基于人工萤火虫的模糊聚类算法研究

骆东松，李雄伟，赵小强

（兰州理工大学 电气工程与信息工程学院，兰州 730050）

摘要： 模糊 C－均值（FCM）聚类算法是数据挖掘中常用的方法之一，但往往受到初始聚类中心影响，收敛结果易陷入局部极小值的问题。该文提出了一种基于人工萤火虫（GSO）的模糊聚类算法（GSFM）。该算法引入了全局寻优能力强的人工萤火虫算法来求得最优解作为 FCM 算法的初始聚类中心，然后利用 FCM 算法优化初始聚类中心，最后求得全局最优解，从而有效克服了 FCM 算法的缺点。实验结果表明，新算法与 FCM 聚类算法相比，提高了算法的寻优能力，并且迭代次数更少，收敛速度更快，聚类效果更好。

关键词： 数据挖掘；模糊 C－均值聚类；人工萤火虫算法；GSFM

中图分类号： O159　　**文献标志码：** A　　**文章编号：** 1000－0682（2013）02－0003－04

Research on fuzzy clustering algorithm based on GSO

LUO Dongsong，LI Xiongwei，ZHAO Xiaoqiang

（*College of Electrical and Information Engineering，Lanzhou Univ. of Tech.* ，*Lanzhou* 730050，*China*）

Abstract: Fuzzy C-means（FCM）clustering algorithm is one of the most commonly used methods in data mining，such as being sensitive to initial conditions，usually leading to local minimum results. Therefore，a new glowworm swarm optimization（GSO）-based fuzzy algorithm（GSFM）is put forward in this paper. GSFM algorithm uses the capacity of global search in GSO algorithm to seek optimal solution as initial clustering-centers of FCM algorithm，and then use FCM algorithm to optimize initial clustering-centers，so as to get the global optimum. Above all，it solves the problems of FCM. According to the test，compared with the FCM clustering algorithm，the new algorithm improves the optimization ability of the algorithm，the number of iterations is fewer，and the convergence speed is faster. In addition，there is also a large improved at the clustering result.

Keywords: data mining；fuzzy C-mean clustering；glowworm swarm optimization；GSFM

0　引言

聚类（Clustering）分析是数据挖掘技术的重要组成部分，它能从潜在的数据中发现新的、有意义的数据分布模式。聚类是在事先不规定分组规则的情况下，将数据按照其自身特征划分成不同的群组。其重要特征是“物以类聚”，即要求在不同群组的数据之间差距越大、越明显越好，而每个群组内部的数据之间要尽量相似，差距越小越好。因此这种类别的划分界限是分明的。但是在现实世界中却有许多实际问题并没有严格的属性，它们在性态和类属方面存在着模糊性，具有“亦此亦彼”的性质。因此，人们就提出了要对待处理的对象进行软划分，L. A. Zedeh 提出的模糊集理论为软划分提供了有力的分析工具，并把用模糊的方法处理聚类问题称之为模糊聚类分析。

模糊 C－均值（FCM，其中 C 表示聚类类别数）聚类算法是由 Dunn[1] 和 Bezdek[2] 建立的，是基于目标函数的聚类算法中理论最为完善，应用最为广泛的一种方法。FCM 算法计算简单且快速，具有比较直观的几何意义，而且已被应用于模式识别、图像处理以及计算机视觉等众多邻域中。但其缺点也是显而易见的，概括起来主要有以下几个方面：（1）初

收稿日期：2012－09－11

基金项目：国家自然科学基金项目（61005026）；甘肃省自然科学基金项目（1112RJZA028）

作者简介：李雄伟（1984），男，陕西周至人，硕士研究生，研究方向为数据挖掘。

始中心个数需要预先给出，而且没有准则可遵循；(2) 只能识别团状的簇，不能识别不规则簇和带状簇，很多情况下对噪声敏感；(3) 对初始聚类中心敏感，容易陷入局部最优，难以取得全局最优或者整个聚类过程需要很长时间才能收敛到全局最优，从而影响聚类效果。文献[3]中针对 C-means 算法提出的多中心思想并结合 DBSCAN 算法在一定程度上解决了(1)和(2)问题。近年来许多文献提出了针对问题(3)的改进方法，归纳起来主要有：结合遗传算法的改进，结合免疫算法的改进，结合群智能的改进等。如一种基于改进型遗传算法的模糊聚类[4]、基于人工免疫粒子群优化算法的动态聚类分析[5]、基于 APSO 的模糊聚类算法[6]等。这些算法采用不同的方式对 FCM 算法进行改进，在一定程度上使 FCM 算法的性能更好。该文把具有全局搜索能力强且易于并行实现的人工萤火虫算法引入到 FCM 算法中，来改进 FCM 算法易陷入局部最优和对初始聚类中心敏感的缺陷。通过经典的数据集测试，实验结果表明这种改进是有效的。

1 FCM 算法

模糊 C-均值聚类算法是一种迭代优化算法，可以描述为最小化指数函数。设集合 $X=\{x_1,x_2,\cdots,x_n\}$ 是特征空间 R^n 上的一个有限数据集合，再把 X 划分为 c 类($2\leqslant c\leqslant n$)，设有个数为 c 的聚类中心 $V=\{v_1,v_2,\cdots,v_c\}$。$n\times c$ 维矩阵 $U=(u_{ij})$，$u_{ij}\in[0,1]$ 表示每个样本的隶属度矩阵。其中，$i=1,2,\cdots,n$；$j=1,2,\cdots,c$。

FCM 算法的目标函数如下：

$$J_{\mathrm{FCM}}(U,V)=\sum_{i=1}^{n}\sum_{j=1}^{c}u_{ij}^{m}\|x_i-v_j\|^2 \tag{1}$$

在式(2)的约束下取得极小值：

$$\sum_{j=1}^{c}u_{ij}=1,u_{ij}\in[0,1] \tag{2}$$

应用拉格朗日乘法，结合式(2)的约束条件对式(1)求导，得：

$$v_j=\frac{\sum_{i=1}^{n}u_{ij}^{m}x_i}{\sum_{i=1}^{n}u_{ij}^{m}} \tag{3}$$

$$u_{ij}=\frac{1}{\sum_{k=1}^{c}\left(\frac{\|x_i-v_j\|^2}{\|x_i-v_k\|^2}\right)^{\frac{2}{m-1}}} \tag{4}$$

其中：m 为模糊加权指数。

模糊 C 均值算法(Fuzzy C-Mean)简称为 FCM 算法，该算法利用式(1)的目标函数 $J_{\mathrm{FCM}}(U,V)$ 的迭代优化从而获取对数据集的模糊分类，使 $J_{\mathrm{FCM}}(U,V)$ 收敛到一个局部极小点。从模糊聚类理论的研究现状来看，现在人们已经提出了诸多算法，这也说明了现有算法还存在着种种不尽如人意的地方。在众多的模糊聚类算法中，模糊 C-均值(FCM)类型的算法理论最为完善，它不仅有着深厚的数学基础(正交投影和均方逼近理论)，而且在众多领域上已经获得了成功的应用，是目前最实用的算法之一，但是即便如此，该算法仍不是无懈可击的，还遗留了许多问题以待解决。FCM 算法对聚类效果的优劣和稳定性是否可靠在很大程度上取决于参数的初值选取，合适的 m 值具有抑制噪音、平滑隶属函数等功效，但是如何优选参数 m，尚缺乏理论指导。尽管存在一些经验值或经验范围，但没有面向问题的优选方法，也缺少有效的评价准则。另外，FCM 类型的聚类算法属于划分方法，对于一组给定的样本集，不管数据中有无聚类结构，也不问分类结果是否有效，总把数据划分到 C 个子类中。换言之，现有的聚类分析与聚类趋势以及有效性分析是隔断的、分离的。还有初始聚类中心的选择等问题。当然，FCM 类型的模糊聚类算法存在的问题远不止上述这些，但这些问题是最基本的。其实，这些问题是基于目标函数的模糊聚类所共有的，归根到底是由目标函数的 5 大参数所引起的。因此，要想解决这些问题，完善 FCM 类型的模糊聚类算法，就必须优化相关的参数。

2 基于人工萤火虫的 FCM 算法

2.1 人工萤火虫算法

人工萤火虫算法(Glowworm Swarm Optimization，GSO)[7-8]是模拟自然界中萤火虫成虫发光的生物学特性发展而来，也是基于群体搜索的随机优化算法。由印度学者 Krishnanand 等人于 2005 年提出。

在基本 GSO 中，把 n 个萤火虫个体随机分布在一个 D 维目标搜索空间中。每个萤火虫都携带了萤光素 l_i。萤火虫个体都发出一定量的萤光相互影响周围的萤火虫个体，并且拥有各自的决策域 r_s($0<r_d^i\leqslant r_s$)。萤火虫个体的萤光素大小与自己所在位置的目标函数有关，荧光素越大、越亮的萤火虫表示它所在的位置越好，即有较好的目标值。萤火虫会在决策域内寻找邻居集合 N_i，在集合中，荧光

素越大的邻居拥有越高的吸引力，吸引萤火虫往这个方向移动，每一次移动的方向会随着选择的邻居不同而改变。另外，决策域的大小会受到邻居数量的影响，邻居密度越小，萤火虫的决策半径会加大以便寻找更多的邻居；邻居密度越大，它的决策半径则会缩小。最后，大部分萤火虫会聚集在多个位置上。初始萤火虫时，每个萤火虫个体都携带了相同的萤光素浓度 l_0 和感知半径 r_0。

1) 萤光素更新

每只萤火虫 i 在 t 迭代的位置 $x_i(t)$ 对应的目标函数值 $f(x_i(t))$ 转化为荧光素值 $l_i(t)$

$$l_i(t) = (1-\rho)\, l_i(t-1) + \gamma f(x_i(t)) \tag{5}$$

其中：ρ 为荧光素消失率，γ 为荧光素更新率。

2) 概率选择

每个个体在其动态决策域半径 r_d^i 内，选择荧光素值比自己高的个体组成其邻域集 $N_i(t) = \{j: \| x_j(t) - x_i(t) \| < r_d^i; l_i(t) < l_j(t)\}$，其中 $(0 < r_d^i \leqslant r_s)$，$r_s$ 为萤火虫个体的感知半径。选择移向邻域集 $N_i(t)$ 内个体 j 的概率 $p_{ij}(t)$

$$p_{ij}(t) = \frac{l_j(t) - l_i(t)}{\sum_{k \in N_i(t)} l_k(t) - l_i(t)} \tag{6}$$

位置更新

$$x_i(t+1) = x_i(t) + s\left(\frac{x_j(t) - x_i(t)}{\| x_j(t) - x_i(t) \|}\right) \tag{7}$$

其中：s 为移动步长。

动态决策域半径更新

$$r_d^i(t+1) = \min\{r_s, \max\{0, r_d^i(t) + \beta(n_t - |N_i(t)|)\}\} \tag{8}$$

简单地说，GSO 算法主要包括萤火虫的初始分布、萤光素更新、萤火虫的移动和决策域的更新 4 个阶段[8]。

2.2　基于人工萤火虫的 FCM 算法

在人工萤火虫聚类中，设样本空间 $X = \{x_1, x_2, \cdots, x_n\}$，其中 x_i 为 d 维向量。用向量 $V = \{v_1, v_2, \cdots, v_c\}$ 来表示一个聚类中心，也就是一个萤火虫。其 v_j 是与 x_i 同维的向量。对于人工萤火虫中每个解（聚类中心）的评价，定义一个个体适应度函数：

$$f(x_i) = \frac{1}{1 + J_{FCM}(U, V)} \tag{9}$$

其中：$J_{FCM}(U, V)$ 为式（1）中定义的目标函数，聚类效果越好，$J_{FCM}(U, V)$ 越小，$f(x_i)$ 就越高。

人工萤火虫聚类算法 GSFM 主要思想：首先以 GSO 算法求得最优解（聚类中心）作为 FCM 算法的初始聚类中心，然后利用 FCM 算法优化初始聚类中心，最后求得最优解。此算法是一种快速聚类算法，所以该算法具有很好的时间性能，缩短了 FCM 算法的收敛时间，具体步骤描述为：

1) 给定类别数 c，群体规模 N，萤光素浓度 l_0，动态决策域初值 r_0，邻域阈值 n_t，移动的步长 s，参数 β, ρ, γ；

2) 根据式（4）随机初始化隶属度矩阵 U^0，作为初始的聚类划分；根据式（3）计算初始聚类中心，即产生初始解集 c_{ij}，然后根据式（9）计算各个解 c_{ij} 的适应度；

3) 根据式（5）将每只萤火虫 i 在 t 迭代的位置 $x_i(t)$ 对应的目标函数值 $f(x_i(t))$ 转化为荧光素值 $l_i(t)$；

4) 每只萤火虫 i 在其动态决策半径 $r_d^i(t)$ 内，选择荧光素值比自己高的个体组成其邻域集 $N_i(t)$；

5) 根据式（6）选出符合条件的萤火虫；

6) 进行移动，根据式（7）更新位置；

7) 按式（8）更新决策半径；

8) 根据式（4）更新隶属度矩阵 U；

9) 根据式（3）更新聚类中心。计算相邻两代隶属度矩阵之差 E，若 $E < \varepsilon$，停止；否则转到 8）。

3　实验结果

为了验证 GSFM 算法有效性和可行性，分别采用 UCI 标准数据库中的 IRIS 数据集、Wine 数据集和人造 KDD CUP99 数据集对算法进行测试，如表 1 所示。

表 1　实验样本数据集的组成

数据集名称	样本个数	类	维数
IRIS	150	3	4
人造 KDD	961	2	7
Wine	178	3	13

其中，IRIS 数据是提取 3 种不同类别花的花瓣和花萼的特征所构成的特征向量。共 150 个样本，每一个样本的 4 个分量分别表示 IRIS 的 Petal Length（花萼的长度），Petal Width（花萼的宽度），Sepal Length（花瓣的长度）和 Sepal Width（花瓣的宽度）。整个样本集包含了 3 个 IRIS 种类：Setosa，Versicolor 和 Virginica，每类各有 50 个样本。其中 Setosa 与其他两类间较好地分离，即为线性的，而 Versicolor 和 Virginica 之间存在交迭，则为非线性的。

从 KDD CUP99 的训练样本集中随机抽取离散属性维构成实验所需的人造样本集，类 normal（正常数据）占 14%；类 abnormal（异常数据）占 86%。与

文献[9]一样，所选的 7 维离散属性中 3 维为分类属性，相应的值域为{tcp，udp，icmp}、{SF，S0，REJ，RSTR，RSTO}、{http，ftp，ftp_data，smtp，pop_3，telnet，private，ecr_I，domain_u，other，finger，auth，mtp，urp_I，imap4，eco_I，systat，exec，login，Z39_50，discard}，另外 4 维是二值属性，相应的值域是{0，1}。

Wine 数据集为产于意大利同一地区不同种植园的 3 种葡萄酒化学分析结果的样本，以 Alcohol，Ash，Malic acid，Alkalinity of ash，Magnesium，Total phenols，Flavanoids，Nonflavanoid phenols，Proanthocyaninsm，Color intensity，OD280/OD315 of diluted wines and Praline 等 13 个参数作为特征，分为 3 类，每类分别有 59、71、48 个共 178 个样本，数据为 178×13 维矩阵。

分别用 FCM 算法和 ABFM 算法对 IRIS 数据和人造 KDD CUP99 数据进行聚类分析。各算法中允许最小误差 $\varepsilon = 10^{-3}$，模糊加权指数 $m = 2$。GSFM 算法的参数设置为：种群规模 $SN = 20$，最大迭代次数 $iter\_\max = 50$，$Limit = 100$。荧光素初始值 $l_0 = 5$，动态决策域初值 $r_0 = 5$，动态决策域的更新率 $\beta = 0.08$，邻域个数阈值 $n_t = 5$，步长 $s = 0.08$，荧光素更新率 $\gamma = 0.6$，荧光素消失率 $\rho = 0.4$。将两种算法分别运行 20 次，实验结果如表 2、表 3 和表 4 所示。

表 2 IRIS 数据的聚类结果

算法	平均错分数			平均准确率/%	迭代次数
	Setosa	Versicolo	Virginica		
GSFM	0/50	2/50	3/50	97.33	10
FCM	0/50	6/50	10/50	89.33	15

表 3 KDD CUP99 数据的聚类结果

算法	平均错分数		平均准确率/%	迭代次数
	类(normal)	类(abnormal)		
GSFM	13/143	9/818	97.71	14
FCM	32/143	19/818	94.69	23

表 4 Wine 数据的聚类结果

算法	平均错分数			平均准确率/%	迭代次数
	类 1	类 2	类 3		
GSFM	3/59	4/71	3/48	95.33	16
FCM	5/59	9/71	3/48	90.45	22

从表 2、表 3 和表 4 的实验结果可以看出，GSFM 算法优于传统的 FCM 算法，不仅分类准确率提高了，而且迭代次数更少，收敛速度加快，聚类效果更好。

4 结论

该文提出的 GSFM 算法将人工萤火虫算法与 FCM 算法结合，先以 GSO 算法求得近似最优解，将其作为 FCM 算法的初始值，继续进行局部搜索，就可以求得全局最优解。该算法克服了 FCM 算法易陷入局部最优解的不足，弥补了 FCM 算法对初始值和噪声点比较敏感的缺陷。实验表明，GSFM 算法具有很强的全局搜索能力，同时也使 FCM 算法对初始划分不再敏感，加快了收敛速度，聚类效果得到明显改善。

参考文献：

[1] Dunn J C. A fuzzy relative of the ISODATA process and it is use in detecting compact well separated cluster [J]. J Cybernet，1974(3)：32－57.

[2] Bezdek J. Pattern recognition with fuzzy objective function algorithms [M]. New York：Plenum，1981.

[3] 周涓，熊中阳，张玉芳，等. 基于最大最小距离法的多中心聚类算法 [J]. 计算机应用，2006(6)：1425－1427.

[4] 殷晓明，顾幸生. 一种基于改进型遗传算法的模糊聚类 [J]. 华东理工大学学报：自然科学版，2006(7)：849－851.

[5] 王磊，吉欢，徐庆征. 基于人工免疫粒子群优化算法的动态聚类分析 [J]. 西安理工大学学报，2008，24(4)：390－394.

[6] 李金霞，宋淑娜，胡学坤，等. 基于 APSO 的模糊聚类算法 [J]. 科学技术与工程，2009，9(19)：5696－5699.

[7] Krishnanand K N D，Ghose D. Glowworm swarm optimization：a new method for optimizing multi－modal functions [J]. Computational Intelligence Studies，2009，1(1)：93－119.

[8] Krishnanand K N. Glowworm swarm optimization：a multimodal function optimization paradigm with applications to multiple signal source localization tasks [D]. Indian：Department of Aerospace Engineering，Indian Institute of Science，2007.

[9] 曾国泰. 萤火虫最佳化分群演算法 [D]. 台北：台湾大同大学资讯经营研究所，2008.