文章编号：1000-8608(2012)05-0749-06

基于属性加权的不完全数模糊 c 均值聚类算法

李　丹*，顾　宏，张立勇

（大连理工大学 控制科学与工程学院，辽宁 大连　116024）

摘要： 针对现有的不完全数模糊聚类算法未考虑样本各维属性对聚类贡献不同的问题，提出了基于属性加权的不完全数模糊 c 均值聚类算法. 利用 ReliefF 算法评价各维属性的重要程度，通过加权欧式距离将属性权重结合入聚类，并能实现在聚类迭代过程中的缺失属性、隶属度及聚类中心的一体化求解. 实验结果表明，该算法强调了重要属性在不完全数模糊聚类中的作用，能够得到更为准确的聚类结果.

关键词： 模糊聚类；模糊 c 均值；属性加权；不完全数据；缺失属性

中图分类号： TP181　　**文献标志码：** A

0　引　言

聚类是数据挖掘、模式识别等领域的重要研究内容之一，在识别数据内在结构方面具有重要作用. 模糊 c 均值(fuzzy c-means, FCM)算法[1]是一种基于目标函数的有效的聚类算法，该算法能够将 p 维完整的目标数据集 $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}\subset\boldsymbol{R}^p$ 划分为若干模糊子类. 然而，在模式分类应用中，由于数据采集失败及随机噪声等原因，很多数据集是不完全的，即数据集包含缺失部分(非全部)属性值的样本，而 FCM 算法不能直接应用于不完全数据集的聚类分析.

为了消除缺失属性对聚类分析的影响，有关文献提出了各种不完全数据集的模糊聚类算法[2-8]. Miyamoto 等[2]提出了几种在模糊 c 均值算法中处理不完全数据的方法，其中一种简单方法是以相应属性的加权平均值估算缺失属性值，另一种方法则忽略数据集中的缺失属性并通过完备属性进行距离测算，这两种方法即后来解决不完全数据聚类问题的常用策略：估算(imputation)法及舍弃(discarding/ignoring)法. 之后，Hathaway 等[3]提出了基于 FCM 进行不完全数据聚类的几种具体方法，其中 WDS(whole data strategy)是一种最简单的方法，该方法舍弃样本集中所有包含缺失属性的样本，利用剩余完备样本得到聚类中心及完备样本的类属，不完全样本的类属则通过其与各聚类中心的局部距离[9]确定，但这种对数据集的约简将造成信息的大量丢失；另一种方法为 PDS(partial distance strategy)，即忽略不完全样本的缺失属性，FCM 算法中的距离仍采用 Dixon 提出的局部距离[9]计算；文献[3]还提出了两种依据数据集信息对不完全样本的缺失属性进行估算的方法，其中 OCS(optimal completion strategy)方法将缺失属性的估算看作优化问题，并在聚类迭代过程中逐步寻求更优的估计值，而 NPS(nearest prototype strategy)方法则在每次聚类迭代中将缺失属性设置为距离最近的聚类中心的相应属性值. 此外，考虑到样本集中属性的缺失原因，Timm 等提出了一种基于 Gath-Geva 算法的不完全数聚类方法[4]；对于不完全关系数据集，Hathaway 等根据三角不等式近似规则提出了一种基于 FCM 的聚类方法[5]；Honda 等则通过局部主成分分析法将不完全数据集划分到若干

收稿日期： 2011-02-12；　**修回日期：** 2012-06-22.
基金项目： 国家地震行业科研专项基金资助项目(200808075).
作者简介： 李　丹*(1977-)，女，博士，讲师，E-mail：ldan@dlut.edu.cn；顾　宏(1961-)，男，教授，博士生导师.

线性模糊子类中[6]；Lim 等提出通过神经网络训练缺失属性以解决不完全数聚类问题[7]. 鉴于不完全数据中缺失属性的不确定性，作者在前期工作中提出了缺失属性的最近邻区间描述[8]，通过将不完全数据集转化为区间型数据集求解不完全数据集的聚类问题.

上述算法均隐含假定待分析样本的各维属性对聚类的贡献均匀，在实际应用中有一定的局限性. 目前，针对完全数据集的属性加权聚类研究十分活跃，研究者已提出了多种属性加权模糊聚类算法[10-13]，通过极小化属性评价函数[10]、利用 ReliefF 算法[14]进行属性赋权[11]、聚类与属性权重协同学习[12]，以及对权重进行区间监督[13]等策略强调重要属性的作用，有效提高了完全数据集的聚类准确率. 本文将属性加权的思路引入不完全数据集的模糊聚类问题，利用 ReliefF 算法分析各维属性的聚类贡献并结合入模糊 c 均值聚类，进而改善不完全数据集的聚类效果.

1 属性加权模糊 c 均值聚类算法

属性加权模糊 c 均值(weighted fuzzy c-means, WFCM)算法[10-13]将完备数据集 $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}\subset\boldsymbol{R}^p$ 划分为 c 个模糊类，对于给定属性权重 $\boldsymbol{w}=(w_1 \quad w_2 \quad \cdots \quad w_p)^{\mathrm{T}}$，$\forall j: w_j>0$，通常 $\sum_{j=1}^{p} w_j=1$，使得聚类目标函数

$$J(\boldsymbol{U},\boldsymbol{V})=\sum_{i=1}^{c}\sum_{k=1}^{n}u_{ik}^{m}\|\boldsymbol{x}_k-\boldsymbol{v}_i\|_{\boldsymbol{w}}^{2} \tag{1}$$

取得极小值. 其中 $\boldsymbol{x}_k=(x_{1k} \quad x_{2k} \quad \cdots \quad x_{pk})^{\mathrm{T}}$ 为样本数据；$\boldsymbol{v}_i\in\boldsymbol{R}^p$ 为第 i 类的聚类中心，记 $\boldsymbol{V}=(v_{ji})\in\boldsymbol{R}^{p\times c}$ 为聚类中心矩阵；$u_{ik}\in[0,1]$ 表示样本 $\boldsymbol{x}_k$ 隶属于第 i 类的程度，且满足 $\sum_{i=1}^{c}u_{ik}=1$，记 $\boldsymbol{U}=(u_{ik})\in\boldsymbol{R}^{c\times n}$ 为模糊划分矩阵；$m>1$ 为模糊化参数；$\|\cdot\|_{\boldsymbol{w}}$ 表示如下加权欧式距离：

$$\|\boldsymbol{x}_k-\boldsymbol{v}_i\|_{\boldsymbol{w}}=\sqrt{(\boldsymbol{x}_k-\boldsymbol{v}_i)^{\mathrm{T}}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}(\boldsymbol{x}_k-\boldsymbol{v}_i)} \tag{2}$$

式中：$\boldsymbol{W}=\mathbf{diag}\{w_1,w_2,\cdots,w_p\}$ 为对角阵.

采用拉格朗日乘子法，可得聚类目标函数取得极小值的必要条件为

$$\boldsymbol{v}_i=\frac{\sum_{k=1}^{n}u_{ik}^{m}\boldsymbol{x}_k}{\sum_{k=1}^{n}u_{ik}^{m}};\ i=1,2,\cdots,c \tag{3}$$

$$u_{ik}=\left[\sum_{t=1}^{c}\left(\frac{\|\boldsymbol{x}_k-\boldsymbol{v}_i\|_{\boldsymbol{w}}^{2}}{\|\boldsymbol{x}_k-\boldsymbol{v}_t\|_{\boldsymbol{w}}^{2}}\right)^{\frac{1}{m-1}}\right]^{-1};$$

$$i=1,2,\cdots,c,\ k=1,2,\cdots,n \tag{4}$$

属性加权模糊 c 均值算法采用交替优化(alternating optimization, AO)策略对聚类目标函数(1)进行优化，当各属性权重相等时，算法为基本 FCM 算法.

2 ReliefF 属性评价算法

Relief 系列算法是基于样本类内相似性及类间相异性的属性评价方法[14,15]，最初的 Relief 算法仅局限于求解两类的分类问题[15]，后扩展为 ReliefF 算法[14]以解决有噪声及多类问题的属性评价.

设 $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}\subset\boldsymbol{R}^p$ 为待评价的完整数据集，$\boldsymbol{x}_k=(x_{1k} \quad x_{2k} \quad \cdots \quad x_{pk})^{\mathrm{T}}$，为样本数据，对于任意选取的基准样本 $\boldsymbol{x}_k(1\leqslant k\leqslant n)$，ReliefF 算法首先在同类中搜索其 z 个最近邻样本 $\boldsymbol{h}_r(r=1,2,\cdots,z)$，然后在各不同类中分别搜索其 z 个最近邻样本 $m_{lr}(r=1,2,\cdots,z;l\neq class(\boldsymbol{x}_k))$，并通过下式度量基准样本 $\boldsymbol{x}_k$ 与近邻样本 $\boldsymbol{x}_b$ 在第 j 个属性上的差异：

$$diff(j,\boldsymbol{x}_k,\boldsymbol{x}_b)=\frac{|x_{jk}-x_{jb}|}{\max(\boldsymbol{\lambda}_j)-\min(\boldsymbol{\lambda}_j)} \tag{5}$$

其中 $\boldsymbol{\lambda}_j=(x_{j1} \quad x_{j2} \quad \cdots \quad x_{jm})^{\mathrm{T}}$，为数据集中各样本第 j 个属性组成的向量，$\max(\boldsymbol{\lambda}_j)$ 和 $\min(\boldsymbol{\lambda}_j)$ 分别表示各样本第 j 个属性取值的最大及最小值. 通过取 n_{r} 个随机基准样本，ReliefF 算法中权值由下式更新[14]：

$$w_j:=w_j-\sum_{r=1}^{z}diff(j,\boldsymbol{x}_k,\boldsymbol{h}_r)/(n_{\mathrm{r}}\cdot z)+$$

$$\sum_{l\neq class(\boldsymbol{x}_k)}\left[\frac{P(l)}{1-P(class(\boldsymbol{x}_k))}\times\right.$$

$$\left.\sum_{r=1}^{z}diff(j,\boldsymbol{x}_k,m_{lr})/(n_{\mathrm{r}}\cdot z)\right] \tag{6}$$

其中 $P(l)$ 为第 l 类出现的概率.

3　基于属性加权的不完全数模糊 c 均值聚类算法

本文采用 ReliefF 算法获得属性权重向量 $\boldsymbol{w}$ 用于无监督的不完全数据加权聚类分析，但该算法是一种有监督学习模式下的属性评价方法，即在已知样本类属的基础上评价属性的重要程度. 针对完全数据集，文献[11]将有监督的 ReliefF 算法应用于无监督聚类分析，本文借鉴文献[11]的方法，采用如下针对不完全数据集的属性加权策略：鉴于经典的 OCS-FCM 算法能够同时获取数据集划分及对缺失属性的估算，因而，可首先对不完全数据集利用 OCS-FCM 算法进行聚类，获得样本类属及由缺失属性估算值“还原”的完备数据集，则在此基础上可通过 ReliefF 算法实现对属性较为合理的评价，使得将属性加权的思路引入不完全数据集的模糊聚类问题成为可能.

在已确定的属性权重基础上，本文采用如式(2)所示的加权欧式距离作为距离度量，将属性加权引入经典的 OCS-FCM 算法，提出了基于属性加权的不完全数模糊 c 均值(WOCS-FCM)聚类算法. 令 $\widetilde{\boldsymbol{X}}=\{\widetilde{\boldsymbol{x}}_1,\widetilde{\boldsymbol{x}}_2,\cdots,\widetilde{\boldsymbol{x}}_n\}$ 为不完全数据集，WOCS-FCM 算法将缺失属性作为变量优化如下目标函数：

$$J(\boldsymbol{U},\boldsymbol{V},\widetilde{\boldsymbol{X}})=\sum_{i=1}^{c}\sum_{k=1}^{n}u_{ik}^{m}\|\widetilde{\boldsymbol{x}}_k-\boldsymbol{v}_i\|_{\boldsymbol{w}}^{2} \tag{7}$$

借鉴文献[3]，可得使得目标函数(7)达到极小值的必要条件为式(3)、(4)以及下式：

$$\widetilde{x}_{jk}=\frac{\sum_{i=1}^{c}u_{ik}^{m}v_{ji}}{\sum_{i=1}^{c}u_{ik}^{m}} \tag{8}$$

WOCS-FCM 算法流程如下：

(1) 利用 ReliefF 算法得属性权重向量 $\boldsymbol{w}=(w_1\quad w_2\quad\cdots\quad w_p)^{\mathrm{T}}$；

(2) 设定聚类类别数 c，设定迭代停止阈值 ε，初始化缺失属性及划分矩阵 $\boldsymbol{U}^{(0)}$；

(3) 当迭代次数为 $l(l=1,2,\cdots)$ 时，根据 $\boldsymbol{U}^{(l-1)}$，利用式(3)更新聚类原型 $\boldsymbol{V}^{(l)}$；

(4) 根据 $\boldsymbol{V}^{(l)}$，利用式(4)更新划分矩阵 $\boldsymbol{U}^{(l)}$；

(5) 利用式(8)估算缺失属性 $\widetilde{x}_{jk}$；

(6) 若 $\forall i,k;\max|u_{ik}^{(l)}-u_{ik}^{(l-1)}|<\varepsilon$，则算法停止，输出划分矩阵 $\boldsymbol{U}$ 和聚类原型 $\boldsymbol{V}$；否则 $l=l+1$，返回(3).

4　数值实验

4.1　数据集

本文对两个著名数据集 IRIS[16] 及 Crude-Oil[17] 进行了聚类分析，这两个数据集常被用作检验聚类算法性能的标准数据.

IRIS 数据集由 150 个样本组成，每个样本有 4 个属性，分别表示 IRIS 的 Petal Length、Petal Width、Sepal Length 和 Sepal Width. 整个样本集包含了 3 个 IRIS 种类，分别为 Setosa、Versicolor 和 Virginica，每类各有 50 个样本，其中 Setosa 与其他两类间能较好地分离，而 Versicolor 和 Virginica 之间存在交叠. Hathaway 等[18] 给出这组数据的实际类原型位置为

$$\boldsymbol{V}^{*}=\begin{pmatrix}5.00 & 5.93 & 6.58\\ 3.42 & 2.77 & 2.97\\ 1.46 & 4.26 & 5.55\\ 0.24 & 1.32 & 2.02\end{pmatrix} \tag{9}$$

Crude-Oil 数据集由 56 个样本组成，每个样本有 5 个属性，用于描述产自 3 个地区原油的化学成分. 数据集中，7 个样本来自 Wilhelm，另各有 11 个和 38 个样本分别来自 Sub Mulinia 及 Upper Mulinia.

本文讨论属性随机缺失(missing completely at random，MCAR)不完全数据集的模糊聚类问题. 通过人为随机舍弃完备数据集 $\boldsymbol{X}$ 中若干属性可得不完全数据集 $\widetilde{\boldsymbol{X}}$，缺失属性的随机选取应满足下述条件[3]：

(1)数据集中任意样本 $\widetilde{\boldsymbol{x}}_k$ 至少保留一个完备属性；

(2)任意属性在数据集中至少保留一个完备值.

4.2 实验结果

为了测试 WOCS-FCM 算法的聚类性能，将其与 WDS-FCM、PDS-FCM、OCS-FCM 及 NPS-FCM 算法[3]对不完全 IRIS、Crude-Oil 数据集的聚类结果进行比较. 本文中，各种聚类算法均选取模糊化参数 $m=2$，迭代停止阈值 $\varepsilon=10^{-5}$，使用满足 $\sum_{i=1}^{c}u_{ik}=1$ 的划分矩阵 $\boldsymbol{U}^{(0)}$ 作为初始值，相应的算法结束条件设置为 $\|\boldsymbol{U}^{(l)}-\boldsymbol{U}^{(l-1)}\|<\varepsilon$，并将 OCS-FCM、NPS-FCM 及 WOCS-FCM 算法中缺失属性的初始值设置为随机值. 在 ReliefF 算法中，选取所有样本为基准样本评价属性权重，对于 IRIS 数据集，ReliefF 算法搜索基准样本的近邻样本数 $z=10$；由于 Crude-Oil 数据集样本较少，近邻样本数取为 $z=\min\{$各类中样本数目最小值，$5\}$.

在实验过程中发现，在相同的属性缺失程度(如 IRIS 数据缺失 5%的属性值)下，数据集缺失不同的属性将可能造成聚类结果的差异. 因此，为避免这种差异对算法性能评价的影响，本文在每种属性缺失程度下依 4.1 节中所述方法随机生成 10 个不完全数据集，取其聚类结果平均值用于算法分析.

以 IRIS 数据缺失 5%属性值的情况为例，图 1 所示为在随机生成的 10 个不完全数据集(集 1～10)上 ReliefF 算法所得各维属性权值.

由图 1 可以看出，当 IRIS 数据缺失 5%的属性值时，在随机生成的 10 个数据集中，ReliefF 算法确定属性 Sepal Length 及 Sepal Width 的权值均较大，而属性 Petal Length 及 Petal Width 的权值相对较小. 因而，在 WOCS-FCM 算法中，属性 Sepal Length 及 Sepal Width 的作用将通过如式(2)所示的加权欧式距离被强调，进而引导聚类过程. 类似地，在其他缺失程度的不完全 IRIS 数据集及不完全 Crude-Oil 数据集上，重要属性也将获得较大权重以强调其在聚类过程中的作用.

表 1 及表 2 统计了两个数据集在不同属性缺失程度下 10 个不完全数据集的聚类结果平均值. 由于已知如式(9)所示的 IRIS 实际类原型位置 $\boldsymbol{V}^{*}$，表 1 的后 5 列为各算法的聚类中心误差平方和，其计算如下[3]：

$$\|\boldsymbol{V}-\boldsymbol{V}^{*}\|_F^2=\sum_{j=1}^{p}\sum_{i=1}^{c}(v_{ji}-v_{ji}^{*})^2 \qquad (10)$$

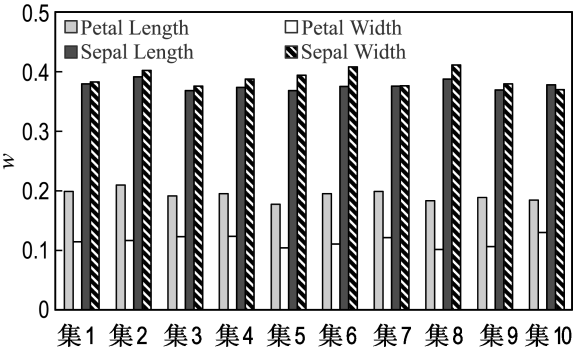

图 1 IRIS 数据缺失 5%属性值时在 10 个数据集上所得各维属性权重

Fig. 1 The attribute weights on 10 incomplete IRIS datasets with 5% attribute values missing

表 1 不完全 IRIS 数据集的聚类结果平均值

Tab. 1 Averaged results of clustering using incomplete IRIS datasets

属性缺失比例/%	平均迭代次数					平均错分数					聚类中心误差平方和				
	WDS	PDS	OCS	NPS	WOCS	WDS	PDS	OCS	NPS	WOCS	WDS	PDS	OCS	NPS	WOCS
0	25.2	25.1	26.3	25.4	22.9	16.0	16.0	16.0	16.0	6.0	0.050	0.050	0.050	0.050	0.038
5	25.4	25.2	30.2	26.8	23.3	16.6	16.4	16.5	16.3	6.4	0.069	0.057	0.055	0.057	0.038
10	25.7	24.9	43.7	29.0	22.5	16.4	16.7	16.3	16.5	7.9	0.080	0.051	0.049	0.052	0.035
15	26.6	23.5	32.0	27.8	23.6	16.6	16.6	17.0	16.2	8.1	0.094	0.066	0.063	0.062	0.043
20	25.9	25.6	41.2	29.3	26.4	18.2	18.0	18.1	18.0	10.5	0.214	0.074	0.070	0.071	0.045

表 2　不完全 Crude-Oil 数据集的聚类结果平均值

Tab. 2　Averaged results of clustering using incomplete Crude-Oil datasets

属性缺失比例/%	平均迭代次数					平均错分数				
	WDS	PDS	OCS	NPS	WOCS	WDS	PDS	OCS	NPS	WOCS
0	33.7	31.7	33.0	33.6	46.3	23.0	23.0	23.0	23.0	21.0
5	41.7	33.9	35.6	36.2	39.3	21.7	21.4	21.4	21.4	20.6
10	40.1	37.1	40.8	38.8	40.2	21.0	21.6	22.2	22.1	20.0
15	38.3	39.5	62.5	43.3	42.3	21.5	21.2	22.1	21.6	18.7
20	37.6	37.7	47.9	44.7	47.2	21.8	21.6	22.4	22.5	20.1

由表 1、2 可以看出，在不完全 IRIS 及 Crude-Oil数据集的不同属性缺失程度下，基于属性加权的 WOCS-FCM 算法在迭代次数方面与其他 4 种算法基本相当；而在两个数据集的错分数及 IRIS 数据集聚类中心误差平方和方面，WOCS-FCM 算法通过强调重要属性的贡献得到了更合理的数据集划分，改善了不完全数据集的聚类效果. 以上结果表明本文提出的基于属性加权的模糊 c 均值聚类算法是合理有效的.

5　结　语

本文从强化重要属性在不完全数据集聚类分析中作用的角度考虑，提出了一种基于属性加权的不完全数模糊 c 均值聚类算法. 该算法在 ReliefF 算法获得的属性权重基础上，将属性加权思路引入不完全数据集模糊聚类，能够实现对缺失属性及聚类结果的一体化求解. 实验结果表明了本文所提算法的有效性和相比已有结果的优越性. 下一步工作是将属性加权思路应用于部分缺失图像的处理中.

参考文献：

[1] Bezdek J C. **Pattern Recognition with Fuzzy Objective Function Algorithms** [M]. New York: Plenum Press, 1981.

[2] Miyamoto S, Takata O, Uayahara K. Handling missing values in fuzzy c-means [C] // **Proceedings of the 3rd Asian Fuzzy System Symposium**. Seoul: Packsan Publishers Inc., 1998: 139-142.

[3] Hathaway R J, Bezdek J C. Fuzzy c-means clustering of incomplete data [J]. **IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics**, 2001, **31**(5): 735-744.

[4] Timm H, Doring C, Kruse R. Different approaches to fuzzy clustering of incomplete datasets [J]. **International Journal of Approximate Reasoning**, 2004, **35**(3): 239-249.

[5] Hathaway R J, Bezdek J C. Clustering incomplete relational data using the non-Euclidean relational fuzzy c-means algorithm [J]. **Pattern Recognition Letters**, 2002, **23**(1-3): 151-160.

[6] Honda K, Ichihashi H. Linear fuzzy clustering techniques with missing values and their application to local principle component analysis [J]. **IEEE Transactions on Fuzzy System**, 2004, **12**(2): 183-193.

[7] Lim C P, Leong J H, Kuan M M. A hybrid neural network system for pattern classification tasks with missing features [J]. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, 2005, **27**(4): 648-653.

[8] LI Dan, GU Hong, ZHANG Li-yong, *et al*. A fuzzy c-means clustering algorithm based on nearest-neighbor intervals for incomplete data [J]. **Expert Systems with Applications**, 2010, **37**(10): 6942-6947.

[9] Dixon J K. Pattern recognition with partly missing data [J]. **IEEE Transactions on Systems, Man, and Cybernetics**, 1979, **9**(10): 617-621.

[10] WANG X Z, WANG Y D, WANG L J. Improving fuzzy c-means clustering based on feature-weight learning [J]. **Pattern Recognition Letters**, 2004, **25**(10): 1123-1132.

[11] 李　洁，高新波，焦李成. 基于特征加权的模糊聚类

新算法[J]. 电子学报, 2006, **34**(1):89-92.
LI Jie, GAO Xin-bo, JIAO Li-cheng. A new feature weighted fuzzy clustering algorithm [J]. **Acta Electronica Sinica**, 2006, **34**(1):89-92. (in Chinese)

[12] ZHANG Li-yong, LI Dan, ZHONG Chong-quan. Collaborative optimization of clustering by fuzzy *c*-means and weight determination by ReliefF [C] // **Proceedings of the 6th International Conference on Fuzzy Systems and Knowledge Discovery**. Minneapolis: IEEE Computer Society Press, 2009: 454-459.

[13] 李 丹,顾 宏,张立勇. 基于属性权重区间监督的模糊 *C* 均值聚类算法[J]. 控制与决策, 2010, **25**(3):457-460.
LI Dan, GU Hong, ZHANG Li-yong. A fuzzy *c*-means algorithm with interval-supervised attribute weights [J]. **Control and Decision**, 2010, **25**(3): 457-460. (in Chinese)

[14] Kononenko I. Estimating attributes: Analysis and extensions of Relief [C] // **Proceedings of the 7th European Conference on Machine Learning**. Catania: Springer-Verlag, 1994:171-182.

[15] Kira K, Rendell L A. A practical approach to feature selection [C] // **Proceedings of the 9th International Conference on Machine Learning**. San Mateo: Morgan Kaufmann Publishers, 1992: 249-256.

[16] Blzke C L, Merz C J. UCI repository of machine learning databases [EB/OL]. [2010-02-12]. http://www. ics. uci. edu/~ mlearn. MLResitory. html.

[17] Johnson R A, Wichern D W. **Applied Multivariate Statistical Analysis** [M]. New Jersey: Prentice-Hall, 1982.

[18] Hathaway R J, Bezdek J C. Optimization of clustering criteria by reformulation [J]. **IEEE Transactions on Fuzzy Systems**, 1995, **3**(2): 241-245.

An attribute weighted fuzzy *c*-means algorithm for incomplete data clustering

LI Dan*, GU Hong, ZHANG Li-yong

(School of Control Science and Engineering, Dalian University of Technology, Dalian 116024, China)

Abstract: In view of the problem that the existing algorithms for incomplete data fuzzy clustering generally view each dimensional attribute as equally important in contribution of clustering, an attribute weighted fuzzy *c*-means algorithm for incomplete data clustering is proposed. In the proposed algorithm, the important degree of each dimensional attribute is evaluated by the ReliefF algorithm and combined into fuzzy clustering by weighted Euclidean distance, and missing attribute values, membership and clustering centers can be obtained simultaneously. The experimental results show that the proposed algorithm can emphasize the important attributes in clustering, and better clustering results can be obtained.

Key words: fuzzy clustering; fuzzy *c*-means; attribute weighting; incomplete data; missing attribute