一种基于密度函数的直觉模糊聚类初始化方法

申晓勇 雷英杰 蔡 茹 雷 阳
(空军工程大学导弹学院 三原 713800)

摘 要 针对基于目标函数的直觉模糊聚类方法容易陷于局部最优值的问题,提出了一种改进的密度函数初始化方法。该方法首先利用样本密度函数在较高局部密度的区域中选取c个样本,然后遍历剩余样本进行粗归类,并计算每类各维数据的平均值作为初始聚类中心。最后通过典型实例验证,该方法不仅解决了容易陷入局部极小值的问题,同时迭代次数减少,收敛速度加快,提高了聚类性能。

关键词 直觉模糊集合,直觉模糊聚类,目标函数,密度函数
中图分类号 TP182, TP391 **文献标识码** A

Initialization Method for Intuitionistic Fuzzy Clustering Based on Density Function

SHEN Xiao-yong LEI Ying-jie CAI Ru LEI Yang
(Air Force Engineering University, Missile Institute, Sanyuan 713800, China)

Abstract To the problems of the technique of Intuitionistic fuzzy clustering based on objective function immersed in partial optimization figure, an improved initialization method based on density function was proposed. First, the proposed method have selected c samples in the upper partial density making use of sample density function. Then the other samples were visited and classified roughly. Moreover, the every kind of average values of all dimensions were calculated, which are counted as initial clustering center. At last, the validity of the technique proposed was checked with an classical instance, and the technique not only solved the problem of functions easily immersed in partial optimization figure, which takes on, but also reduced iterative times and improve convergent speed, so as to improve clustering capability.

Keywords Intuitionistic fuzzy set, Intuitionistic fuzzy clustering, Objective function, Density function

1 引言

Atanassov 直觉模糊集合(Intuitionistic Fuzzy Sets, IFS)[1]是对 Zadeh 模糊集合最有影响的一种扩充和发展。IFS增加了一个新的属性参数——非隶属度函数,进而可以描述"非此非彼"的"模糊概念",更加细腻地刻画客观世界的模糊性本质[2],因而引起众多学者的关注。

近年来,基于直觉模糊集合的模式识别、数据挖掘、计算机视觉和模糊控制等许多领域的研究,均要求对数据进行聚类分析,因而直觉模糊聚类方法成为新的研究热点。文献[3]在FCM聚类算法的基础上,提出了一种基于目标函数的IFS数据的聚类方法,可用于解决大数据量以及实时性要求很高的聚类问题。但是该方法本质上是一种局部寻优技术,它的迭代过程采用爬山技术来寻找最优解,因此该算法对初始聚类中心比较敏感,容易陷入局部极小值,而得不到全局最优解。同时算法的收敛速度受初始值影响较大,特别是在聚类数较大的情况下,这一缺点更为突出。如果选取的初始聚类中心接近数据真正的类别中心,算法的收敛速度和准确性都将得到改善。

目前,基于传统模糊聚类中心初始化的方法很多,如网格密度法[4]、密度评估法[5]、密度指针法[6]和距离优化法[7]等。文献[4,6]提出的方法对高维数据很难奏效,因为高维空间中样本的分布比较分散,很难形成支持度很高的聚类,给网格划分造成难度。并且随着维数的增加,网格单元也剧增,提高了算法复杂度。文献[7]通过选取足够体现类间相异性的点作为初始聚类中心,但对输入参数和实例的输入顺序都非常敏感。文献[5]是一种基于密度评估的方法,通过两两样本距离的比较完成对密度的评估,从输入样本中具有较高局部密度的区域中选取初始聚类中心。

鉴于此,本文受文献[5]的启发,对其改进并进行直觉化扩展,提出一种改进的基于密度函数的直觉模糊聚类初始化方法,最后通过实例进行验证。

2 直觉模糊集合理论

为了描述方便,我们首先给出直觉模糊集的定义。Atanassov对直觉模糊集给出如下定义[1]:

定义1(直觉模糊集合) 设 X 是一个给定论域,则 X 上的一个直觉模糊集 A 为

$$A=\{\langle x,\mu_A(x),\gamma_A(x)\rangle \mid x\in X\} \quad (1)$$

其中 $\mu_A(x):X\to[0,1]$ 和 $\gamma_A(x):X\to[0,1]$ 分别代表 A 的隶

到稿日期:2008-06-18 本文受国家自然科学基金资助项目(No. 60773209),陕西省自然科学基金资助项目(No. 2006F18)资助。
申晓勇(1982-),男,博士生,主要从事智能信息处理研究,E-mail:yuzhong25@163.com;雷英杰(1956-),男,博士,教授,博士生导师,主要从事智能信息处理与智能决策等研究。

属函数$\mu_A(x)$和非隶属函数$\gamma_A(x)$，且对于A上的所有$x\in X$，$0\leqslant\mu_A(x)+\gamma_A(x)\leqslant1$成立。

直觉模糊集A有时可以简记作$A=\langle x,\mu_A,\gamma_A\rangle$或$A=\langle\mu_A,\gamma_A\rangle/x$。显然，每一个一般模糊子集对应于下列直觉模糊子集$A=\{\langle x,\mu_A(x),1-\mu_A(x)\rangle|x\in X\}$。

对于X中的每一个直觉模糊子集，称$\pi_A(x)=1-\mu_A(x)-\gamma_A(x)$为$A$中$x$的直觉指数(Intuitionistic Index)，它是x对A的犹豫程度(Hesitancy degree)的一种测度。显然，对于每一个$x\in X$，$0\leqslant\pi_A(x)\leqslant1$，对于X中的每一个一般模糊子集A，$\pi_A(x)=1-\mu_A(x)-[1-\mu_A(x)]=0$，$\forall x\in X$。

3 直觉模糊聚类初始化算法

给定样本集$X=\{x_1,x_2,\cdots,x_n\}\subset R^s$为模式空间中$n$个模式的一组有限观测样本集，$x_j=(\langle x\mu_{j1},x\gamma_{j1},x\pi_{j1}\rangle,\langle x\mu_{j2},x\gamma_{j2},x\pi_{j2}\rangle,\cdots,\langle x\mu_{js},x\gamma_{js},x\pi_{js}\rangle)^T$为观测样本$x_j$的特征矢量，特征矢量每维特征上的赋值$\langle x\mu_{jk},x\gamma_{jk},x\pi_{jk}\rangle$均为一个直觉模糊数。$P=\{p_1,p_2,\cdots,p_c\}$是$c$个聚类原型，$c$为聚类类别数，$p_i$表示第$i$类的聚类原型矢量，$p_i=\{\langle p\mu_{i1},p\gamma_{i1},p\pi_{i1}\rangle,\langle p\mu_{i2},p\gamma_{i2},p\pi_{i2}\rangle,\cdots,\langle p\mu_{is},p\gamma_{is},p\pi_{is}\rangle\}$，$p_i$第$k$维特征上的赋值$p_{ik}=\langle p\mu_{ik},p\gamma_{ik},p\pi_{ik}\rangle$也为直觉模糊数。

$$D_w(x_j,p_i)=\sqrt{\frac{(x\mu_j-p\mu_i)A(x\mu_j-p\mu_i)^T+(x\gamma_j-p\gamma_i)A(x\gamma_j-p\gamma_i)^T+(x\pi_j-p\pi_i)A(x\pi_j-p\pi_i)^T)}{2}}$$

$$=\frac{1}{\sqrt{2s}}\sqrt{\sum_{k=1}^{s}\omega(k)(|x\mu_{jk}-p\mu_{ik}|^2+|x\gamma_{jk}-p\gamma_{ik}|^2+|x\pi_{jk}-p\pi_{ik}|^2)}\tag{5}$$

其中$x\mu_j$，$p\mu_i$表示隶属度矢量，$x\gamma_j$，$p\gamma_i$表示非隶属度矢量，$x\pi_j$，$p\pi_i$表示犹豫度矢量，且$x\mu_j+x\gamma_j+x\pi_j=I$，$p\mu_i+p\gamma_i+p\pi_i=I$($I$为$s$维单位矢量)；$x\mu_{jk}+x\gamma_{jk}+x\pi_{jk}=1$，$p\mu_{ik}+p\gamma_{ik}+p\pi_{ik}=1$；矩阵$A$为加权对角矩阵，$\omega(k)\geqslant0(k=1,2,\cdots,s)$是加于第$k$维特征上的权数，$\omega(k)$满足归一化条件：

$$\frac{1}{s}\sum_{k=1}^{s}\omega(k)=1\tag{6}$$

具体算法如下。

初始化：给定聚类类别数c，$2\leqslant c\leqslant n$，n为样本数据个数，设定迭代停止阈值ε，平滑参数m，初始聚类原型模式$P^{(0)}$，设置迭代计数器$b=0$。

步骤1　用式(7)计算、更新划分隶属矩阵U_μ，划分非隶属矩阵U_γ：

$$\begin{cases}\mu_{ij}=\dfrac{1}{\sum\limits_{k=1}^{c}\left[\dfrac{D_w(x_j,p_i)}{D_w(x_j,p_k)}\right]^{\frac{2}{m-1}}},\ \gamma_{ij}=1-\pi_{ij}-\dfrac{1}{\sum\limits_{k=1}^{c}\left[\dfrac{D_w(x_j,p_i)}{D_w(x_j,p_k)}\right]^{\frac{2}{m-1}}}\\ \quad\forall k,D_w(x_j,p_k)\neq0\\ \mu_{ij}=1,\ \gamma_{ij}=0\ \ \exists k,D_w(x_j,p_k)=0,\text{且 }i=k\\ \mu_{ij}=0,\ \gamma_{ij}=1\ \ \exists k,D_w(x_j,p_k)=0,\text{且 }i\neq k\end{cases}\tag{7}$$

对于$\forall i,j$，如果$D_w(x_j,p_k)^{(b)}>0$，则有

$$\begin{cases}\mu_{ij}^{(b)}=\left\{\sum\limits_{k=1}^{c}\left(\dfrac{D_w(x_j,p_k)^{(b)}}{D_w(x_j,p_k)^{(b)}}\right)^{\frac{2}{m-1}}\right\}^{-1}\\ \gamma_{ij}=1-\pi_{ij}-\left\{\sum\limits_{k=1}^{c}\left(\dfrac{D_w(x_j,p_k)^{(b)}}{D_w(x_j,p_k)^{(b)}}\right)^{\frac{2}{m-1}}\right\}^{-1}\end{cases}\tag{8}$$

如果$\exists k$，使得$D_w(x_j,p_k)^{(b)}=0$，则有

$$\begin{aligned}&\mu_{ij}=1,\ \gamma_{ij}=0\ \ i=k\\&\mu_{ij}=0,\ \gamma_{ij}=1\ \ i\neq k\end{aligned}\tag{9}$$

步骤2　用式(10)、(11)、(12)更新聚类原型模式矩阵$P_i^{(b+1)}$，分别求得$p\mu_i^{(b+1)}$，$p\gamma_i^{(b+1)}$和$p\pi_i^{(b+1)}$：

$$p\mu_i=\frac{1}{\sum\limits_{j=1}^{n}((\mu_{ij})^m/2+(1-\gamma_{ik})^m/2)}\sum_{j=1}^{n}((\mu_{ij})^m/2+(1-\gamma_{ik})^m/2)x\mu_j\tag{10}$$

$$p\gamma_i=\frac{1}{\sum\limits_{j=1}^{n}((\mu_{ij})^m/2+(1-\gamma_{ik})^m/2)}\sum_{j=1}^{n}((\mu_{ij})^m/2+(1-\gamma_{ik})^m/2)x\gamma_j\tag{11}$$

$$p\pi_i=I-p\mu_i-p\gamma_i\tag{12}$$

步骤3　如果$\|P^{(b)}-P^{(b+1)}\|>\varepsilon$，则令$b=b+1$，转向步骤1，否则算法停止，并输出划分隶属矩阵$U_\mu$，划分非隶属矩阵$U_\gamma$和聚类原型$P$。其中$\|\cdot\|$为某种合适的矩阵范数。

首先给出文献[3]提出的基于目标函数的直觉模糊聚类算法。

3.1 基于目标函数的直觉模糊聚类算法

首先给出基于目标函数的直觉模糊聚类的描述形式：

$$\begin{cases}J_m(U_\mu,U_\gamma,P)=\sum\limits_{j=1}^{n}\sum\limits_{i=1}^{c}((\mu_{ij})^m/2+(1-\gamma_{ik})^m/2)\\ \quad D_w(x_j,p_i)^2,m\in[1,\infty),U_\mu\in M_{IFC},U_\gamma\in M_{IFC}\\ M_{IFC}=\{U_\mu\in R^{cn},U_\gamma\in R^{cn}|\mu_{ik}\in[0,1],\gamma_{ik}\in[0,1],\\ \quad 0<\sum\limits_{k=1}^{n}\mu_{ik}<n,0<\sum\limits_{k=1}^{n}\gamma_{ik}<n,\forall i,\forall k\}\end{cases}\tag{2}$$

其中，m称作平滑参数(不作特殊要求可取$m=2$)，U_μ为模糊划分隶属矩阵，U_γ为模糊划分非隶属矩阵，而且

$$\mu_{ij}+\gamma_{ij}+\pi_{ij}=1\tag{3}$$

同时

$$\sum_{i=1}^{c}\mu_{ik}=1\tag{4}$$

还有

3.2 初始聚类中心的选择

如上所提算法对于初始聚类原型的选择采用随机样本法，对初始参数比较敏感，故需要选择合理的初始聚类中心，从而减少迭代次数，降低算法的计算复杂度。

定义2(密度函数)　样本点x_j处的密度函数定义如下：

$$D_j^{(0)}=\sum_{i=1}^{n}\frac{1}{1+r_dD_w(x_i,x_j)^2}\tag{13}$$

其中，r_d是领域密度有效半径，它的选择与数据集合的分布特性有关。这里取r_d为n个样本的距离之和，即

$$r_d=\sum_{j=1}^{n}\sum_{i=1}^{n}D_w(x_i,x_j)^2\tag{14}$$

由式(14)可知，在x_i周围样本点越密集，则$D_j^{(0)}$值越大。故可以用来表示在样本空间中样本点的密集程度。

下面给出具体的初始聚类中心选择算法。

步骤1　给定n个样本，用式(14)计算每一个样本点处的密度函数值，取

$$D_1^*=\max\{D_i^{(0)},i=1,2,\cdots,n\}\tag{15}$$

则第一个模拟聚类中心x_1^*为D_1^*对应的样本点；

步骤2　根据式(16)、(17)迭代计算其余模拟聚类中心：

$$D_j^{(k)}=\left|D_j^{(k-1)}-D_k^*\frac{1}{1+f_dD_w(x_j,x_k^*)^2}\right|\tag{16}$$

其中，$k=1,2,\cdots,c-1$，c 为聚类数目，j 不等于已选模拟聚类中心样本对应的下标，而且，

$$D_m^*=\max\{D_i^{(m-1)},i=1,2,\cdots,n\} \tag{17}$$

其中，$m=2,\cdots,c$，i 不等于已选模拟聚类中心样本对应的下标，则 D_m^* 对应的样本点 x_m^* 取为第 m 个模拟聚类中心；

步骤3　定义集合 $I=X-\{I_k\}$，$I_k=\{x_k^*\}(k=1,2,\cdots,c)$，通过距离度量式(5)，分别计算集合 I 中元素离哪个模拟聚类中心 x_k^* 最近，将其划入对应集合 I_k，同时去掉集合 I 中该元素，直到 $I=\varnothing$ 循环停止；

步骤4　分别计算集合 I_k 中所有元素的平均值 $\overline{x}_k^*$，作为初始聚类中心。

该算法时间复杂度为 $O(n^2)$，只与样本数有关，而与样本维数无关，对于高维数据可以取得良好的聚类效果。

4　实例验证

为了验证算法可行，现对某一时刻来袭的20批空中目标进行聚类分析，目标的聚类指标和特征信息经标准化处理之后，每批目标针对各影响因子的隶属度及非隶属度数据如表1所列。

表1　聚类指标和特征信息表

目标＼特征	目标类型	距离	高度	速度	武器类型	干扰能力	航向角
X_1	〈0.50,0.50〉	〈0.19,0.79〉	〈0.91,0.06〉	〈0.55,0.41〉	〈0.10,0.86〉	〈0.05,0.90〉	〈0.05,0.90〉
X_2	〈0.50,0.49〉	〈0.26,0.72〉	〈0.86,0.12〉	〈0.45,0.51〉	〈0.10,0.86〉	〈0.13,0.85〉	〈0.12,0.85〉
X_3	〈0.50,0.47〉	〈0.63,0.36〉	〈0.60,0.38〉	〈0.57,0.42〉	〈0.10,0.86〉	〈0.67,0.32〉	〈0.32,0.64〉
X_4	〈0.50,0.46〉	〈0.57,0.42〉	〈0.41,0.56〉	〈0.38,0.61〉	〈0.10,0.86〉	〈0.68,0.32〉	〈0.52,0.47〉
X_5	〈0.50,0.42〉	〈0.19,0.86〉	〈0.94,0.03〉	〈0.55,0.41〉	〈0.10,0.88〉	〈0.05,0.93〉	〈0.05,0.93〉
X_6	〈0.50,0.45〉	〈0.29,0.71〉	〈0.83,0.13〉	〈0.45,0.51〉	〈0.10,0.88〉	〈0.13,0.84〉	〈0.12,0.84〉
X_7	〈0.50,0.43〉	〈0.45,0.52〉	〈0.60,0.37〉	〈0.57,0.41〉	〈0.10,0.86〉	〈0.67,0.32〉	〈0.32,0.65〉
X_8	〈0.50,0.44〉	〈0.51,0.47〉	〈0.41,0.56〉	〈0.38,0.62〉	〈0.10,0.86〉	〈0.68,0.32〉	〈0.52,0.47〉
X_9	〈0.50,0.44〉	〈0.19,0.80〉	〈0.90,0.05〉	〈0.55,0.41〉	〈0.10,0.86〉	〈0.05,0.91〉	〈0.05,0.91〉
X_{10}	〈0.50,0.45〉	〈0.26,0.72〉	〈0.89,0.10〉	〈0.47,0.51〉	〈0.10,0.86〉	〈0.13,0.84〉	〈0.12,0.84〉
X_{11}	〈0.10,0.83〉	〈0.44,0.53〉	〈0.20,0.76〉	〈0.13,0.81〉	〈0.50,0.48〉	〈0.67,0.32〉	〈0.18,0.82〉
X_{12}	〈0.10,0.84〉	〈0.50,0.50〉	〈0.27,0.70〉	〈0.08,0.92〉	〈0.50,0.48〉	〈0.53,0.46〉	〈0.08,0.88〉
X_{13}	〈0.10,0.86〉	〈0.46,0.52〉	〈0.18,0.80〉	〈0.19,0.78〉	〈0.50,0.48〉	〈0.65,0.31〉	〈0.10,0.90〉
X_{14}	〈0.10,0.88〉	〈0.49,0.50〉	〈0.17,0.83〉	〈0.15,0.81〉	〈0.50,0.48〉	〈0.63,0.33〉	〈0.12,0.83〉
X_{15}	〈0.10,0.85〉	〈0.38,0.61〉	〈0.11,0.87〉	〈0.12,0.86〉	〈0.50,0.48〉	〈0.67,0.32〉	〈0.05,0.92〉
X_{16}	〈0.10,0.87〉	〈0.46,0.52〉	〈0.26,0.71〉	〈0.18,0.81〉	〈0.50,0.48〉	〈0.56,0.42〉	〈0.12,0.86〉
X_{17}	〈0.90,0.03〉	〈0.63,0.31〉	〈0.44,0.53〉	〈0.43,0.52〉	〈0.10,0.86〉	〈0.68,0.32〉	〈0.42,0.55〉
X_{18}	〈0.10,0.84〉	〈0.43,0.55〉	〈0.26,0.70〉	〈0.11,0.85〉	〈0.10,0.88〉	〈0.55,0.41〉	〈0.12,0.87〉
X_{19}	〈0.90,0.07〉	〈0.56,0.42〉	〈0.54,0.43〉	〈0.48,0.51〉	〈0.10,0.86〉	〈0.63,0.35〉	〈0.32,0.65〉
X_{20}	〈0.90,0.05〉	〈0.63,0.34〉	〈0.52,0.43〉	〈0.44,0.55〉	〈0.10,0.86〉	〈0.63,0.34〉	〈0.52,0.44〉

通过初步分析，聚类类别数为3。下面检验基于密度函数法初始聚类中心的选择是否有效，计算可知3个初始聚类中心，如表2所列。

表2　初始聚类中心

	第一维	第二维	第三维	第四维	第五维	第六维	第七维
1	〈0.1,0.8529〉	〈0.4514,0.5329〉	〈0.2071,0.7671〉	〈0.1371,0.8343〉	〈0.4429,0.5371〉	〈0.6086,0.3671〉	〈0.11,0.8686〉
2	〈0.5,0.4583〉	〈0.23,0.7667〉	〈0.8883,0.0817〉	〈0.5033,0.46〉	〈0.1,0.8667〉	〈0.09,0.8783〉	〈0.085,0.8783〉
3	〈0.6714,0.2786〉	〈0.5686,0.4057〉	〈0.5029,0.4657〉	〈0.4643,0.52〉	〈0.1,0.86〉	〈0.6629,0.3271〉	〈0.42,0.55529〉

下面分别通过随机样本法与密度函数法获得初始聚类中心，进行直觉模糊聚类。试验中取平滑参数 $m=2$，$\varepsilon=10^{-5}$。为简单起见，不妨令 $\pi_{ij}=0$，由于各维特征的权重相同，故令 $\omega(k)=1$，比较结果如表3所列。

表3　聚类结果比较

	迭代次数	目标函数初始值	目标函数值
随机样本法	15	0.6632	0.0892/0.0905(极小值)
密度函数法	4	0.0897	0.0892(最小值)

可见，采用本文提出的基于密度函数的聚类初始化算法，可以极大地提高直觉模糊聚类算法性能，不仅解决了易陷入局部极小值的问题，同时减少了迭代次数，目标函数初始值较小，加快了收敛速度。

结束语　针对基于目标函数的直觉模糊聚类初始化问题，本文提出一种改进的密度函数初始化方法。利用样本密度函数在较高局部密度的区域中选取 c 个样本，并通过求其周围所有样本点的平均值来提取样本的聚类中心，其算法复杂度只与样本数有关，而与维数无关，尤其对高维数据可以取得良好的聚类效果。实验表明，用其初始化聚类中心可以有效地改善聚类效果，不仅解决了基于目标函数的直觉模糊聚类算法容易陷入局部极小值的问题，同时减少了迭代次数，使得收敛更快。

参考文献

[1] Atanassov K. Intuitionistic Fuzzy Sets[J]. Fuzzy Sets and Systems, 1986, 20(1): 87-96

[2] 雷英杰，王涛，赵晔，等. 直觉模糊匹配的语义距离与贴近度[J]. 空军工程大学学报，2005，6(1)：69-72

[3] 申晓勇，雷英杰，蔡茹，等. 基于目标函数的直觉模糊集合数据的聚类方法[J]. 系统工程与电子技术，2008

[4] 盛莉，邹开其，邓冠男. 基于网格和密度的模糊 c 均值聚类初始化方法[J]. 计算机应用与软件，2008，25(3)：22-23

[5] 宋清昆，郝敏. 一种改进的模糊 C 均值聚类算法[J]. 哈尔滨理工大学学报，2007，12(4)：8-10

[6] 牛琨，张舒博，陈俊亮. 融合网格密度的聚类中心初始化方案[J]. 北京邮电大学学报，2007，30(2)：6-10

[7] He Ji, Lan M, Tan C L, et al. Initialization of cluster refinement algorithms: a review and comparative study[C]// Proceedings of International Joint Conference on Neural Networks. Budapest: [s. n.], 2004: 297-302