聚类中心个体, 初始化各参数。

第 2 步 对每个个体用式(2)计算模糊矩阵, 用式(9)计

算适应度值 $f(x_i)$ 。

第 3 步 按适应值降序排列, 将每个个体按 SFLA 分配到 q 子群。根据适应度值计算每个子群的 X_{best} 、 X_{worst} 、 X_{gbest} 。

第 4 步 用式(4)、(5)更新 V_j 、 X_j 。

第 5 步 如果 $run < subrunMax$, 转第 4 步; 否则转第 6 步。

第 6 步 用式(7)、(8)更新 X_{worst} , 更新成功转第 8 步; 否则转第 7 步。

第 7 步 用式(12)更新 X_{worst} 。

第 8 步 更新 $f(x_j)$, 如果 $run < runMax$, 转第 3 步; 否则转第 9 步。

第 9 步 找出适应值最好的 X_{gbest} , 即最优聚类模糊矩阵中心集合。

第 10 步 根据最优聚类中心计算模糊矩阵, 输出结果。

3 SP-FCM 算法的软件实现

本算法具体功能由 4 大模块分别实现: Init 模块、SP 模块、Fcm 模块和 Output 模块。各模块具体功能描述如下。

3.1 Init 模块

本模块的功能就是初始化 Q 个聚类中心, 设聚类样本集为 $X_t = \{x_1, x_2, \dots, x_i, \dots, x_n\}$, 其中 x_i 为 d 维向量。以一个个体代表一个簇的中心集合 $A_t = \{c_1, c_2, \dots, c_i, \dots, c_k\}$, 其中 c_i 为第 i 个聚类中心的编码, 也是 d 维。每类 d 维聚类中心采用实数编码, 并计算模糊矩阵及适应度。初始化 Q 个聚类中心, 对于个体 c_i 的评价的适应值函数定义为:

$$f(x_i) = \frac{1}{J(\mu, A) + 1} \quad (9)$$

3.2 SP 模块

本模块的功能是寻找最优的聚类中心。根据适应值, 寻找最优聚类中心, 通过设计的一个搜索粒度系数将 PSO 算法与 SFLA 相结合, 通过将 PSO 算法嵌套到 SFLA 内, 并对 SFLA 的更新算子改进的综合算法。该模块基本流程如图 1 所示。

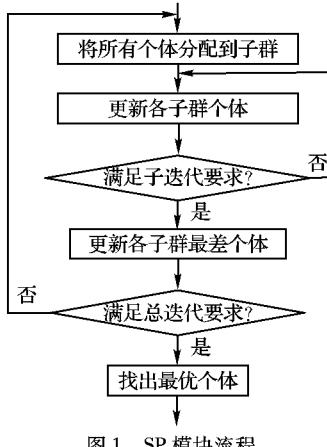


图 1 SP 模块流程

3.2.1 搜索粒度系数

由于前期需要充分利用 PSO 算法较强的搜索能力, 缩小搜索范围, 加快收敛的优点, 迭代后期则需要充分利用 SFLA 较强的跳出局部最优的能力, 扩大搜索范围, 避免陷入局部最优。PSO 算法与 SFLA 各有优缺点。简单讲二者嵌套结合起

来虽然利用到了二者的优点,但同时也会继承二者的缺点,每次搜索时都通过各子群使用 PSO 算法的更新策略,同时再对各子群进行蛙跳算法混合更新操作,无疑降低了算法的收敛速度,同时也削弱了 PSO 算法的局部搜索能力。因此本文设置一个搜索粒度系数,以充分利用二者的优点进行搜索,搜索粒度系数定义:

$$subrunMax = \left[\sqrt{\frac{runMax}{run + \varepsilon}} \right] \quad (10)$$

其中: ε 为初始搜索粒度(本文取 $\varepsilon = 2$), η 为全局搜索粒度(本文取 $\eta = 2$), $subrunMax$ 为搜索粒度系数, $runMax$ 为最大迭代次数, run 为当前迭代次数。

3.2.2 SFLA 的改进

对于基本的 SFLA,在算法执行第 7 步和第 8 步更新失败后,迭代时随机生成一个新解取代原来的 X_{worst} ,可能导致陷入局部最优,收敛速度减慢。为了解决该问题,使个体避免在已搜索过的区域重复搜索,需要缩小随机生成新解的范围,即在当前 X_{worst} 附近随机产生新解。

由于高斯算子同样具有很好的随机性,以及较强局部搜索能力,但同样也有陷入局部最优的可能。混沌变异算子具有伪随机性,同时能避免陷入局部最优解,但是由于混沌变异算子对初始值相当敏感,每两次迭代的差值方差较大,对于局部搜索能力较弱。本文将高斯算子和混沌变异算子两者结合,利用混沌变异算子跳出局部最优能力,再对个体进行搅动。本文采用 Logistic 映射函数进行变异操作:

$$z_n = \xi z_n (1 - z_n) \quad (11)$$

其中 $\xi \in [1, 4]$ 为控制参数(本文取 $\xi = 4$)。利用该映射随机生成一个与样本同维的向量 z_n ($n = 0, 1, 2, \dots$) 对 X_{worst} 进行更新:

$$X_{worst} = X_{worst} + N(\delta, \sigma^2) \lambda \quad (12)$$

其中: δ 为均值, σ^2 为方差的高斯算子。

3.3 FCM 模块与 Output 模块

FCM 模块的功能是在得到最优聚类中心 A 之后,通过 FCM 算法更新模糊矩阵,Output 模块根据模糊矩阵输出聚类结果。

4 仿真实验

本文在 Windows XP 系统下利用 Matlab 7.0 作为算法的现实平台,利用 UCI 数据库的 Iris、Wine 数据集和利用二维随机产生的分布数的数据检验两个方面来验证改进算法的聚类效果。

4.1 检查聚类效果

利用 UCI 数据库的 Iris、Wine 数据集检查聚类效果。在本实验中,分别用 FCM、PSO-FCM、SP-FCM 对 Iris 数据进行聚类分析,实验参数见表 1。

实验结果如图 2~3 和表 2 所示。图 2 中,FCM 算法在 15 代左右就趋向于收敛,PSO-FCM 与 SP-FCM 算法在 20 代左右才收敛于最优,图 3 中,FCM 算法在 15 代左右就趋向于收敛,PSO-FCM 算法在 30 代左右才收敛于最优,SP-FCM 算法,在 40 代左右才收敛于最优。可见 SP-FCM 算法有效地克服了 FCM 算法陷入局部最优的缺点。

由图 2~3 可看出:迭代初期 SP-FCM 算法与 PSO-FCM 算法的收敛速度近似,中后期 SP-FCM 算法收敛速度高于 PSO-FCM 算法。SP-FCM 算法收敛的目标函数值比 PSO-FCM 及 FCM 算法均小,可见 SP-FCM 算法收敛也较快,聚类效果也比 PSO-FCM 算法及 FCM 算法好。

表 1 实验参数表

| 参数 | 算法 | | |
|-------------------------|-----|---------|--------|
| | FCM | PSO-FCM | SP-FCM |
| 种群规模 Q | — | 80 | 80 |
| 子群规模 q | — | — | 4 |
| 模糊指数 m | 2 | 2 | 2 |
| 最大惯性权重系数 ω_{max} | — | 0.9 | 0.9 |
| 最小惯性权重系数 ω_{min} | — | 0.1 | 0.1 |
| 加速常数 r_1, r_2 | — | 2 | 2 |
| Logistic 映射控制参数 ξ | — | — | 4 |
| 高斯算子均值 δ | — | — | 1 |
| 高斯算子方差 σ^2 | — | — | 0 |
| 初始搜索粒度 ε | — | — | 2 |
| 全局搜索粒度 η | — | — | 2 |
| 总迭代次数 $runMax$ | 100 | 100 | 100 |
| 运行次数 | 100 | 100 | 100 |

注:“—”表示无该项参数。

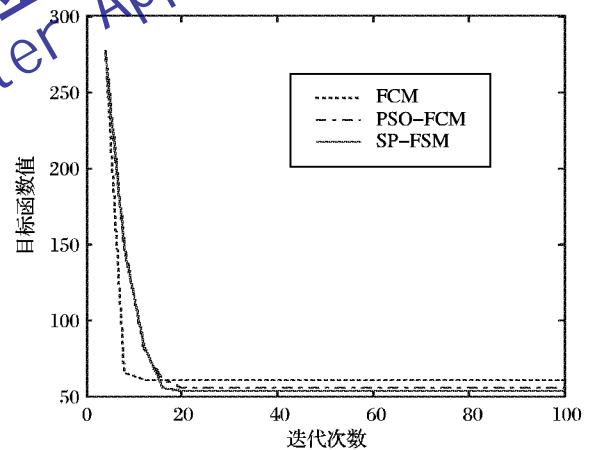


图 2 Iris 数据集测试数据的收敛曲线

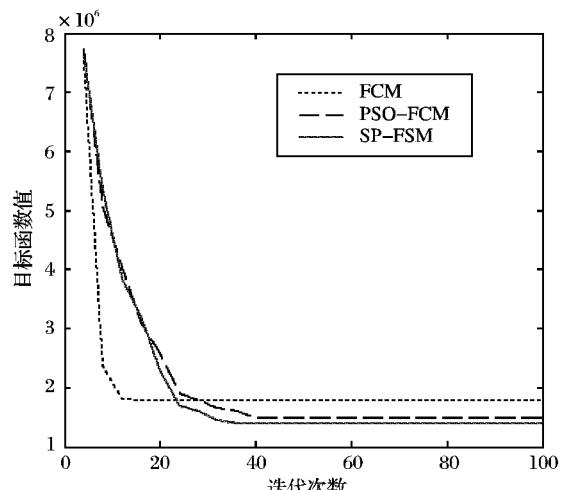


图 3 Wine 数据集测试数据的收敛曲线

表 2 显示,SP-FCM 算法对于 Iris 数据集与 Wine 数据集

聚类的平均聚类正确率分别达到 92.67% 和 75.84%, 均比 PSO-FCM 算法与 FCM 算法高。特别是对 Wine 数据集聚类, 聚类效果有明显改善。

4.2 利用二维随机产生的分布数的数据检验聚类效果。

在本实验中, 分别用 FCM、PSO-FCM、SP-FCM 算法对在

[0.1,1] 内随机生成的 100 个二维数据进行聚类分析, 将样本数据聚成 3 类。实验参数同表 1, 实验数据见图 4。实验结果见表 3。由表 3 可看出, SP-FCM 算法的目标函数平均值比 FCM 算法与 PSO-FCM 算法更好, 说明 SP-FCM 算法的聚类效果有明显改善。

表 2 FCM、PSO-FCM、SP-FCM 聚类对比分析表

| 算法 | Iris | | | Wine | | |
|---------|---------|---------|-----------|---------|---------|-----------|
| | 平均聚类正确数 | 平均聚类错误数 | 平均聚类正确率/% | 平均聚类正确数 | 平均聚类错误数 | 平均聚类错误数/% |
| FCM | 134 | 16 | 89.33 | 122 | 56 | 68.54 |
| PSO-FCM | 138 | 12 | 92.00 | 128 | 50 | 71.90 |
| SP-FCM | 139 | 11 | 92.67 | 135 | 43 | 75.84 |

表 3 FCM、PSO-FCM、SP-FCM 聚类对比分析表

| 算法 | 聚类中心 1 | 聚类中心 2 | 聚类中心 3 | 目标函数平均值 |
|---------|----------------|----------------|----------------|-----------|
| FCM | (0.236, 0.277) | (0.618, 0.594) | (0.820, 0.812) | 0.150 424 |
| PSO-FCM | (0.231, 0.269) | (0.620, 0.640) | (0.835, 0.824) | 0.153 547 |
| SP-FCM | (0.230, 0.260) | (0.620, 0.636) | (0.836, 0.828) | 0.163 242 |

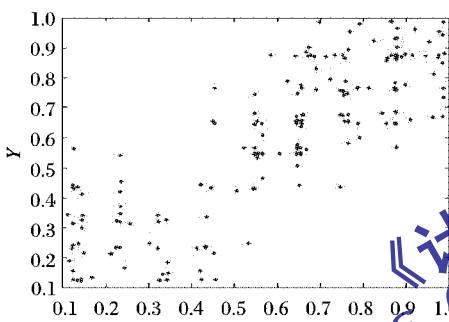


图 4 随机测试数据

5 结语

本文根据 SFLA 与 PSO 算法及 FCM 算法的特点, 在已有文献的基础上, 通过对搜索粒度系数参数的设置, 将 PSO 算法和 SFLA 有机地结合应用在 FCM 算法上, 提出了 SP-FCM 算法, 较好地解决了由于 FCM 算法对初始值敏感、陷入局部最优而引起聚类效果不佳这一问题。将其成功地运用到 FCM 算法中, 达到了预期效果。仿真实验表明: 运用 SP-FCM 算法较其他方法所得聚类效果更好, 跳出局部最优能力更强, 具有较好的全局收敛性。

参考文献:

- HAN JIAWEI, KAMBER M. 数据挖掘: 概念与技术 [M]. 范明, 孟小峰, 译. 北京: 机械工业出版社, 2005.
- BEZDEK J C. Pattern recognition with fuzzy objective function algorithms [M]. New York: Plenum Press, 1981.
- 刘向东, 沙秋夫, 刘勇奎. 基于粒子群算法的聚类分析 [J]. 计算机工程, 2006, 32(6): 201–202.
- 宋娇, 葛临东. 一种遗传模糊聚类算法及其应用 [J]. 计算机应用, 2008, 28(5): 1197–1199.
- 杜长海, 黄席樾, 杨祖元, 等. 改进的 FCM 聚类在交通时段自动划分中的应用 [J]. 计算机工程与应用, 2009, 45(24): 190–193.
- 况夯, 罗军. 基于遗传 FCM 算法的文本聚类 [J]. 计算机应用, 2009, 29(12): 558–560.
- TANG HAI-YAN, DING BAO, QI WEI-GUI. Research on traffic mode of elevator applied fuzzy C-means clustering algorithm based on PSO [C]// IEEE International Conference on Measuring Technology and Mechatronics Automation. Washington, DC: IEEE, 2009: 582–585.
- 李丽丽, 刘希玉, 刘涛, 等. 一种基于粒子群优化的 FCM 聚类方法 [J]. 计算机应用技术, 2008(1): 89–92.
- LIU HSIANG-CHUAN, YIH JENG-MING, WU DER-BANG, et al. Fuzzy C-mean clustering algorithms based on picard iteration and particle swarm optimization [C]// IEEE International Conference on Education Technology and Training. Washington, DC: IEEE, 2008: 838–842.
- KENNEDY J, EBERHART R C, SHI Y. Swarm intelligence [M]. San Francisco: Morgan Kaufmann Publisher, 2001.
- JING WEI, ZHAO HAI, SONG CHUN-HE, et al. A optimized particle filter based on PSO algorithm [C]// IEEE International Conference on BioMedical Information Engineering. Washington, DC: IEEE, 2009: 122–125.
- EUSUFF M M, LANSEY K E. Optimization of water distribution network design using the shuffled frog leaping algorithm [J]. Journal of Water Sources Planning and Management, 2003, 129(3): 210–225.
- ZHEN ZIYANG, WANG DAOBO, LIU YUANYUAN. Improved shuffled frog leaping algorithm for continuous optimization problem [C]// IEEE Congress on Evolutionary Computation. Washington, DC: IEEE, 2009, 5: 2992–2995.
- SHI Y, EBERHART R. A modified particle swarm optimizer [C]// IEEE International Conference on Evolutionary Computation. Washington, DC: IEEE, 1998: 69–73.