一种基于粒子群优化的可能性C均值聚类改进方法

陈东辉 刘志镜 王纵虎
(西安电子科技大学计算机学院 西安 710071)

摘 要 提出了一种基于粒子群优化的可能性C均值(Possibilistic C-means,PCM)聚类改进方法。该方法首先通过改进PCM算法的目标函数来计算数据模式的隶属度矩阵和聚类中心完成粒子编码,从而降低算法对初始中心的敏感,提高聚类的精度;其次,通过粒子群优化(Particle Swarm Optimization,PSO)算法对编码进行优化,以有效地克服PCM聚类算法容易导致聚类一致性和陷入局部最优解的缺点,减少算法的迭代次数。通过人造数据集和UCI数据集上的实验,表明该算法在计算复杂度、聚类精度和全局寻优能力方面表现得较为突出。

关键词 模糊聚类,粒子群优化,模糊C均值,可能性C均值
中图法分类号 TP18 **文献标识码** A

Improved Possibilistic C-means Clustering Algorithm Based on Particle Swarm Optimization

CHEN Dong-hui LIU Zhi-jing WANG Zong-hu
(Department of Computer Science and Technology, Xidian University, Xi'an 710071, China)

Abstract An improved possibilistic C-means(PCM) algorithm based on particle swarm optimization (PSO) was presented. This algorithm consists of two steps: first, using the improved PCM to calculate the degree of membership matrix and cluster centroid to encode particles, which can low the influence of initialized centroid and improve clustering precision. In the second, using PSO to optimize the encoded data points, which can overcome the coincident clusters and avoid easily falling into local optimum. The experimental results on the synthetic data sets and UCI data sets show that the proposed algorithm has less computational complexity, higher clustering precision and greater searching capability.

Keywords Fuzzy clustering, Particle swarm optimization, Fuzzy C-means clustering, Possibilistic C-means clustering

1 引言

聚类分析是通过将有限未标记数据集分成有限离散"自然"数据集来发现数据的结构信息,是统计模式分类中无监督分类的一个重要分支。聚类分析是按照某个特定标准把一个数据集分成若干个不同的子类,使得在同一类内的样本相似性尽可能的大,不同类的样本差异性也尽可能地大[1]。换句话说,聚类后同一类别的数据样本尽可能的聚集在一起,而不同的样本尽量分离。聚类方法按照隶属度的取值范围大体上分为硬聚类方法和模糊聚类方法。

硬聚类方法是一种硬划分,它将每个待辨识的数据样本对各个类的隶属度取值0和1两种,值为0表示该样本不属于这一类,值为1表示该样本属于这一类。传统的硬聚类算法大致分为两大类:启发式和划分式。启发式方法将数据进行树状分类,常常给出数据的几种可能的分类情况;划分式则不同,它将数据按照某种标准划分成单一的结果。划分技术包括目标函数法(平方误差)、密度估计(模型搜寻)、图结构和最近邻法。硬聚类算法具有花费时间少的优点,但是硬聚类割裂了样本与样本之间的联系,无法准确地表达它们在性态和类属方面存在着的中介性,容易陷入局部最优解,使所得的聚类结果误差偏大。

模糊聚类算法[2]扩展了隶属度的取值范围,具有更好的聚类效果与数据表达能力。该类方法是基于Zadeh教授在1965年提出的模糊理论,它是模糊理论与聚类分析相结合的产物。模糊聚类方法能够对类与类之间有交叉的数据样本集进行有效的聚类,所得的聚类结果明显优于硬聚类方法。模糊聚类由于建立起了数据样本对于类别的不确定性的描述,表达了样本类属的模糊性,因此能够更客观地反应现实世界,成为聚类分析研究的热点和主流。著名学者Ruspini[3]首先提出了模糊划分的概念,把模糊集理论引入到聚类分析中。随后研究者提出了多种模糊聚类分析方法,比较典型的有基于模糊等价关系的传递闭包方法[4,5]、基于相似性关系[6]和模糊关系[7,8]的方法、基于模糊图论的最大数方法[9]、基于数据集的凸分解[10]、动态规划[11]和难以辨别关系[12]等方法。但是以上的模糊聚类方法不适用于大数据量情况,难以满足实时性要求高的场合,并且计算复杂度较高,故在实际应用和研究中已经逐渐减少。文献中研究最多、实际中应用最广的是基于目标函数的模糊聚类方法,该类方法把聚类问题描述

到稿日期:2012-01-06 返修日期:2012-06-06 本文受国家自然科学基金项目(61173091),国家科技支撑计划项目(2007BAH08802),陕西省13115科技创新工程重大专项(2007ZDKG-57)资助。

陈东辉(1984—),男,博士生,主要研究方向为数据挖掘,E-mail:chen-donghui@163.com;刘志镜(1957—),男,教授,博士生导师,主要研究方向为数据挖掘和视觉计算;王纵虎(1984—),男,博士生,主要研究方向为文本挖掘。

为一个带约束的优化问题，通过求解优化问题的解来确定数据集的模糊划分和聚类结果。此类算法设计简单、易于应用且聚类性能良好，并借助于经典的数学非线性规划理论来求解优化问题，容易编程实现。

模糊 C 均值(FCM)聚类算法[13,14]是应用最广的一种模糊聚类算法，但是它也存在一些不足：(1)FCM 使数据点在所有类中的隶属度之和为 1，因此产生的隶属度结果不能真实反应数据点对其所属类的依附关系；(2)由于 FCM 产生的划分矩阵不能真实反应数据样本点与类的隶属关系，这必然导致 FCM 没有较好的稳健性，它对噪声是敏感的。为了克服 FCM 的这些缺点，Krishnapuram 和 Keller 放弃了 FCM 的可能性约束条件，构造了一个新的目标函数，提出了 PCM 聚类算法[15]。该算法能够有效地聚类包含噪声或野值点的数据，它使噪声数据或野值点具有很小的隶属度值，从而它们对聚类结果的影响可以忽略不计，但是 PCM 对初始聚类中心很敏感，容易陷入局部最优，导致聚类结果一致性等问题[16]。

在聚类方法的研究中，将智能优化算法与传统聚类算法相结合的方法已有一些研究。例如文献[17]中，傅景广等人将遗传算法与传统的 K-均值聚类算法相结合，利用遗传算法中的选择、交叉和变异操作对聚类中心的编码进行优化，得到了明显优于传统 K-均值算法的聚类划分效果；在文献[18]中，刘向东等人在基于遗传算法的 K-均值聚类算法基础上，提出了基于粒子群优化算法的聚类方法，得到的结果明显优于前者；在文献[19]中，王玲、贺兴时将粒子群算法与模糊 C-均值聚类算法结合，克服了传统模糊 C-均值聚类算法的缺陷，同时其在收敛速度方面也明显优于基于遗传算法的模糊 C-均值聚类算法。粒子群优化(PSO)算法[20]是一种基于群体智能的全局寻优算法。该算法由于收敛速度快，需要设定的参数少，且编程实现简单，多数情况下比遗传进化算法更快地收敛于最优解，而且可以避免完全随机寻优的退化现象，因此近年来得到学术界的广泛关注。

本文提出了一种基于 PSO 改进的可能性 C 均值模糊聚类方法。其利用改进的 PCM 算法完成粒子编码，利用改进的 PSO 算法对 PCM 进行优化，有效地解决了 PCM 算法对初始条件敏感、容易陷入局部最优以及导致聚类结果一致性的问题。实验结果也证明，本文算法在计算复杂度、聚类精度和全局寻优能力方面都有了较大的改进。

2 基本概念

本节将介绍模糊 C 均值、可能性 C 均值算法和粒子群算法的相关基础概念。

2.1 FCM

FCM 是 Bezdek 于 1981 年提出的，是目前广泛采用的一种聚类算法。算法从一个初始划分开始，需要预先指定聚类数目，还需要定义一个最优化聚类标准的目标函数，来作为各数据模式分布的代价函数。FCM 把 N 个数据向量分成 C 个模糊簇，用每个簇的聚类中心代表该类。通过反复的迭代运算，逐步降低目标函数的误差值，当目标函数收敛时，得到最终的聚类结果。

模糊聚类在数学上可以表示为如下目标函数求极值的问题：

$$\min J_{FCM}=\sum_{i=1}^{C}\sum_{k=1}^{N}(u_{ik})^m d^2(\boldsymbol{x}_k,\boldsymbol{v}_i) \tag{1}$$

其中隶属度的约束条件为：

$$\sum_{i=1}^{C}u_{ik}=1,i=1,2,\cdots,N \tag{2}$$

应用 Lagrange 乘数法，结合式(2)的约束条件对式(1)求解，可得聚类中心和隶属度的更新公式如下：

$$v_i=\frac{\sum_{k=1}^{N}u_{ik}^m\cdot\boldsymbol{x}_k}{\sum_{k=1}^{N}u_{ik}^m} \tag{3}$$

$$u_{ik}=\frac{d(\boldsymbol{x}_k,\boldsymbol{v}_i)^{-\frac{2}{m-1}}}{\sum_{j=1}^{C}d(\boldsymbol{x}_k,\boldsymbol{v}_j)^{-\frac{2}{m-1}}} \tag{4}$$

式(1)中，u_{ik} 表示的是数据模式 $\boldsymbol{x}_k$ 属于第 k 类的隶属度，$\boldsymbol{v}_i$ 为聚类中心，$C\geqslant 2$ 为聚类簇数，$d(\boldsymbol{x}_k,\boldsymbol{v}_i)$ 表示数据模式 $\boldsymbol{x}_k$ 到聚类中心 $\boldsymbol{v}_i$ 的欧式距离，m 是模糊加权指数，取值在[1.5，2.5]之间。

在 FCM 算法中，约束条件式(2)是假定每个数据点的影响力是相同的，这使得样本的隶属度不但与该类的中心有关，而且受到其他类中心位置的影响，因为它们要分享和为 1 的隶属度，这往往会影响聚类结果。图 1 所示为类 1 和类 2 含有两个噪声点分别为 A 和 B，A 和 B 的位置恰好位于两个类的中心线上，由于受到隶属度和为 1 的限制，噪声 A 和 B 被赋予了较高的隶属度。从图中可以看出，A 比 B 更靠近两个类，但是它们的隶属度却是相同的，都为 0.5，因此 FCM 算法对噪声数据敏感，而且无法反映出真实情况。

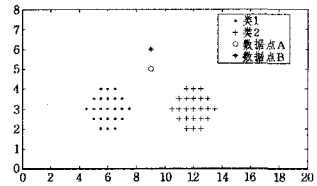

图 1 含噪声数据的两个类

2.2 PCM

PCM 算法正是为了解决 FCM 算法的上述缺点被 Krishnapuram and Keller 提出来的，他们放松了隶属度的约束条件，从而解决了 FCM 对噪声数据点敏感的问题。

PCM 的目标函数定义如下：

$$J_{PCM}=\sum_{j=1}^{C}\sum_{k=1}^{N}(t_{jk})^m d_{jk}^2+\sum_{j=1}^{C}\eta_j\sum_{k=1}^{N}(1-t_{jk})^m \tag{5}$$

PCM 的聚类中心和隶属度的更新公式如下：

$$\boldsymbol{v}_i=\frac{\sum_{k=1}^{N}t_{jk}^m x_k}{\sum_{k=1}^{N}t_{jk}^m} \tag{6}$$

$$t_{jk}=\left[1+\left(\frac{d(\boldsymbol{x}_k,\boldsymbol{v}_j)^2}{\eta_i}\right)^{\frac{1}{m-1}}\right]^{-1} \tag{7}$$

式中，t_{ik} 表示第 i 类的典型值，即数据模式 $\boldsymbol{x}_k$ 属于第 i 类的概率。它只依赖于 $\boldsymbol{x}_k$ 与 $\boldsymbol{v}_i$ 的距离，与其他类中心的位置无关。比较式(4)和式(7)可知，式(7)的 $d^2(\boldsymbol{x}_k,\boldsymbol{v}_j)$ 可以取 0 值，因此 PCM 不存在 FCM 的奇异性问题。但是该算法的不足之处在于计算复杂度大，对选择的参数初始值敏感。由于 PCM 放松了约束条件，可能会使样本点无限逼近若干个簇，从而导致聚类一致性的问题。

2.3 PSO

PSO 算法是由 Kennedy 和 Eberhart 受鸟群觅食行为的启发，于 1995 年提出，用于解决复杂优化问题。它是一种基

于群体智能的随机全局优化技术，被成功地应用于聚类问题、图像处理、函数优化等。

在 PSO 算法中，粒子通过不断调整自己的位置来搜索新解，每个粒子都能记住自己搜索到的最优解，记作 p_{id}，以及整个粒子群的最优解 p_{gd}。每个粒子都有一个速度，记作 v_{id}。每个粒子通过下列公式来更新自身的位置：

$$v_{id}^{(t+1)}=\omega v^{(t)}+c_1 r_1(p_{id}^{(t)}-s_{id}^{(t)})+c_2 r_2(p_{gd}^{(t)}-s_{id}^{(t)}) \tag{8}$$

式中，v_{id} 表示第 i 个粒子在第 d 维上的速度；ω 是惯性权重；c_1 和 c_2 分别是个体和全局的加速系数；r_1 和 r_2 分别是 0 和 1 之间的随机数。则粒子移动的下一个位置为：

$$s_{id}^{(t+1)}=s_{id}^{(t)}+v_{id}^{(t)} \tag{9}$$

作为一个粒子，每个粒子在解空间中不断进化搜索直到达到终止条件。粒子进化过程评估自身性能的标准是适应度函数。使用 PSO 算法作为本文算法的载体，是因为粒子群算法被证明有着良好的寻优性能和收敛性。

3 基于 PSO 的 PCM 改进算法

本节将提出一个基于粒子群优化的 PCM 改进方法 IPCM-PSO(Improved Possibilistic C-means Method based on PSO)。

3.1 PCM 改进算法

在第 2.1 节和 2.2 节已经讨论了 FCM 和 PCM 算法的优缺点，FCM 算法虽然计算上比 PCM 简单，但是它对噪声数据敏感，不能真实反映数据样本点与类的隶属关系，缺乏较好的稳健性。PCM 虽然克服了 FCM 算法对噪声敏感，却带来了聚类一致性的问题，即当类与类之间有重叠时可能得到错误的结果；另外它的惩罚因子 η_i 的计算依赖 FCM 算法，并且在实际应用中大大增加了算法的复杂度。因此结合它们的优点并且用文献[21]的协方差矩阵来优化 η_i。给出 PCM 改进算法 IPCM，定义的目标函数定义如下：

$$J_{IPCM}=\sum_{i=1}^{C}\sum_{k=1}^{N}(u_{ik}^{m}+t_{ik}^{q})d_{ik}^{2}+\frac{\beta}{mq\sqrt{c}}\sum_{i=1}^{C}\sum_{k=1}^{N}(t_{ik}^{q}\log t_{ik}^{q}-t_{ik}^{q}) \tag{10}$$

式中，$\beta=\frac{1}{N}\sum_{k=1}^{N}\|x_k-\bar{x}\|^2$，$\bar{x}=\frac{1}{N}\sum_{k=1}^{N}x_k$。

由式(10)可以看出，目标函数前一项是 FCM 和 PCM 目标函数的结合，后一项是目标函数的惩罚项，这是为了避免目标函数产生无意义的平凡解。另外，引入协方差矩阵可以很好地反映样本数据集的紧凑和分离程度，使聚类效果更好[8]。

要使目标函数求解最优化(最小化)，前一项要求数据集的特征向量到聚类中心的距离尽可能的小，后一项是单调递减的函数，所以 t_{ik} 要尽可能的大。因此式(10)最优解的必要条件是：

$$u_{ik}=\frac{\|x_k-v_i\|^{-\frac{2}{m-1}}}{\sum_{j=1}^{C}\|x_k-v_j\|^{-\frac{2}{m-1}}} \tag{11}$$

$$t_{ik}=\exp\left(-\frac{mq\sqrt{c}\,\|x_k-v_i\|^2}{\beta}\right) \tag{12}$$

$$v_i=\frac{\sum_{k=1}^{N}(u_{ik}^{m}+t_{ik}^{q})x_k}{\sum_{k=1}^{N}(u_{ik}^{m}+t_{ik}^{q})} \tag{13}$$

3.2 PSO 改进算法

由式(8)可知，因为 r_1 和 r_2 是随机的，所以在每次迭代中的学习能力也是随机的。然而为了使优秀的粒子更容易地被其他粒子学习，本文借鉴文献[10]中的继承机制改进粒子群优化算法：

$$\omega_1=c_1 r_1=\frac{p_{id}^{(t)}}{f_{id}^{(t)}}+\omega,\omega_2=c_2 r_2=\frac{p_{gd}^{(t)}}{p_{id}^{(t)}} \tag{14}$$

式中，$f_{id}^{(t)}$ 表示粒子 i 在第 t 代的适应度值。引入继承机制能够加快整个群体的收敛速度。

惯性权重 ω 较大，则算法具有较强的全局搜索能力；反之，则算法倾向于局部搜索。通常将 ω 由初始的最大值 0.9 随着迭代次数的增加线性递减至 0.4，所以 ω 设置如下：

$$\omega=0.9-\frac{t}{iter_{total}}\times 0.5 \tag{15}$$

3.3 基于 PSO 的 PCM 改进算法

对 d 维的样本空间 $X=\{x_1,x_2,\cdots,x_n\}$ 进行聚类，即要找到并计算各个聚类中心，按照各样本与该聚类中心的度量来计算隶属度。粒子群中一个粒子的位置和速度都是 $c\times d$ 维向量(c 表示类的个数)。用改进的粒子群算法，粒子通过每一次迭代改变聚类中心的取值来产生多种聚类结果，直到找到可接受的聚类中心，即适应度函数达到终止条件。

本文基于粒子群优化的 PCM 改进算法的具体步骤如下：

(a)初始化聚类中心和粒子群算法的参数，包括聚类数目 c、模糊指标参数 m 和 $q(m,q>1)$、群体规模 n，根据式(14)设置学习因子 ω_1 和 ω_2，根据式(15)设置惯性权重 ω、迭代的最大次数 $iter_{total}$，阈值 $\varepsilon=0.0001$；

(b)将每个样本随机分为某一类，初始化聚类中心 $v_i^{(0)}=\{v_{i1}^{(0)},v_{i2}^{(0)},\cdots,v_{id}^{(0)}\}$，$i=1,2,\cdots,c$，完成粒子编码，则会形成 n 个初始粒子；对于每一个粒子计算出它的个体极值 $p_{id}^{(0)}$ 和在整个粒子群中的最优解 $p_{gd}^{(0)}$；

(c)由式(11)、式(12)来计算或更新可能性划分矩阵 u_{ik} 和 t_{ik}；

(d)由式(13)更新每个聚类中心 $v_i^{(t)}=\{v_{i1}^{(t)},v_{i2}^{(t)},\cdots,v_{id}^{(t)}\}$；

(e)根据式(10)计算每个粒子的适应度值，比较以前的粒子适应度值与当前的适应度值，如果当前值更好(值更小)，则用当前的值 $p_{id}^{(t+1)}$ 取代 $p_{id}^{(t)}$ 和 $p_{gd}^{(t+1)}$ 取代 $p_{gd}^{(t)}$；否则保持 $p_{id}^{(t)}$ 和 $p_{gd}^{(t)}$；

(f)根据式(8)和式(9)分别更新粒子的速度和粒子位置；

(g)如果达到最大迭代次数或者达到足够好的位置(最优解对应的目标函数值不变或者变动小于阈值 ε)，或者持续迭代次数达到设定值 $iter_{total}$，则结束；否则转步骤(c)。最终得到整个搜索空间找到的最小适应度值对应的簇划分。

4 实验结果

为了验证本文算法的有效性，对算法进行了实验仿真。实验环境为 CPU：Intel(R) Core2 Duo 2.93GHz；内存：4G；硬盘 500G。实验平台：Microsoft Visual C# 和 Matlab。

实验 1　采用了一个简单人工数据集 X_{12}[11] 进行实验，以验证算法对噪声点的处理效果。数据集包含 12 个样本，每个样本 2 个属性，整体分布如图 2 所示，x_{11} 和 x_{12} 是预设的两个噪声点。

这里，令 $c=2$，$m=q=2.0$，$\varepsilon=0.0001$。初始化聚类中心：

$$v^{(0)}=\begin{pmatrix}-3.34 & 1.67\\ 1.67 & 0.00\end{pmatrix}$$

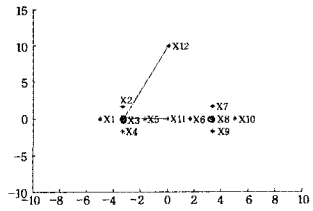

图 2　数据集 X_{12} 分布

根据初始中心分别用 FCM、PCM 和 IPCM-PSO 来计算隶属度矩阵和更新聚类中心，最终的隶属度矩阵 $\boldsymbol{U}$ 和 $\boldsymbol{T}$ 如表 1 所列。

表 1　FCM、PCM 和 IPCM-PSO 算法对 X_{12} 的隶属度值

样本	FCM		PCM		IPCM-PSO			
	U_1^T	U_2^T	T_1^T	T_2^T	U_1^T	U_1^T	T_1^T	T_2^T
x_1	0.94	0.06	0.49	0.13	0.96	0.04	0.79	0.10
x_2	0.97	0.03	0.66	0.20	0.94	0.06	0.70	0.15
x_3	0.99	0.01	0.85	0.21	0.99	0.01	0.97	0.17
x_4	0.90	0.10	0.65	0.19	0.93	0.07	0.75	0.19
x_5	0.92	0.08	0.97	0.35	0.94	0.06	0.78	0.20
x_6	0.08	0.92	0.35	0.97	0.06	0.94	0.20	0.78
x_7	0.03	0.97	0.19	0.66	0.06	0.94	0.15	0.70
x_8	0.01	0.99	0.21	0.85	0.01	0.99	0.17	0.97
x_9	0.10	0.90	0.19	0.65	0.07	0.93	0.19	0.75
x_{10}	0.06	0.94	0.13	0.49	0.04	0.96	0.10	0.79
x_{11}	0.50	0.50	0.63	0.63	0.50	0.50	0.24	0.24
x_{12}	0.50	0.50	0.07	0.07	0.50	0.50	0.01	0.01

从表 1 的结果可以看出，利用 FCM 算法得到的数据点 x_{11} 和 x_{12} 的隶属度值都为 0.5，然而由图 2 可知，x_{12} 到聚类中心距离是 x_{11} 到聚类中心的距离的 3 倍多。故 FCM 对噪声点是不能区分的。PCM 算法虽然克服了 FCM 的隶属度之和为 1 的约束条件，但是对 x_{11} 的隶属度取值过大(0.63)。IPCM-PSO 对两个噪声点的典型值都很小，分别为 0.24 和 0.01，因此该算法很好地避免了对噪声数据点的敏感，具有很好的鲁棒性。

由图 2 可以看出，X_{12} 的理想聚类中心为点 x_3 和 x_8，即 $\boldsymbol{v}_{ideal}=\begin{pmatrix}-3.34 & 3.34\\ 0.00 & 0.00\end{pmatrix}$ 表 2 是各算法经过若干次迭代后的聚类中心以及该中心到理想中心的欧氏距离。由表 2 的结果可以看出，本文算法 IPCM-PSO 计算的聚类中心与理想中心更为接近，因此聚类精度比其他两种算法更高。

表 2　各算法运行的聚类中心及其到理想中心的欧氏距离

算法	FCM	PCM	IPCM-PSO
聚类中心	$\begin{pmatrix}-2.99 & 2.99\\ 0.54 & 0.54\end{pmatrix}$	$\begin{pmatrix}-2.15 & 2.15\\ 0.02 & 0.02\end{pmatrix}$	$\begin{pmatrix}-3.33 & 3.33\\ 0.02 & 0.02\end{pmatrix}$
欧氏距离	0.9101	1.6832	0.0316

实验 2　采用 UCI 机器学习数据库[24]的 Iris 标准数据集进行试验，以验证聚类算法的有效性。与本文算法进行对比的方法包括 FCM、PCM 及其改进算法 UPCM[21]、IPCM[22]。Iris 数据集共有 150 个样本，每个样本有 4 个特征，分别为花瓣长度、花瓣宽度、萼片长度和萼片宽度；共包含 3 个种类，每类 50 个样本，其中第一类与其他两个类完全分离，其余两个类之间有交叉。实验结果如表 3 所列。Iris 数据集是作为检验聚类算法效率的一个较标准的测试数据，在 Zedeh 等学者提出的聚类算法中也主要是用该数据集结果作比较。国外文献中给出 Iris 数据实际的聚类中心为(v_1=(6.58, 2.97,5.55,2.02)，v_2=(5.93,2.77,4.26,1.32)，v_3=(5.00, 3.42,1.46,0.24))。

表 3　各算法在 Iris 数据集上的聚类结果比较

算法	聚类中心	误分数	平均耗时	聚类精度(%)
FCM	(6.77,3.05,5.65,2.05) (5.88,2.76,4.36,1.40) (5.00,3.41,1.48,0.25)	16	0.27	89.33
PCM	(6.17,2.92,4.86,1.67) (6.17,2.92,4.86,1.67) (5.01,3.41,1.48,0.25)	23	1.54	84.67
UPCM	(6.60,3.01,5.50,1.94) (5.87,2.76,4.32,1.38) (5.03,3.48,1.46,0.25)	**10**	**0.92**	**93.33**
IPCM	(6.58,3.00,5.51,2.05) (5.91,2.79,4.29,1.38) (4.96,3.40,1.49,0.27)	8	0.75	94.67
IPCM-PSO	(6.61,2.98,5.53,2.04) (5.92,2.78,4.29,1.34) (5.01,3.40,1.47,0.25)	7	0.45	95.33

误分数表示没有被正确划分所属簇类的数据对象的数目，聚类精度表示实验数据对象被正确划分所属簇类的数目占数据集所有样本的比例。从表 3 的实验结果可以看出，FCM 算法在聚类过程中平均耗时最小，收敛速度最快，但是由于第 2 节讨论过的样本的隶属度不但与该类的中心有关，而且受到其他类中心位置的影响，因此其往往会影响聚类结果，导致聚类精度最低。PCM 算法改善了 FCM 算法的不足，聚类精度有所提高，但是它的第一个和第二个聚类中心出现了重合，与实际的聚类中心不符，出现了聚类一致性的问题。UPCM、IPCM 与本文算法都达到了 90%以上，并且未产生重合聚类。IPCM-PSO 算法不仅对 PCM 作了改进，而且引入了 PSO 算法，聚类结果有了很大的提高，全局搜索能力明显增强，而且节省了时间。

图 3 为 Iris 数据集的适应度收敛曲线，其展示了种群中随机 5 个粒子的最优适应度变化曲线和粒子群全局最优变化曲线。从图中可以看出，由于本文算法具有较强的全局寻优能力，每代粒子之间可以共享社会信息及各个粒子的自我经验，不存在随机寻优的退化现象，因此收敛比较平稳，无震荡现象，且有较快的收敛速度，实验中的数据集在 50 次迭代内适应度函数值能都到达收敛状态。

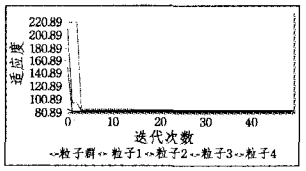

图 3　Iris 数据集粒子群适应度收敛曲线

为了能够直观显示聚类结果，利用 matlab 编程工具对本文算法进行仿真。取 Iris 数据集的花萼长度和花瓣长度这两个属性分别作为横轴和纵轴，先对数据进行标准化预处理，将其映射到[0,1]范围之间，处理后的数据分布如图 4 所示。

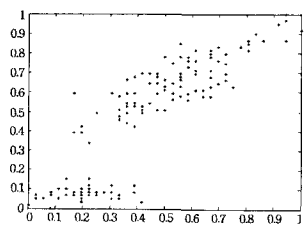

图 4　Iris 数据的二维映射分布

从图5可以很清晰地看出用IPCM-PSO算法对Iris数据集聚类的结果，聚类中心用红色的圆圈显示在图中，蓝色弧线把数据集分成了3类，避免了PCM算法出现簇类重叠的现象。可见该算法得到了正确的聚类结果。

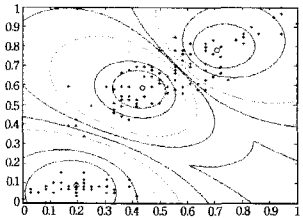

图5 IPCM-PSO算法产生的结果

结束语 基于粒子群优化的可能性C-均值聚类算法结合了FCM和PCM算法中的优点，并且引入协方差矩阵来构造目标函数，很好地反映了数据集的紧凑程度，提高了聚类的精度；利用粒子群优化算法，提高了全局搜索能力，减少了算法迭代次数，加快了算法的收敛。通过实验结果和分析证明了本文算法的有效性。

参 考 文 献

[1] 高新波. 模式聚类分析及应用[M]. 西安：西安电子科技大学出版社，2004

[2] 李相镐，等. 模糊聚类分析及其应用[M]. 贵州：贵州科技出版社，1994

[3] Ruspini E H. A new approach to clustering [J]. Information and control，1969，15(1)：22-32

[4] Dunn J C. A graph theroretic analysis of pattern classification via tamura's fuzzy relation [J]. IEEE Trans. SMC，1974，4(3)：310-313

[5] Le Z. Fuzzy relation compositions and pattern recognition [J]. Information Sciences，1996，89：107-130

[6] Tamra S，et al. Pattern classification based on fuzzy relations [J]. IEEE Trans. SMC，1971，1(1)：217-242

[7] Backer E，Jain A K. A clustering performance measure based on fuzzy set decomposition[J]. IEEE Trans. PAMI，1981，3(1)：66-74

[8] Zadeh L A. Similarity relations and fuzzy orderings[J]. Inf Sci，1971，3：177-200

[9] Leahy R，Wu Z. An optimal graph theoretic approach to data clustering：theory and its application to image segmentation [J]. IEEE Trans on PAMI，1993，15(11)：1103-1113

[10] Bezdek J C，Harris J O. Convex decompositions of fuzzy ISODATA clustering algorithm[J]. IEEE Tans. PAMI，1980，1(2)：1-8

[11] Esogbue A O. Optimal clustering of fuzzy data via fuzzy dynamic programming[J]. FSS，1986，26(1)：127-130

[12] Mantaras R L D，Valverde L. New results in fuzzy clustering based on the concept of indistinguishablity relation[J]. IEEE Trans. PAMI，1988，10(5)：754-757

[13] Bezdek J C. Pattern Recognition with Fuzzy Objective Function Algorithms[M]. Plenum，New York，1981

[14] 彭代强，李家强，林幼权. 基于模糊隶属度空间约束的FCM图像分割[J]. 计算机科学，2010，37(10)：257-259

[15] Krishnapuram R，Keller J M. A possibilistic approach to clustering [J]. IEEE Trans Fuzzy System，1993，1(2)：98-110

[16] Barni M，Cappellini V，Ecocci A M. Comments on A possibilistic approach to clustering [J]. IEEE Trans Fuzzy Systems，1996，4(3)：393-396

[17] 傅景广，许刚，王裕国. 基于遗传算法的聚类分析[J]. 计算机工程，2004，30(4)：122-124

[18] 刘向东，沙秋夫，等. 基于粒子群优化算法的聚类分析[J]. 计算机工程，2006，32(6)：201-202

[19] 王玲，贺兴时. 基于粒子群优化的模糊聚类分析[J]. 价值工程，2007，11：96-98

[20] 李朔枫，李太勇. 一种基于距离的自适应模糊粒子群优化算法[J]. 计算机科学，2011，38(8)：257-259

[21] Yang M S，Wu K L. Unsupervised possibilistic clustering[J]. Pattern Recognition，2006，39(1)：5-21

[22] Zhang J S，Yeung Y W. Improved possibilistic C-means clustering algorithms[J]. IEEE Transactions on Fuzzy Systems，2004，12(2)：209-217

[23] Yu Jin，Qian Feng，Qi Rong-bin. Improvement of Stochastic Particle Swarm Optimization by Succession Strategy[J]. Communications of the Systemics and Informatics World Network，2008，3：155-159

[24] Pal N R，Pal K，Bezdek J C，et al. A possibilistic fuzzy C-Means clustering algorithm [J]. IEEE Trans Fuzzy Systems，2005，13(4)：517-530

[25] UCI Repository of Machine Learning Databases retrieved from the World Wide Web[OL]. http://www. ics. uci. edu/~ mlearn/MLRepository. html，1998

[26] Gao Ying，Wang Xiu-liang，Lu Xu-qing，et al. Possibilistic C-Means Clustering Algorithm Based on Particle Swarm Optimization [J]. Computer Simulation，2010(9)：177-180

[27] Niu Qiang，Huang Xin-jian. An improved fuzzy C-means clustering algorithm based on PSO [J]. Journal of Software，2011，6(5)：873-879

(上接第101页)

[2] Zheng Zi-bin，Lyu M R. WS-DREAM：A distributed reliability assessment Mechanism for Web Services[C]//DSN. 2008：392-397

[3] 邵凌霜，李田，赵俊峰，等. 一种可扩展的Web Service QoS管理框架[J]. 计算机学报，2008，31(8)：1458-1470

[4] ISO/IEC (2003) ISO/IEC 9126-2 Technical Report：Software engineering. Product quality[S]. Part 2：External metrics. 86

[5] Stefani A，Xenos M N. E-commerce system quality assessment using a model based on ISO 9126 and Belief Networks[J]. Software Quality Journal，2008，16(1)：107-129

[6] Poole D，Smyth C，Sharma R. Ontology Design for Scientific Theories That Make Probabilistic Predictions[J]. IEEE Intelligent Systems，2009，24(1)：27-36

[7] Leontidis M，Halatsis C. Supporting Learner's Needs with an Ontology-Based Bayesian Network[C]//ICALT. 2009：579-583

[8] 张连文，郭海鹏. 贝叶斯网络引论[M]. 北京：科学出版社，2006