基于属性权重的 Fuzzy C Mean 算法

王丽娟[1),2)] 关守义[3)] 王晓龙[1)] 王熙照[2)]

[1)](哈尔滨工业大学计算机科学与技术学院 哈尔滨 150001)

[2)](河北大学数学与计算机学院 保定 071002)

[3)](河北师范大学学位办公室 石家庄 050016)

摘 要 提出 CF-WFCM 算法,该算法分为属性权重学习算法和聚类算法两部分.属性权重学习算法,从数据自身的相似性出发,通过梯度递减算法极小化属性评价函数 $CFuzziness(w)$,为每个属性赋予一个权重.将属性权重应用于 Fuzzy C Mean 聚类算法,得到 CF-WFCM 算法的聚类算法.CF-WFCM 算法强化重要属性在聚类过程中的作用,消减冗余属性的作用,从而改善聚类的效果.我们选取了部分 UCI 数据库进行实验,实验结果证明:CF-WFCM 算法的聚类结果优于 FCM 算法的聚类结果.函数 $CFuzziness(w)$不仅可以评价属性的重要性,而且可以评价属性评价函数的优劣.实验说明了这一问题.最后我们对 CF-WFCM 算法进行了讨论.

关键词 梯度递减算法;Fuzzy C Mean 算法;属性权重学习算法;聚类有效性函数

中图法分类号 TP18

Fuzzy C Mean Algorithm Based on Feature Weights

WANG Li-Juan[1),2)] GUAN Shou-Yi[3)] WANG Xiao-Long[1)] WANG Xi-Zhao[2)]

[1)](*School of Computer Science and Technology, Harbin Institute of Technology, Harbin* 150001)

[2)](*School of Mathematics and Computer Science, Hebei University, Baoding* 071002)

[3)](*Office of Academic Degrees Committee, Hebei Normal University, Shijiazhuang* 050016)

Abstract This paper proposes CF-WFCM algorithm including feature weight learning algorithm and clustering algorithm. According to data's similarity, feature weight learning algorithm gives each feature a feature weight by minimizing the feature evaluation index $CFuzziness(w)$ through gradient descent technique. When the feature weight is applied in the Fuzzy C Mean (FCM) clustering algorithm, it forms the clustering algorithm of CF-WFCM algorithm. CF-WFCM emphasizes the important feature's effect and lessens the redundant feature's effect in the procedure of clustering so that the performance of clustering has been improved. Experiments on some UCI databases show that the result of CF-WFCM is better than that of FCM. In addition, the index $CFuzziness(w)$ not only can be used to learn feature weight, but also is a valid entropy function to evaluate the feature evaluation indexes. If we can choose a better validity index to learn the feature weight before clustering, large computation will be avoided, which is showed in an example. In the end, the authors discuss the CF-WFCM algorithm.

Keywords gradient descent algorithm; Fuzzy C Mean algorithm; feature weight learning algorithm; cluster validity index

---

收稿日期:2004-11-09;修改稿收到日期:2006-04-18.本课题得到两项国家自然科学基金(60435020,60473045)和河北省自然科学基金(603137)资助.**王丽娟**,女,1978 年生,博士研究生,主要研究方向为属性选取、聚类算法、数据挖掘.E-mail:bssswlj@126.com.**关守义**,男,1970 年生,政工师,主要研究方向为数据挖掘、教育学.**王晓龙**,男,1955 年生,教授,博士生导师,主要研究领域为人工智能、机器学习、计算语言学、中文信息处理.**王熙照**,男,1963 年生,教授,博士生导师,主要研究领域为机器学习.

1 引 言

Fuzzy C Mean(FCM)算法在无监督的情况下,根据数据自身分布,将数据划分到不同的集合中.自从 1974 年 Dunn[1] 和 Bezedk[2] 提出 FCM 算法后,众多学者在这一领域进行了大量的研究,其中一个研究重点为度量方式的改进[3~5].

FCM 算法基于传统的欧式距离.欧式距离假设每一个属性在聚类过程中的重要性均相同,即数据空间为球形空间,在这种情况下 FCM 算法可以得到较好的聚类结果;而实际情况中,有些属性在聚类过程中发挥重要的作用,有些属性的作用次要甚至可以忽略,那么数据空间可能呈现椭球形,在这种情况下 FCM 算法得不到较好的聚类结果[3].文献[3]提出 WFCM 算法,该算法通过极小化属性评价函数 $E(w)$ 为每个属性学习权重,构造加权的欧式距离,得到了较好的结果.

本文在文献[3]的基础上,通过梯度递减算法极小化属性评价函数 $CFuzziness(w)$[6],为每个属性赋予一个权重;权重和欧式距离结合,得到基于属性权重的欧式距离;将基于属性权重欧式距离应用于 FCM 算法,得到 CF-WFCM 算法.CF-WFCM 算法强化重要属性在聚类过程中的作用,消减冗余属性的作用,从而改善聚类的效果.我们选取了部分 UCI 数据库进行实验,实验结果证明:CF-WFCM 算法的聚类结果优于 FCM 算法的聚类结果.除此之外,本文比较了函数 $E(w)$ 构造的 WFCM 算法和函数 $CFuzziness(w)$ 构造的 CF-WFCM 算法,实验结果表明:这两个函数构造的聚类算法均优于 FCM 算法;而 CF-WFCM 算法的聚类结果略优于 WFCM 算法的结果.

函数 $CFuzziness(w)$ 不仅可以评价属性的重要性,而且可以评价属性评价函数的优劣[7].在聚类过程之前,用函数 $CFuzziness(w)$ 度量众多属性评价函数的模糊度,从中选取模糊度较小的属性评价函数优化聚类,避免了属性评价函数选取不当造成的大量计算.

本文首先介绍 FCM 算法及其有效性函数,接着定义数据的相似性度量、分析属性评价函数 $CFuzziness(w)$ 的性质,同时介绍 CF-WFCM 算法的属性权重学习算法及其聚类算法;第 4 节介绍如何根据 $CFuzziness(w)$ 函数值选取较优的属性评价函数;第 5 节比较 CF-WFCM 算法、WFCM 算法和 FCM 算法聚类结果;最后对 CF-WFCM 算法进行讨论.

2 FCM 算法及其评价函数

2.1 FCM 算法

FCM[8] 算法把 n 个数据 $\boldsymbol{X}=(X_1,X_2,\cdots,X_n)$ 分为 c 个模糊组,并求每组的聚类中心 $v_i(i=1,2,\cdots,c)$,使得价值函数 J 达到最小.价值函数 J 的定义如下:

$$J(\boldsymbol{U},v_1,v_2\cdots,v_c;X)=\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^{m}d_{ij}^{2} \tag{1}$$

并且需要满足下式:

$$\sum_{i=1}^{c}u_{ij}=1,\quad \forall j=1,2,\cdots,n \tag{2}$$

这里 $u_{ij}\in[0,1]$ 表示第 j 个数据点属于第 i 个聚类中心的隶属度;每个数据点与相应聚类中心的隶属度构成了隶属矩阵 $\boldsymbol{U}$;v_i 为第 i 个模糊聚类中心;d_{ij} 为第 i 个聚类中心与第 j 个数据点的欧几里德距离;$m\in[1,\infty)$ 是一个加权指数,随着 m 的增大,聚类的模糊性增大.在本文中 $m=2$.对所有输入参量求导,使式(1)达到最小同时满足式(2)的必要条件为

$$v_i=\frac{\sum_{j=1}^{n}u_{ij}^{2}x_j}{\sum_{j=1}^{n}u_{ij}^{2}} \tag{3}$$

$$u_{ij}=\frac{1}{\sum_{k=1}^{c}\left(\frac{d_{ij}}{d_{kj}}\right)^{2}} \tag{4}$$

算法的描述如下:

1. 用值在区间[0,1]内的随机数初始化隶属矩阵 $\boldsymbol{U}$,使其满足约束条件式(2);

2. 根据式(3)计算 c 个聚类中心 v_i,$i=1,2,\cdots,c$;

3. 根据式(1)计算价值函数 J,如果相对上次价值函数值的改变量小于某个阈值 ε,则算法停止;

4. 根据式(4)更新 $\boldsymbol{U}$ 阵,返回步 2.

在上述算法中,有两个参数需要事先指定:阈值 ε 和聚类个数 c.其中阈值 ε 可以人为指定一个较小的数值,只要能够满足所要求的精度即可.聚类个数 c 通常根据先验知识确定其大体范围,在该范围内逐一计算不同聚类个数 c 所对应聚类结果的有效性函数值,从中选取出较优的 c 值.

2.2 FCM 算法的有效性函数

聚类有效性函数可以评价不同聚类算法的结果

以及同一算法在不同参数情况下得到的聚类结果.在这里我们希望通过聚类有效性函数：(1)比较FCM算法在不同聚类个数情况下得到的结果，从中选择较优的聚类，作为最终的结果；而最终结果对应的聚类个数被确定为较优的 c 值；(2)比较 FCM 算法及其优化算法得到的聚类结果.

与 FCM 算法有关的有效性函数分成两类.一类有效性函数基于数据集的模糊划分，其基本观点是：一个能较好分类的数据集应该是较“分明”的，因此，模糊划分的模糊性越小，聚类结果越好.这一类的代表函数如：Bezdek 提出的分割系数[9](partition coefficient)F 和分割熵[9](partition entropy)H.另一类有效性函数基于数据集的几何结构，其基本观点是：一个能较好分类的数据集的每一个子类应该是紧致的，且子类与子类尽可能分离，以紧致性和分离性作为聚类的有效性标准.这一类的代表函数如：Xie 和 Beni 在 1991 年定义的紧致分离函数 CS(Compactness Separability)[10].

上面提到的 3 个有效性函数定义和性质如表 1 所示.在评价聚类结果时，这 3 个有效性函数并不一定同时达到最优，通常认为：多个有效性函数取得最优值的聚类结果为较优的结果，该结果对应的聚类个数 c 为较优的聚类个数.

表 1　聚类有效性函数

有效性函数名称	函数描述	最优的聚类以及聚类个数
分割系数(partition coefficient)	$F=\dfrac{\sum_{j=1}^{n}\sum_{i=1}^{c}u_{ij}^2}{n}$	$\max(F,\boldsymbol{U},c)$
分割熵(partition entropy)	$H=-\dfrac{1}{n}\left\{\sum_{j=1}^{n}\sum_{i=1}^{c}[u_{ij}\log u_{ij}]\right\}$	$\min(H,\boldsymbol{U},c)$
紧致分离函数(compactness separability)	$CS(\boldsymbol{U},V;X)=\dfrac{\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^2\lVert X_j-v_i\rVert^2}{n\times(\min\limits_{i\neq k}\{\lVert v_i-v_k\rVert^2\})}=\left[\dfrac{\left[\frac{\sigma}{n}\right]}{Sep(V)}\right]$	$\min(CS,\boldsymbol{U},c)$

3　CF-WFCM 算法

3.1　相似性度量

通过比较数据间的相似性，能够为每个属性赋予不同的重要性.相似性度量的方法如欧氏距离、相关系数法、两夹角余弦等.本文定义一种灵活可变的基于欧式距离的相似性度量[3,6]，定义如下：

$$\rho_{pq}^{(w)}=\frac{1}{1+\beta\times d_{pq}^{(w)}} \tag{5}$$

β 是[0,1]的一个常数，我们希望通过调整 β 能够使得 $\rho_{pq}^{(w)}$ 的值近似均匀地分布在[0,1]内，因而 β 可按下式确定：

$$\frac{2}{n(n-1)}\sum_{p<q}\frac{1}{1+\beta\times d_{pq}}=0.5 \tag{6}$$

式(6)中 d_{pq} 为普通的欧氏距离，而式(5)中 $d_{pq}^{(w)}$ 是基于属性权重的欧式距离，定义如下：

$$d_{pq}^{(w)}=\sqrt{\left(\sum_{k=1\cdots m}w_k^2(x_{pk}-x_{qk})^2\right)} \tag{7}$$

式(7)中 $\boldsymbol{w}=(w_1,w_2,\cdots,w_m)$ 是与输入属性相对应的一个权重矢量，$w_k\in[0,1]$ 描述第 k 维属性在聚类中的重要性[3,6].通过调整 w_k 的值，数据 p 与数据 q 间的相似度在不断地变化.当 $w=(1,\cdots,1)_m$ 时，数据位于原始空间，即球形空间[3]，所有属性的重要性相等，数据间的相似度基于普通的欧式距离 d_{pq}，此时的相似度 $\rho_{pq}^{(w)}$ 记做 ρ_{pq}，称 ρ_{pq} 为原始相似度；当权重分量不全等于 1 时，数据处于压缩后的空间，即椭球空间[3]，$\rho_{pq}^{(w)}$ 不再等于 ρ_{pq}，$\rho_{pq}^{(w)}$ 称为基于属性权重相似度.当 $w_k=1$ 时，第 k 维属性对应的坐标轴在数据空间中不变，在聚类过程中发挥其全部的作用；w_k 的值越小，该属性对应的坐标轴被压缩的程度也就越大，在聚类过程中发挥的作用就越小；当 $w_k=0$ 时，该属性从数据空间中消失，在聚类过程中不发挥任何作用.

3.2　函数 *CFuzziness*(*w*)性质及其 CF-WFCM 算法的属性权重学习算法

文献[3,6]根据信息论的观点得到如下结论：当数据间的相似度在 0.5 附近时，模糊性较大；当数据间的相似度远离 0.5 时，模糊性较小.事实上，一个好的聚类结果应该具有模糊性小的性质，即数据间的相似度应该远离 0.5.我们希望通过调整属性权重，相似数据间的距离靠近，不相似数据间的距离拉大，即 $\rho_{pq}^{(w)}\to 0$ 或 1，这样得到的聚类结果模糊性小，

而且类内相似性大,类间相似性小.基于以上分析,本文提出用函数 $CFuzziness(w)$ 作为属性的评价函数.函数 $CFuzziness(w)$ 定义[6]如下:

$$CFuzziness(w)=\frac{-2}{n(n-1)}\sum_{q<p}\frac{1}{2}(\rho_{pq}^{(w)}\times\log\rho_{pq}+(1-\rho_{pq}^{(w)})\times\log(1-\rho_{pq})) \quad (8)$$

当 $\rho_{pq}^{(w)}=1$ 时,定义 $CFuzziness(w)=0$.函数 $CFuzziness(w)$ 满足 Deluca 提出的熵函数的几条性质[7,11],即

(i) $\rho_{pq}^{(w)}=\rho_{pq}=0.5$,函数 $CFuzziness(w)$ 取得最大值,即 $CFuzziness(w)=1$.

(ii) $(\rho_{pq}^{(w)}=\rho_{pq})=0$ 或 1,函数 $CFuzziness(w)$ 取得最小值,即 $CFuzziness(w)=0$.

(iii) 对于任两个整数 p,q,有两个模糊集 S 和 S',如果 $\rho_{pq}^{(w)}(s)\geq\rho_{pq}^{(w)}(s')\geq0.5$ 或 $\rho_{pq}^{(w)}(s)\leq\rho_{pq}^{(w)}(s')\leq0.5$,那么 $CFuzziness(w_{S'})\geq CFuzziness(w_S)$;对 ρ_{pq} 同理.

(iv) 当 $0<\rho_{pq}^{(w)},\rho_{pq}<0.5$,函数 $CFuzziness(w)$ 单调递增;当 $1>\rho_{pq}^{(w)},\rho_{pq}>0.5$,函数 $CFuzziness(w)$ 单调递减.

因而,函数 $CFuzziness(w)$ 可以作为有效的熵函数度量空间变换模糊度[7].在第 4 节中我们将详细讨论这一点.同时,函数 $CFuzziness(w)$ 还具有如下的性质:

$$\lim_{[\rho_{pq}^{(w)}\to0,\rho_{pq}<0.5]\text{或}[\rho_{pq}^{(w)}\to1,\rho_{pq}>0.5]} CFuzziness(w)=\min(CFuzziness(w)) \quad (9)$$

极小化函数 $CFuzziness(w)$ 得到的属性权重使得相似的数据($\rho_{pq}>0.5$)更相似($\rho_{pq}^{(w)}\to1$),不相似的数据($\rho_{pq}<0.5$)更不相似($\rho_{pq}^{(w)}\to0$).由于 $\rho_{pq}^{(w)}\to0$ 或 1,所以学习后的数据分布清晰,模糊性小,聚类结果能够改善.

极小化函数 $CFuzziness(w)$ 学习属性权重的算法即 CF-WFCM 算法的属性权重学习算法,本文选用文献[6]的梯度递减算法极小化该函数,详细内容参考文献[6].

3.3 CF-WFCM 算法的聚类算法

属性权重学习算法为待聚类数据的每个属性赋予权重;通过 FCM 聚类算法将未知数据划分到不同的类中.由于权重的引入,FCM 算法的度量方式由传统的欧式距离变成基于属性权重欧式距离,因而 FCM 算法的计算公式需要重新定义.首先算法中的新的价值函数 J[3]定义如下:

$$J(\boldsymbol{U},v_1,v_2,\cdots,v_c)=\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^2(d_{ij}^{(w)})^2 \quad (10)$$

由于 FCM 算法要求每个数据对所有类的隶属度和为 1,即满足式(2),在满足式(2)的条件下,极小化式(10)得到聚类中心和隶属度矩阵的计算公式[3]:

$$v_i=\frac{\sum_{j=1}^{n}u_{ij}^2x_j}{\sum_{j=1}^{n}u_{ij}^2} \quad (11)$$

$$u_{ij}=\frac{1}{\sum_{k=1}^{c}\left(\frac{d_{ij}^{(w)}}{d_{kj}^{(w)}}\right)^2} \quad (12)$$

根据上面的 3 个式子就可以得到 CF-WFCM 算法的聚类算法,计算步骤仍为 FCM 算法的 4 步.综上所述,我们得到 CF-WFCM 算法的属性权重学习算法和聚类算法.

例 1. 以 UCI 数据库中的 BUPA liver disorders 数据库为例,说明(1)如何通过有效性函数评价聚类结果,选取聚类个数 c;(2)比较 FCM 算法和 CF-WFCM 算法聚类结果.

BUPA liver disorders 数据库通过 6 个属性描述一个成年男子是否得肝病的 2 种情况,共 345 个数据.其中前 5 个属性是和肝病密切相关的血液指标,过量饮酒使得这 5 个指标值明显上升,最后一个属性是每天的饮酒量.根据先验知识,我们估计 BUPA liver disorders 数据库的大致有 2～10 类.因此,我们分别测试 BUPA liver disorders 数据库 2～10 类的有效性函数值如表 2.从表 2 中,我们可以发现,当 $c=2$ 时,各个有效性函数均取得最优值,所以我们断定 BUPA liver disorders 数据库应该分成 2 类,这与实际情况是一致的.

表 2 BUPA liver disorder 数据库取不同类数时有效性函数的值

类数	F	H	CS
2	0.829	0.416	0.126
3	0.628	0.899	0.599
4	0.552	1.149	0.662
5	0.468	1.452	0.673
6	0.390	1.728	1.875
7	0.360	1.907	1.246
8	0.338	2.045	1.018
9	0.297	2.245	1.978
10	0.275	2.392	2.058

通过属性权重学习算法得到 Bupa liver disorder 数据库 6 个属性的权重如表 3.当把该数据库分成 2 类的时候,FCM 算法和 CF-WFCM 算法聚类结果有效性函数值如表 4 所示.从表 4 中可以看出:CF-WFCM 算法的聚类结果优于 FCM 算法的聚类结果.

表 3　BUPA liver disorder 数据库属性的权重

属性	权重
mcv	0.063
alkphos	1.000
sgpt	1.000
sgot	0.266
gammagt	1.000
drinks	0.337

表 4　聚类 BUPA liver disorder 数据库聚类结果有效性函数值

算法	F	H	CS
FCM	0.829	0.416	0.126
CF-WFCM	0.840	0.392	0.112

4　属性评价函数的评价

文献[3]用属性评价函数 $E(w)$ 为每个属性学习权重，优化 FCM 算法得到较好的效果. 文献[4,5]通过优化 FCM 算法的价值函数 J，得到一种新的距离度量，改善聚类结果. 除了上面提到的属性评价函数，信息熵函数 $Fuzziness(w)$[3,6] 也可以为属性评价函数学习权重，其定义为

$$Fuzziness(w)=\frac{-2}{n(n-1)}\sum_{q<p}\frac{1}{2}(\rho_{pq}^{(w)}\log\rho_{pq}^{(w)}+(1-\rho_{pq}^{(w)})\log(1-\rho_{pq}^{(w)})) \quad (13)$$

文献[6]用信息熵函数 $Fuzziness(w)$ 学习权重，优化传递闭包聚类. 在这一过程中产生错误的划分结果，文献[6]根据所得结果直观地分析错误，并没有从量的角度说明该函数不适用于学习属性权重，优化聚类算法.

根据不同的优化原则，我们可以设计出不同的属性评价函数，优化聚类算法，那么如何从这些函数中选取较优的属性评价函数？根据聚类结果的优劣可以评价属性评价函数的优劣. 如果我们选取了较差的属性评价函数，在聚类结束后会发现聚类结果较差，达不到优化聚类的目的，此时浪费了大量的计算时间和计算资源. 因而我们需要一种方法在聚类之前就可以比较属性评价函数的优劣.

文献[7]认为：函数 $CFuzziness(w)$ 不仅可以评价属性的重要性，而且还具有熵函数的性质，能够度量空间变换模糊度. 也就是说，如果在原空间中相似(不相似)的数据，通过某个属性评价函数在变换后空间更相似(更不相似)，那么这个变换的模糊性较小，用于变换的属性评价函数较好，此时函数 $CFuzziness(w)$ 的值也较小；否则，称这个变换的模糊性较大，用于变换的属性评价函数较差，此时函数 $CFuzziness(w)$ 的值较大. 下面，我们用文献[3,6,7]中的一组人造数据说明：不同的属性评价函数学习的权重是不同的，导致空间变换的模糊性也不同，即函数 $CFuzziness(w)$ 的值. 只有模糊性小的空间变换，才能够达到优化聚类的目的，否则可能产生错误聚类结果. 函数 $CFuzziness(w)$ 的值在聚类过程之前就可以得到，也就是说我们在不知道聚类结果的前提下，通过计算函数 $CFuzziness(w)$ 的值就能够比较属性评价函数的优劣，从而避免了大量计算时间和计算资源的浪费.

例 2.　设 $CL=\{X_1,X_2,X_3,X_4,X_5\}$ 为待聚类的数据[3,6,7]，其中 $X_1=\{4.8,5.0,3.0,2.0\}$，$X_2=\{2.0,3.0,4.0,5.0\}$，$X_3=\{5.0,5.0,2.0,3.0\}$，$X_4=\{1.0,5.0,3.0,1.0\}$，$X_5=\{1.0,4.9,5.0,1.0\}$. 极小化函数 $E(w)$ 和函数 $Fuzziness(w)$ 为每个属性学习权重，优化 FCM 算法. 这两个函数空间变换模糊性的大小，即函数 $CFuzziness(w)$ 的值，如表 5 所示. 从表 5 中可以看出函数 $E(w)$ 空间变换的模糊性小于函数 $Fuzziness(w)$ 的模糊性. 这两组属性权重用于 FCM 算法得到的聚类结果有效性函数值如表 6 所示. 从表 6 中可以看出在各个聚类个数的情况下，通过函数 $E(w)$ 优化的 FCM 算法聚类结果优于函数 $Fuzziness(w)$ 优化的 FCM 算法聚类结果(除了类数为 2 的 V_{xb} 值). 根据函数 $E(w)$，我们判断这组数据应该分成 3 类；而根据函数 $Fuzziness(w)$ 的聚类结果认为这组数据应该分成 2 类. 文献[6]中传递闭包聚类的最优聚类结果为 3 类，因此函数 $Fuzziness(w)$ 的聚类结果产生错误，从而证明：函数 $E(w)$ 比函数 $Fuzziness(w)$ 更适合优化 FCM 算法，这与函数 $CFuzziness(w)$ 在聚类之前的判断结果一致.

表 5　函数 $E(w)$ 和 $Fuzziness(w)$ 学习的属性权重及其 $CFuzziness(w)$ 值

	$E(w)$	$Fuzziness(w)$
$w1$	0.488	0.251
$w2$	1	0.059
$w3$	0	0.061
$w4$	0	0
$CFuzziness(w)$	0.441	0.497

表 6　基于函数 $E(w)$ 和 $Fuzziness(w)$ 的 FCM 算法聚类结果

	聚类个数	V_{pc}	V_{pe}	V_{xb}
$E(w)$	2	0.87	0.31	0.027
	3	0.99	0.01	0.00009
	4	0.91	0.23	0.002
$Fuzziness(w)$	2	0.55	0.91	0.017
	3	0.45	1.30	0.159
	4	0.39	1.64	0.127

5 实验结果

从 UCI 数据库中选取 5 个数据库分别是 Glass，Iono，Iris，Pima 和 Vehicle. 表 7 介绍这 5 个数据库的特征. 这 5 个数据库通过 FCM 算法、CF-WFCM 算法和 WFCM 算法聚类结果的有效性函数值如表 8 所示. 从表 8 中可以发现：

(1) CF-WFCM 算法和 WFCM 算法的聚类结果均优于 FCM 算法的聚类结果；CF-WFCM 算法的聚类结果略优于 WFCM 算法的聚类结果.

(2) 针对不同数据库，CF-WFCM 算法对聚类结果的改善程度不一样，如 Glass 数据库的改善较为明显；

(3) 属性权重算法通过梯度递减算法学习权重，该算法可能收敛于局部极小值. 通过实验发现：即使属性权重算法收敛到局部极小值也能达到优化聚类的目的.

(4) 本文进行的所有实验中 CF-WFCM 算法均收敛.

CF-WFCM 算法聚类结果的改善是以学习权重为代价的，其时间复杂度为 $O(cn^2)$. 表 9 说明 CF-WFCM 算法中属性权重学习算法所用的时间.

表 7 数据库的特征

数据库	数据个数	属性个数	属性类型
Glass	345	6	数值型
Iono	351	34	数值型
Iris	150	4	数值型
Pima	768	8	数值型
Vehicle	94	18	数值型

表 8 FCM 算法、CF-WFCM 算法和 WFCM 算法聚类结果的比较

数据库	算法	F	H	CS	聚类个数
Glass	FCM	0.17	2.54	352.4	6
	CF-WFCM	0.57	1.23	0.6020	6
	WFCM	0.57	1.23	0.907	6
Iono	FCM	0.65	0.75	0.72	2
	CF-WFCM	0.65	0.74	0.65	2
	WFCM	0.65	0.75	0.71	2
Iris	FCM	0.78	0.57	0.13	3
	CF-WFCM	0.86	0.35	0.04	3
	WFCM	0.86	0.36	0.05	3
Pima	FCM	0.82	0.43	0.12	2
	CF-WFCM	0.87	0.31	0.08	2
	WFCM	0.87	0.31	0.08	2
Vehicle	FCM	0.70	0.80	0.24	4
	CF-WFCM	0.81	0.54	0.12	4
	WFCM	0.76	0.64	0.16	4

表 9 属性权重学习算法所用的时间

数据库	时间(s)
Glass	152
Iono	975
Iris	20
Pima	1147
Vehicle	35

6 结 论

CF-WFCM 算法强化重要属性在聚类过程中的作用，消减冗余属性的作用，从而改善 FCM 算法聚类的效果. 通过对部分 UCI 数据库的实验证明：CF-WFCM 算法的聚类结果优于 FCM 算法的聚类结果，且 CF-WFCM 算法聚类结果略优于 WFCM 的聚类结果. 函数 $CFuzziness(w)$ 不仅可以评价属性的重要性，而且可以评价属性评价函数的优劣. 本文用属性权重学习算法优化 FCM 算法，文献[6]用该算法优化传递闭包聚类算法[6]. 由此看出：只要能够将属性权重学习算法和聚类或分类算法很好地结合，就可以改善聚类和分类算法的效果. 下一步我们将研究如何用属性权重学习算法优化更多的聚类或者分类算法.

参 考 文 献

1 Dunn J. C.. Some recent investigations of a new fuzzy partition algorithm and its application to pattern classification problems. Journal of Cybernetics, 1974, 4: 1～15

2 Bezdek J. C.. Pattern Recognition with Fuzzy Objective Function Algorithms. New York: Plenum, 1981

3 Wang X. Z., Wang Y. D., Wang L. J.. Improving Fuzzy C-Means clustering based on feature-weight learning. Pattern Recognition Letters, 2004, 25(10): 1123～1132

4 Gustafson D. E., Kessel W.. Fuzzy clustering with a fuzzy covariance matrix. In: Proceedings of the IEEE Conference on Decision Control, San Diego, CA, 1979, 761～766

5 Wu K. L., Yang M. S.. Alternative c-means clustering algorithms. Pattern Recognition, 2002, 35(10): 2267～2278

6 Wang Xi-Zhao, Wang Li-Juan, Wang Li-Wei. The fuzziness analysis of transitive closure clustering. Computer Engineering and Applications, 2003, 39(18): 92～94(in Chinese)
(王熙照，王丽娟，王利伟. 传递闭包聚类中的模糊性分析. 计算机工程与应用, 2003, 39(18): 92～94)

7 Wang L. J., Wang X. Z., Ha M. H., Gu Y. S.. Mining the weights of similarity measure through learning. In: Proceedings of the 1st International Conference on Machine Learning and Cybernetics, Beijing, China, 2002, 4: 1837～1841

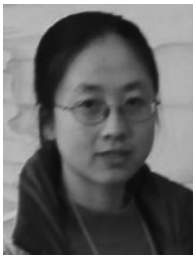
8 Jang J. S. R., Sun C. T., Mizutani E.. Neuro-Fuzzy and Soft

Computing. NJ：Prentice-Hall，1996

9 Pal N. R.，Bezdek J. C.. On cluster validity for the Fuzzy c-means model. IEEE Transactions on Fuzzy Systems，1995，3(3)：370～379

10 Xie X. L.，Beni G. A.. Validity measure for fuzzy clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence，1991，3(8)：841～846

11 Deluca A.，Termini S.. A definition of a nonprobabilistic entropy in the setting of fuzzy set theory. Information and Control，1972，20：301～312



WANG Li-Juan，born in 1978，Ph. D. candidate. Her research interests include feature selection，clustering and data mining.

GUAN Shou-Yi，born in 1970，lecturer. His research interests include data mining and education.

WANG Xiao-Long，born in 1955，professor，Ph. D. supervisor. His research interests include artificial intelligence，machine learning，computational linguistics，and Chinese information processing.

WANG Xi-Zhao，born in 1963，professor，Ph. D. supervisor. His research interests focus on machine learning.

Background

The work was supported by the National Natural Science Foundation of China (60435020 and 60473045) and the Natural Science Foundation of Hebei Province (603137). This paper investigates the clustering algorithm of Fuzzy C Mean(FCM). Each feature is denoted a feature weight by Information theory and gradient descent technique，the procedure of which forms the feature weight learning algorithm. When the feature weight is applied in FCM，it forms the CF-WFCM clustering algorithm，whose performance has been improved. The feature weight learning algorithm is a generalization of feature selection algorithm. CF-WFCM algorithm can be widely used in that variety areas，such as pattern recognition，data mining and artificial intelligence etc.