基于智能优化算法的模糊软子空间聚类方法

张恒巍 何嘉婧 韩继红 王晋东
(解放军信息工程大学 郑州 450001)

摘 要 为解决选定特征上的聚类问题和模糊C-均值聚类存在的初始值敏感、易陷入局部最优的问题,提出了一种基于改进萤火虫算法的模糊软子空间聚类方法。该方法在模糊C-均值聚类算法的基础上,采用基于数据可靠性的k-均值算法中特征权值的计算方法,并结合萤火虫算法的全局搜索能力对所有的特征子空间进行搜索;设计了一种目标函数来对聚类结果和子空间所包含的特征维进行评估,并利用目标函数改进了萤火虫算法的搜索公式。实验结果表明,该方法能有效地收敛于全局最优解,具有良好的聚类效果和抗噪性。

关键词 聚类分析,子空间聚类,模糊C-均值,萤火虫算法,特征权值
中图法分类号 TP181 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.3.047

Fuzzy Soft Subspace Clustering Method Based on Intelligent Optimization Algorithm

ZHANG Heng-wei HE Jia-jing HAN Ji-hong WANG Jin-dong
(PLA Information Engineering University, Zhengzhou 450001, China)

Abstract To solve the issue of clustering on selected characteristics and the problems that fuzzy C-means is sensitive to initial value and easy to fall into local optimum, a new fuzzy subspace clustering method based on improved firefly algorithm was proposed. Based on fuzzy C-means clustering algorithm, the method uses the way to calculate feature weighting in reliability-based k-means algorithm, and combines with the global search capability of firefly algorithm to search for all the subspace. An objective function was designed to evaluate the clustering results and feature-dimension included in subspace, and it was adopted to improve the searching formula of firefly algorithm. Experimental results show that the proposed clustering method can effectively converge to the global optimal solution, and has good clustering effect and noise immunity.

Keywords Clustering analysis, Subspace clustering, Fuzzy C-mean, Firefly algorithm, Feature weighting

1 引言

随着信息技术的发展和信息服务的普及,众多应用领域的数据呈现向高维和大规模方向发展的趋势。作为大数据时代分析数据背后隐藏规律的有效工具,数据挖掘受到众多研究者的关注和重视。聚类分析是数据挖掘领域的关键技术和重要研究方向,其目的在于将数据集合中的对象聚合成不同的簇,并使簇中的对象具有较高的相似度,簇间则差异度较大[1]。聚类分析可以提取不同对象的相似性并将其进行汇聚,有助于发现和分析同类事物间存在的规律。聚类算法在众多研究领域都被广泛使用,如图像处理、资源调度、数据分析、效能评估、机器学习等。

按照聚类模式的不同,聚类算法主要包括硬聚类和软聚类算法。其中,硬聚类规定一个对象只能属于严格划分下的一个类,而软聚类则允许一个对象在不同程度上分属于一个或几个类[2],由于能够更好地描述实际情况,软聚类已成为目前的研究重点并被广泛应用。模糊C-均值聚类算法(Fuzzy C-Mean Clustering Algorithm, FCM)是基于目标函数的模糊聚类算法,属于软聚类算法,具备收敛速度快、局部搜索能力强的优点,但是由于聚类中心选取的随机性,导致算法对初始值和噪声数据敏感,全局优化能力差。近年来,不同研究者利用各种群智能算法对其进行改进,例如将遗传算法(Genetic Algorithm, GA)[3]、蚁群算法(Ant-colony Algorithm)[5]、粒子群算法(Particle Swarm Optimization, PSO)[5]、人工蜂群算法[6](Artificial Bee Colony Algorithm, ABC)等与FCM相结合,提高了聚类结果的稳定性和准确性。剑桥大学的Yang Xin-She教授于2009年提出萤火虫算法[7](Firefly Algorithm, FA),与其它群智能算法相比,该算法具有控制参数少、全局寻优能力强、收敛速度快和精度高的优点,适用于聚类方法的设计。但是,FA存在局部最优值或全局最优值振荡的缺陷[8]。

在对高维数据或数据集中的多个特征进行聚类时,传统算法一般利用全空间聚类,会影响聚类准确性并增加计算量。子空间聚类技术[9]能有效减少冗余和非相关属性对聚类过程

到稿日期:2015-08-28 返修日期:2015-11-08 本文受国家自然科学基金项目(61303074, 61309013),国家重点基础研究发展计划("973"计划)基金项目(2012CB315900),河南省科技计划项目(12210231003, 13210231002)资助。

张恒巍(1978—),男,博士,讲师,主要研究方向为云资源动态管理、网络与信息安全,E-mail:zhw11qd@126.com;**何嘉婧**(1991—),女,硕士生,主要研究方向为服务计算;**韩继红**(1966—),女,博士,教授,博士生导师,主要研究方向为资源管理、网络协议分析;**王晋东**(1966—),男,教授,主要研究方向为资源动态管理、信息安全分析。

的扰乱，有效避免上述问题的影响。其中，模糊子空间聚类(Fuzzy Subspace Clustering，FSC)[10]是典型算法之一，JING等人[11]采用熵加权法改进了文献[10]的目标函数，提出了熵加权 k-均值算法 EWKM。Wang 等人[12]引入功率指数，提出了一种改进模糊子空间聚类算法 DI-FSC。此外，Zhu 等人[13]借鉴在线学习和可扩展聚类技术对软子空间聚类方法进行改进，提出了在线软子空间聚类和数据流软子空间聚类算法。Hu 等人[14]提出基于多目标进化的软子空间聚类算法 MOE-ASSC。Boongoen 等人[15]基于数据可靠性理论，提出 Filter 软子空间聚类技术，与 FSC、EWKM 等 Wrapper 聚类算法相比，其适应性和准确性更好[16]。

本文首先借鉴 RKM 算法中特征权值的计算方法，在聚类过程中引入特征权值；然后对萤火虫算法进行改进，提高其自适应能力；在此基础上，将改进的萤火虫算法和 FCM 算法相结合，提出一种新的聚类方法，有效地提高了聚类能力。实验表明，相比其它算法，本文算法具有更好的聚类效果和抗噪能力。

2 相关工作

2.1 基于数据可靠性的 k-均值算法

作为 Filter 软子空间聚类算法，基于数据可靠性的 k-均值算法(Reliability-based KM，RKM)通过引入数据可靠性提升了聚类效果，适合于不同类型的聚类问题。

令 $a\in\{1,\cdots,(n-1)\}$ 代表样本最近邻居的数量，N_{jk}^{a} 为样本 j 在属性 k 上 a 个近邻的集合，“样本-属性”相关度矩阵 $AS^{a}\in R^{n\times d}$，$AS_{jk}^{a}\in[0,1]$ 代表样本 x_j 与近邻集合 N_{jk}^{a} 在属性 k 上的相关度。AS_{jk}^{a} 的计算公式如下：

$$AS_{jk}^{a}=1-\frac{D_{jk}^{a}}{\max\limits_{\forall j,k}D_{jk}^{a}},D_{jk}^{a}=\frac{1}{a}\sum_{q\in N_{jk}^{a}}\sqrt{(x_{jk}-q_k)^2} \tag{1}$$

其中，D_{jk}^{a} 代表在属性 k 上样本 x_j 与 a 个近邻的平均距离，D_{jk}^{a} 越小，AS_{jk}^{a} 越接近于 1，样本 x_j 所在的簇与属性 k 的相关性越高；D_{jk}^{a} 越大，AS_{jk}^{a} 越接近于 0，样本 x_j 所在的簇与属性 k 的相关性越低。对任意给定的 k、j 和 a，AS_{jk}^{a} 仅与样本有关，与聚类中心以及隶属度无关。

$V=[v_{ik}]_{C\times D}$ 代表聚类中心矩阵，矩阵元素为各聚类中心的坐标；$U=[u_{ij}]_{C\times N}$ 代表隶属度矩阵，矩阵元素为样本对各簇的隶属度值；$W=[w_{ik}]_{C\times D}$ 代表权值矩阵，矩阵元素为属性对各簇的权值。其中，N 是数据点数量，C 是聚类的簇数，D 是样本维数，即属性数目。RKM 的目标函数如下：

$$J=\sum_{i=1}^{C}\sum_{k=1}^{D}\sum_{j=1}^{N}u_{ij}w_{ik}(x_{jk}-v_{ik})^2 \tag{2}$$

在初始化阶段，聚类中心为随机选取的 C 个样本 $V=\{v_1,v_2,\cdots,v_C\}$，则簇 i 中的特征权值定义为：

$$w_{ik}=\frac{AS_{jk}^{a}}{\sum\limits_{l=1}^{D}AS_{jl}^{a}} \tag{3}$$

在算法的迭代计算过程中，特征权值 w_{ik} 的更新公式为：

$$w_{ik}=\frac{MA_{jk}^{a}}{\sum\limits_{l=1}^{D}MA_{jl}^{a}} \tag{4}$$

其中 $MA_{ik}^{a}=\min\limits_{q\in C_i}AS_{qk}^{a}$。

基于式(3)、式(4)，定义划分矩阵的更新公式为：

$$u_{ij}=\begin{cases}1, & i=\arg\min\limits_{q=1,\cdots,C}\sum\limits_{k=1}^{D}w_{qk}(x_{jk}-v_{qk})^2\\ 0, & \text{otherwise}\end{cases} \tag{5}$$

与 KM 算法相似，聚类中心矩阵 V 的更新公式为：

$$v_{ik}=\frac{\sum\limits_{i=1}^{N}u_{ij}x_{jk}}{\sum\limits_{j=1}^{N}u_{ij}} \tag{6}$$

RKM 算法的一般流程为：

(1)采用随机方式初始化聚类中心矩阵 V；

(2)利用式(3)计算簇 i 中的初始特征权值；

Repeat

(3)利用式(5)更新隶属度矩阵 U；

(4)利用式(6)更新聚类中心矩阵 V；

(5)利用式(4)更新特征权值矩阵 W；

Until 目标函数 J 达到最小值

RKM 算法在权值更新上不需要依赖聚类中心，能够有效提升聚类的准确性。

2.2 模糊 C-均值算法

FCM 算法基于最小二乘法原理，采用迭代过程不断逼近目标函数的最小值。

定义数据集为 $X=\{X_1,X_2,\cdots,X_n\}$，其中 $X_i=\{X_{i1},X_{i2},\cdots,X_{in}\}$。用目标函数 J 表征划分类的紧密程度：

$$J(U,V)=\sum_{i=1}^{n}\sum_{j=1}^{c}\mu_{ij}^{m}\parallel x_i-v_j\parallel^2 \tag{7}$$

其中，$m\in[1,+\infty)$ 是模糊指数，控制不同类隶属度的权重，在迭代计算中隶属度和聚类中心的求解公式分别如式(8)、式(9)所示：

$$\mu_{ij}=(\sum_{k=1}^{c}\frac{\parallel x_i-v_j\parallel^{2/(m-1)}}{\parallel x_i-v_k\parallel^{2/(m-1)}})^{-1} \tag{8}$$

$$v_j=\frac{\sum\limits_{i=1}^{n}(\mu_{ij})^m x_i}{\sum\limits_{i=1}^{n}(\mu_{ij})^m} \tag{9}$$

FCM 算法的基本步骤是：

(1)设定初始参数，令迭代次数 $t=1$，选择聚类中心的个数 C 以及模糊指数 m。

(2)初始化聚类中心 V，用式(8)计算隶属度 $U=[\mu_{ij}]$。

(3)用式(9)更新聚类中心 $V=\{v_1,v_2,\cdots,v_c\}$。

(4)若 $\max|\mu_{ij}^{k}-\mu_{ij}^{k-1}|\leqslant\varepsilon$，则算法收敛；若迭代次数大于 T，则算法结束。否则，令 $t=t+1$，转到步骤 2。其中 ε 是预先设定的阈值。

模糊 C-均值聚类算法是无监督聚类方法，其迭代计算利用梯度下降法搜索最优解，具有易于实现、收敛速度快、应用范围广且局部搜索能力较强的优点。但是，算法对初始值条件与噪声数据敏感，对不同初始值常会有不同的聚类效果，且容易陷入局部最优，全局寻优能力较差。

2.3 萤火虫算法

萤火虫算法是一种启发式算法[17]，其根据萤火虫利用闪光来捕食、警戒或吸引异性的原理进行设计，算法假设：

(1)不分性别，一个萤火虫发光将会吸引所有的萤火虫。

(2)吸引力与发光亮度成正比，小亮度萤火虫被大亮度萤火虫吸引，并向大亮度萤火虫方向移动。最亮的萤火虫随机移动。

(3)随着距离的增大亮度会减小。

(4)具体应用中，亮度与给定的目标函数相关。

FA 算法采用自身亮度和吸引度作为寻优要素。位置和亮度具有正相关性。萤火虫的吸引力受亮度影响，由于亮度

跟随萤火虫 i 与 j 的距离变化而变化，因此吸引度也相应改变。

亮度 L 与吸引度 β 的计算公式如下：

$$L=L_0\times e^{-\gamma r_{ij}} \tag{10}$$

$$\beta=\beta_0\times e^{-\gamma r_{ij}^2} \tag{11}$$

其中，L_0 是萤火虫自身($r=0$ 处)的荧光亮度，即亮度最大值。β_0 为光源处($r=0$ 处)的吸引度，即吸引度的最大值。γ 为光强吸收系数，β_0 和 γ 均为常数。r_{ij} 代表萤火虫 i 和 j 之间的距离，亮度 L 和吸引度 β 随着 r_{ij} 的增加而减弱。

$$r_{ij}=\| x_i-x_j \|=\sqrt{\sum_{k=1}^{d}(x_{ik}-x_{jk})^2} \tag{12}$$

当萤火虫 i 被 j 吸引时，萤火虫 i 的位移公式为：

$$x_i^{t+1}=x_i^t+\beta(x_j^t-x_i^t)+s(\delta-0.5) \tag{13}$$

其中，x_i、x_j 分别代表萤火虫 i 和 j 的位置；s 是步长因子，为[0,1]上的常数；δ 为随机因子，在[0,1]上均匀分布。

亮度最高的萤火虫随机移动，其位移公式为：

$$x_{best}^{t+1}=x_{best}^t+s(\delta-0.5) \tag{14}$$

萤火虫算法的参数较少，易于实现，具备良好的全局寻优能力和一定的局部寻优能力，收敛到最优解的速度快，在性能上优于很多经典的群智能算法[18]。

3 基于改进萤火虫算法的软子空间聚类方法

3.1 目标函数的设计

目标函数是算法进行结果评价的标准，其设计直接影响算法的搜索方向和收敛程度。具体到聚类问题，设计的准则是使聚类后的簇具有最大的独立性和紧凑性，即尽量最大化簇内相似度和簇间差异度。在实际问题求解中，有时只需对某些特征维进行聚类，为针对所需特征维获取目标簇以及分析相关特征维在子空间聚类中的性能，本文引入贡献率 w_{ik} 作为特征权值，用以描述特征维对子空间聚类的作用程度。

由于模糊聚类将数据对象归属于不同类的不确定性定义为模糊隶属度，比硬聚类更好地描述了现实问题，因此本文借鉴 RKM 和 FCM 算法，设计了一种新的目标函数，并在此基础上提出了模糊子空间聚类算法 MFARCM。

目标函数 J 为：

$$J=\sum_{i=1}^{C}\sum_{k=1}^{D}w_{ik}\sum_{j=1}^{N}u_{ij}^m(x_{jk}-v_{ik})^2 \tag{15}$$

其中，w_{ik} 为特征权值，代表特征维的贡献率。m 为 FCM 算法中的模糊指数且 $m>1$，用于控制算法对划分矩阵 U 的敏感度。当参数 $m\to1$ 时，式(15)与 RKM 算法的目标函数一致；当 $m\to\infty$ 时，样本对聚类中心的隶属度趋于相同。

3.2 萤火虫的编码与算法改进

在本文的 MFARCM 算法中，假设期望聚类数为 C，用一组聚类中心表示萤火虫 i 的位置，则萤火虫的位置为向量 $V=(v_1,v_2,\cdots,v_C)$，其中 v_i 代表第 i 个聚类中心。萤火虫 i 采用如下编码：

$$\{x_{11},x_{12},\cdots,x_{1D},\cdots,x_{C1},x_{C2},\cdots,x_{CD}\}$$

萤火虫 i 的 D 维向量由聚类中心 $v_i=\{x_{i1},x_{i2},\cdots,x_{iD}\}$ 中的元素 x_{iD} 组成。

萤火虫的亮度 L 用目标函数 J 定义，表示为：

$$L=\frac{1}{1+J} \tag{16}$$

由式(16)可知，若目标函数 J 的值越小，则萤火虫亮度 L 越大。聚类目标是最小化目标函数，选取亮度达到 L_{max} 的萤火虫为最优聚类中心。

根据式(13)可知，基本 FA 算法中，萤火虫 i 的移动方向取决于亮度 L，位移的大小取决于吸引度 β，由于 β_0 和 γ 均为常数，β 的大小仅决定于萤火虫之间的距离 r_{ij}。因此，位移大小 x_i^{t+1} 也仅决定于 r_{ij}。此时，萤火虫 i 的位移中只有方向受亮度 L 的影响，而位移大小与亮度 L 无关。

通过分析可知，针对距离为 r 的萤火虫 i 和 j，当亮度分别处于 $L_i<L_j$ 和 $L_i\ll L_j$ 两种情况时，由于距离相同，萤火虫 i 向 j 移动的位移大小是一样的。但是，两种情况下萤火虫 i 和 j 的亮度的差异程度很大，显然萤火虫 i 的位移在这两种情况下应该不同。

基于以上思路，对代表萤火虫 i 向 j 移动的位移公式(13)做如下修改：

$$x_i^{t+1}=x_i^t+\beta\frac{L_j^t-L_i^t}{L_{max}^t-L_{min}^t}(x_j^t-x_i^t)+s(\delta-0.5) \tag{17}$$

其中，L_i^t 和 L_j^t 分别代表第 t 代萤火虫 i 和 j 的亮度，L_{max}^t 和 L_{min}^t 分别代表第 t 代最亮和最暗萤火虫的亮度。修改后的位移公式可以描述亮度对萤火虫位移大小的影响，能够更准确地刻画萤火虫的位移过程，增强了算法的全局搜索能力，第 4 节的实验对此进行了验证。

式(17)中的第三部分 $s(\delta-0.5)$ 代表局部随机搜索移动，展现的是算法的局部寻优能力，其中步长因子 s 用于控制算法的随机性。在寻优过程中，步长设置关系到算法会不会陷入局部最优。s 取值较高时，算法可以较快地跳出局部最优，但是可能会降低算法求解精度；s 取值较低时，能够减弱最优解附近发生的振荡，提高局部搜索能力。

因此，本文采用自适应性步长来平衡算法的全局和局部搜索能力，改进如下：

$$s=s_0e^{-t} \tag{18}$$

可以看出，在优化过程开始时，s 取值较高，个体保持较快的寻优速度与较强的探索能力，避免了算法早熟；在优化过程的后期，随机步长逐渐变小，个体局部搜索能力逐渐增强，倾向于在自身周围搜索精确解，提高了局部搜索能力并抑制了振荡。自适应步长能够有效提升聚类算法的性能，第 4 节的实验对此进行了验证。

3.3 MFARCM 算法的步骤

MFARCM 算法的聚类过程就是将式(15)中的目标函数最小化的过程，即最小化类内模糊加权距离。

基于改进萤火虫算法更新聚类中心矩阵 V，萤火虫种群代表类中心，萤火虫位置表示聚类中心，亮度表示目标函数的值，萤火虫的吸引-移动过程表示聚类的优化过程，最优聚类由亮度最高的萤火虫代表。

采用 FCM 算法中的隶属度更新方法，由拉格朗日定理得到划分矩阵 U 的更新公式：

$$u_{ij}=\frac{(d_{ij})^{-1/(m-1)}}{\sum_{l=1}^{C}(d_{lj})^{-1/(m-1)}} \tag{19}$$

$$d_{ij}=\sum_{k=1}^{D}w_{ik}(x_{jk}-v_{ik})^2 \tag{20}$$

采取 RKM 中的方式更新权值矩阵 W，初始化阶段的计算如式(3)所示，迭代阶段的更新计算如式(4)所示。

基于改进萤火虫算法的软子空间聚类方法的流程如图1所示。

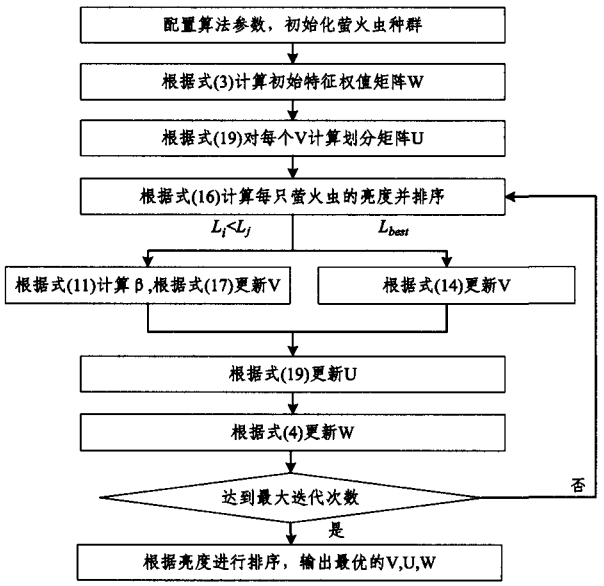

图1 MFARCM算法流程图

在初始化阶段，设置萤火虫的数量为 n，最大吸引度 β_0，吸收因子 γ，步长因子初始参数 s_0，聚类目标簇数 C，最大迭代次数 $maxT$，以及 δ 和 m。初始化萤火虫的位置，在数据集中随机挑选 C 个位置作为初始位置，代表初始聚类中心集。使用RKM的公式计算权值矩阵的初始值 W，保证权值不受初始聚类中心的影响。计算划分矩阵 U 时，采用更能反映现实情况的FCM算法中的模糊划分方法。

在初始聚类的基础上，采用式(16)计算个体的亮度，选取并记录亮度最大的个体的位置、亮度和聚类结果。移动的方向指向亮度最大的个体，并根据式(18)设置个体的步长 s。如果 $L_i<L_j$，表示个体 j 位置较好，按照式(11)计算个体 j 对 i 的吸引度 β_{ij}，按照式(17)更新个体 i 的位置。如果更新后个体 i 的状态优于先前状态，则执行吸引-移动行为；否则，重新选择亮度次大的个体开展追逐。亮度最大的个体则采用式(14)更新位置。

个体位置更新完成后，根据式(19)更新划分矩阵，根据式(4)更新权值矩阵，并重新选取和记录亮度最大个体的位置、亮度和聚类结果。若迭代过程达到 $maxT$，则终止算法。输出最优个体的位置，即代表聚类中心的 V，以及相应的权值矩阵 W 和划分矩阵 U；否则，对个体的位置再次更新，进入下一轮迭代。

4 实验对比及结果分析

为了分析与验证本文提出的基于改进萤火虫算法的子空间聚类算法(MFARCM)的性能，利用UCI机器学习数据库[19]中的Iris、Vehicle两个数据集以及生理医学数据库癌基因表达数据集[20]中的CNS、Leukemia-MLL和Lung harvard 3个数据集进行测试。前两个数据集用于测试算法对低维数据的兼容性，后3个数据集用于测试算法在高维数据中的应用性能，并与标准FCM算法、基于萤火虫算法的模糊聚类算法(FAFCM)[21]、RKM算法、EWKM算法及文献[16]中的DESC算法进行比较。

Iris数据集含3类样本，样本数量为150，每个样本含4个属性。Vehicle数据集含4类样本，样本数量为846，每个样本含18个属性。CNS数据集含2类样本，样本数量为34，每个样本含7129个属性。Leukemia-MLL数据集含3类样本，样本数量为72，每个样本含12582个属性。Lung harvard数据集含5类样本，样本数量为203，每个样本包含12600个属性。对数据集的数据进行线性归一化预处理，使每一维数据在[0,1]内取值。

测试的硬件环境为Intel Core TM i3-2100CPU，4GB内存，所有算法程序在WEKA平台[22]下开发。为了衡量聚类结果，在实验中主要比较RI(Rand Index)和NMI(Normalized Mutual Information)[23]两个指标。

$$RI=\frac{f_{00}+f_{11}}{N(N-1)/2} \tag{21}$$

其中，f_{00} 是不同类且不同簇的样本点数量，f_{11} 是同类且同簇的样本点数量，N 是样本的总数量。

$$NMI=\frac{\sum_{i=1}^{K}\sum_{j=1}^{C}n_{ij}\log\frac{N\cdot n_{ij}}{n_i\cdot n_j}}{\sqrt{(\sum_{i=1}^{K}n_i\log\frac{n_i}{N})(\sum_{j=1}^{C}n_j\log\frac{n_j}{N})}} \tag{22}$$

其中，K 是类别数，N 是样本的总数量，n_i 和 n_j 分别是属于类 i 和簇 j 的样本数，n_{ij} 是属于类 i 且属于簇 j 的样本数。RI和NMI指标都是越大越好。

算法的参数设置如下：最大迭代次数 $maxT=200$；对于FCM算法，$m=2$；对于FAFCM算法，$\beta_0=1.0$，$\gamma=0.9$，$s=0.1$，$m=2$；对于RKM算法，$a=3$；对于EWKM算法，$\gamma=0.5$；对于DESC算法，$m=2$，$\tau=\eta=2$；对于本文算法，$\beta_0=1.0$，$\gamma=0.9$，$s_0=0.1$，$m=2$，$a=3$。6个测试算法分别运行20次，取最好结果，如表1、表2、图2和图3所示。

表1 算法有效性指标(RI)比较

Dataset	FCM	FAFCM	RKM	EWKM	DESC	本文算法	
Vehicle	Mean	0.6493	0.6525	0.6482	0.3886	0.6476	**0.6675**
	Std	0.0072	**0.0036**	0.0091	0.1242	0.0162	0.0087
Iris	Mean	0.8332	0.8423	0.8607	0.8773	**0.9423**	0.8853
	Std	0.0642	0.0653	0.0622	0.0964	**0.0180**	0.0324
Lung harvard	Mean	0.5002	0.5177	0.5633	0.5771	0.5724	**0.5782**
	Std	0.0142	**0.0126**	0.0233	0.0296	0.0234	0.0135
Leukemia-MLL	Mean	0.7024	0.7375	0.7241	0.6196	0.7534	**0.7785**
	Std	0.0923	0.0914	0.0536	0.1107	**0.0173**	0.0231
CNS	Mean	0.5598	0.5764	0.6147	**0.7051**	0.5476	0.5682
	Std	0.0867	0.0812	0.1141	0.1060	0.0787	**0.0655**

表 2　算法有效性指标(NMI)比较

Dataset		FCM	FAFCM	RKM	EWKM	DESC	本文算法
Vehicle	Mean	0.1047	0.1343	0.1310	0.1039	0.1382	**0.1433**
	Std	0.0232	**0.0110**	0.0354	0.0528	0.0306	0.0206
Iris	Mean	0.7100	0.7312	0.7287	0.7828	**0.8529**	0.7908
	Std	0.0651	0.0649	0.0671	0.1172	**0.0319**	0.0537
Lung harvard	Mean	0.1572	0.1694	0.2696	0.2715	0.2840	**0.3001**
	Std	0.0289	**0.0263**	0.0699	0.0835	0.0536	0.0412
Leukemia-MLL	Mean	0.4107	0.4476	0.4520	0.3093	0.4674	**0.4876**
	Std	0.1256	0.1223	0.1125	0.1637	**0.0354**	0.0423
CNS	Mean	0.0144	0.0154	0.1283	0.1351	0.1440	**0.1512**
	Std	0.0211	**0.0157**	0.1360	0.2007	0.0955	0.0876

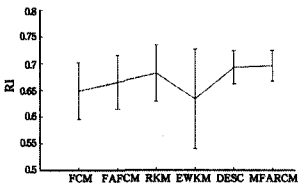

图 2　RI 值及标准差的平均数比较

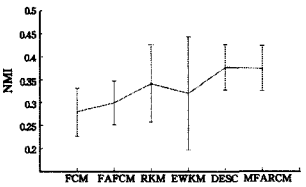

图 3　NMI 值及标准差的平均数比较

可以看出，在 CNS 数据集上，本文算法的 RI 值仅次于 EWKM。在 Iris 数据集上，本文算法的性能仅次于 DESC。其他数据集上，本文算法的 RI 值和 NMI 值均为最大，聚类效果最好，准确性最高。同时，本文算法标准差在各个数据集上均较小，具有更好的稳定性。

为了比较 6 种算法抗噪声数据的能力，在原数据集中加入 10%、20%、30%和 40%的噪声数据，然后分别进行实验，结果如图 4 所示。

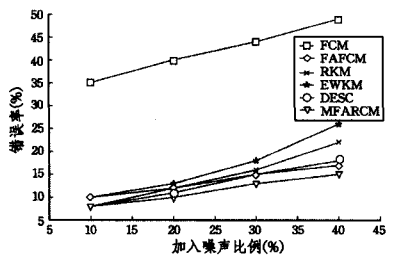

图 4　6 种算法的抗噪性比较

可以看出，FCM 算法的抗噪性最差，本文算法的抗噪性能最好，尤其随着加入的噪声比例的增加，其抗噪性能的优势更加明显。

6 种算法的运行时间如表 3 所列。

表 3　算法运行 20 次的平均时间(单位：s)

算法	Iris	Vehicle	CNS	Leukemia MLL	Lung harvard
FCM	0.0492	1.0034	7.0204	42.7653	212.0097
FAFCM	0.0573	1.2347	7.9593	45.3545	213.2230
RKM	0.0267	0.4973	5.1464	18.7152	79.3096
EWKM	0.0432	0.9591	12.6785	51.6876	199.5557
DESC	0.0661	0.9793	9.7577	53.1457	205.1150
本文算法	0.0652	1.0256	9.3836	48.7243	204.0321

可以看出，在低维数据集上，RKM 算法的运行时间最短，DESC 和本文算法的运行时间最长。在高维数据集上，RKM 算法的运行时间仍最短，本文算法、EWKM 和 DESC 的运算速度依次递减。原因分析如下：FCM 的搜索过程采用最快梯度下降法，全局搜索能力较差，但收敛速度快；RKM 属于 KM 类方法，收敛速度比采用模糊聚类的 FCM 更快，在 6 种算法中运算速度最优，但是两种算法的全局搜索能力不强。本文算法和 DESC 的权值更新计算方式比 EWKM 简捷，因此速度相对更快。

综合分析上述实验结果可知，本文算法在 FCM 的基础上引入特征权值，能够解决对特定特征进行聚类的问题，并能更好地满足高维数据集的聚类要求。改进的萤火虫算法具有更好的全局搜索能力，有效地避免了陷入局部最优的问题，使本文方法在聚类的准确性和抗噪声能力方面优于 FCM、FAFCM 和 RKM 方法。与同样具有较好全局搜索能力的 EWKM 和 DESC 相比，本文方法的特征权值确定过程独立于聚类中心的计算，消除了初始值对聚类效果的影响，具有更好的聚类效果，准确性和稳定性更强。但是，由于本方法比较复杂，运算速度不如 FCM、FAFCM 和 RKM 方法，与 EWKM 和 DESC 方法接近。

结束语　本文针对实际应用中需要对特定特征进行聚类分析的问题，在 FCM 聚类算法的基础上引入了特征权值，设计了新的目标函数。同时，针对萤火虫算法容易在最优值附近反复振荡的问题，对其进行了自适应性改进，利用其具有较快的收敛速度和较好的全局搜索能力的优点，提出一种基于改进萤火虫算法的模糊子空间聚类方法。实验结果表明本文的方法具有更好的聚类效果和抗噪性，但存在运行时间较长的问题。下一步的工作是对算法简化及参数自动配置进行研究，提升算法的操作性。

参 考 文 献

[1] Li Xin, Zhang Ji-fu, Cai Jiang-hui. A fuzzy clustering algorithm based on large density region [J]. Journal of Chinese Computer Systems, 2015, 33(1): 1310-1315

[2] Kuo R J, Huang Y D, Lin C C, et al. Automatic kernel clustering with bee colony optimization algorithm [J]. Information Sciences, 2014, 283: 107-122

[3] Fu Ping, Luo Ke. Clustering Analysis of Immune-genetic Algorithm Based on Information Entropy [J]. Computer Engineering, 2008, 34(6): 227-228, 232(in Chinese)
傅平，罗可. 基于信息熵的免疫遗传算法聚类分析[J]. 计算机工程，2008，34(6)：227-228，232

[4] Wang Fei, Zhang De-xian, Bao Na. Fuzzy document clustering based on ant colony algorithm[C]// Proc of the 6th International Symposium on Neural Networks: Advances in Neural Networks-Part II. 2015: 709-716

[5] Yao Li-juan, Luo Ke, Meng Ying. Clustering algorithm based on particle swarm optimization[J]. Computer Engineering and Applications, 2013, 48(13): 150-153(in Chinese)
姚丽娟，罗可，孟颖. 一种基于粒子群的聚类算法[J]. 计算机工程与应用，2013，48(13)：150-153

[6] Karabogad, BasturkB. On the performance of artifi cial bee co-

lony (ABC) algorithm [J]. Applied Soft Computing, 2014, 8 (1): 687-697

[7] Yang Xin-she. Firefly algorithms for multimodal optimization [C]// Proc of the 5th International Conference on Stochastic Algorithms: Foundations and Applications. Berlin: Springer-Verlag, 2013: 169-178

[8] Senthilnath J, Omkar S N, Mani V. Clustering using firely algorithm: Performance study[J]. Swarm and Evolutionary Computation, 2014, 1(3): 164-171

[9] Chen Li-fei, Guo Gong-de, Jiang Qing-shan. Adaptive algorithm for soft subspace clustering [J]. Journal of Software, 2014, 21 (10): 2513-2523

[10] Gan Guo-jun, Wu Jian-hong, Yang Zi-jiang. A fuzzy subspace algorithm for clustering high dimensional data[M]// Advanced Data Mining and Applications. 2013: 271-278

[11] Jing Li-ping, Ng M K, Huang Zhe-xue. An Entropy Weighting K-means algorithm for subspace clustering of high-dimensional sparse data[J]. IEEE Transactions on Knowledge Data and Engineering, 2015, 19(2): 1026-1041

[12] Wang Jun, Chung Fu-lai, Wang Shi-tong. Double indices-induced FCM clustering and its integration with fuzzy subspace clustering[J]. Pattern Anal Applic, 2014, 17: 549-566

[13] Zhu Lin, Cao Long-bin, Yang Jie, et al. Evolving soft subspace clustering[J]. Applied Soft Computing, 2014, 14: 210-228

[14] Hu Xia, Zhuang Jian, Yu De-hong. Novel soft subspace clustering with multi-objective evolutionary approach for high-dimensional data[J]. Pattern Recognition, 2013, 46: 2562-2573

[15] Boongoen T, Shang Chang-jing, Natthakan I O, et al. Extending Data Reliability Measure to a Filter Approach for Soft Subspace Clustering[J]. IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics, 2012, 41(6): 1705-1714

[16] Bi Zhi-sheng, Wang Jia-hai, Yin Jian. Subspace Clustering Based on Differential Evolution [J]. Chinese Journal of Computers, 2015, 35(10): 2116-2128(in Chinese)
毕志升, 王甲海, 印鉴. 基于差分演化算法的软子空间聚类[J]. 计算机学报, 2015, 35(10): 2116-2128

[17] Yang Xin-she. Nature-inspired metaheuristic algorithms [M]. London: Luniver Press, 2015: 83-96

[18] Lee H S, Tzeng G H, Yeih W C, et al. Revised DEMATEL: resolving the infeasibility of DEMATEL [J]. Applied Mathematical Modelling, 2015, 37(5): 1-12

[19] UCI Database of UCLA University [EB/OL]. (2015-05-22) [2015-08-30]. http://www.UCLA.org/UCIdata

[20] UCLA Physiology and Medicine GeneData [EB/OL]. (2015-02-10) [2015-08-30]. http://datam.i2r.edu/datasets/krbd

[21] Lin Mu-gang, Liu Fang-ju, Tong Xiao-jiao. Fuzzy clustering algorithm based on firefly algorithm[J]. Computer Applications, 2014, 50(21): 35-38(in Chinese)
林睦纲, 刘芳菊, 童小娇. 一种基于萤火虫算法的模糊聚类方法[J]. 计算机应用, 2014, 50(21): 35-38

[22] Hall M, Frank E, Holmes G. The WEKA data mining software version 2.8 [EB/OL]. http://www.weka.an.za.net

[23] Liu J, Mohammed J, Carter J, et al. Distance-based clustering of CGH data[J]. Bioinformatics, 2013, 22(16): 1971-1978

(上接第 255 页)

参 考 文 献

[1] Linguistic Data Consortium. ACE (Automatic Content Extraction) Chinese Annotation Guidelines for Events Version 5.5.1. [OL]. (2009-09-08). http://www.ldc.upenn.edu/Projects/ACE

[2] Zhao Y Y, Qin B, Che W X, et al. Research on Chinese event extraction[J]. Journal of Chinese Information Processing, 2008, 22 (1): 3-8(in Chinese)
赵妍妍, 秦兵, 车万翔, 等. 中文事件抽取技术研究[J]. 中文信息学报, 2008, 22(1): 3-8

[3] Tan H Y. Research on Chinese event extraction [D]. Harbin: Harbin Institute of Technology, 2008(in Chinese)
谭红叶. 中文事件抽取关键技术研究[D]. 哈尔滨: 哈尔滨工业大学, 2008

[4] Zheng Chen, Heng Ji. Language specific issue and feature exploration in Chinese event extraction[C]// Proceeding of the 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics Boulder. Colorado, USA, 2009: 209-212

[5] Ahn D. The Stages of Event Extraction[C]// Proceedings of the Workshop on Annotations and Reasoning about Time and Events. 2006: 1-8

[6] Li Pei-feng, Zhou Guo-dong, Zhu Qiao-ming, et al. Employing Compositional Semantics and Discourse Consistency in Chinese Event Extraction[C]// Proc. EMNLP. 2012: 1006-1016

[7] Hou L B, Li P F, Zhu Q M. Study of Event Recognition Based on CRFs and Cross-event [J]. Computer Engineering, 2012, 38 (24): 191-195(in Chinese)
侯立斌, 李培峰, 朱巧明. 基于 CRFs 和跨事件的事件识别研究[J]. 计算机工程, 2012, 38(24): 191-195

[8] Fu Jian-feng, Liu Zong-tian, Zhong Zhao-man, et al. Chinese Event Extraction Based on Feature Weighting[J]. Information Technology Journal, 2010, 9(1): 184-187

[9] Li Pei-feng, Zhou Guo-dong, Zhu Qiao-ming. Argument Inference from Relevant Event Mentions in Chinese Argument Extraction[C]// Proceedings of ACL. 2013: 1477-1487

[10] Liao Sha-sha, Grishman R. Using Document Level Cross-Event Inference to Improve Event Extraction[C]// Porc. ACL 2010. Uppsala, Sweden, 2010: 789-797

[11] Hong Yu, Zhang Jian-feng, Ma Bin, et al. Using Cross-Entity Inference to Improve Event Extraction[C]// Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA, 2011: 1127-1136

[12] Richardson M, Domingos P. Hybrid markov logic networks[J]. Machine Learning, 2006, 62(1/2): 1106-1111

[13] Poon H, Domingos P. Joint inference in information extraction [C]// AAAI. 2007: 913-918

[14] Poon H, Pedro D. Joint unsupervised coreference resolution with Markov logic[C]// Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2008: 650-659

[15] Poon H, Pedro D. Unsupervised semantic parsing [C]// Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1. 2009: 1-10

[16] Singla, Parag, Pedro D. Entity resolution with markov logic [C]// Sixth International Conference on Data Mining, 2006 (ICDM'06). IEEE, 2006: 572-582

[17] Li Pei-feng, Zhou Guo-dong. Employing Morphological Structures and Sememes for Chinese Event Extraction[C]// COLING. 2012: 1619-1634