ISSN 1000-9825, CODEN RUXUEW
*Journal of Software*, Vol.21, No.8, August 2010, pp.1878−1888
doi: 10.3724/SP.J.1001.2010.03607

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

# 基于混合距离学习的双指数模糊 C 均值算法*

王 骏[1,2,3], 王士同[2,3+]

[1](南京理工大学 计算机科学与技术学院,江苏 南京 210094)

[2](江南大学 信息工程学院,江苏 无锡 214122)

[3](南京大学 计算机软件新技术国家重点实验室,江苏 南京 210093)

## Double Indices FCM Algorithm Based on Hybrid Distance Metric Learning

WANG Jun[1,2,3], WANG Shi-Tong[2,3+]

[1](School of Computer Science and Technology, Nanjing University of Science and Technology, Nanjing 210094, China)

[2](School of Information Technology, Jiangnan University, Wuxi 214122, China)

[3](State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093, China)

+ Corresponding author: E-mail: wxwangst@yahoo.com.cn

**Wang J, Wang ST. Double indices FCM algorithm based on hybrid distance metric learning. *Journal of Software*, 2010,21(8):1878−1888.** http://www.jos.org.cn/1000-9825/3607.htm

**Abstract**: To learn a good distance metric without any class label information, an algorithm named HDDI-FCM (double indices fuzzy C-means with hybrid distance) is proposed in this paper. In detail, the unknown distance metric is firstly represented as the linear combination of several known distance metrics. Then the algorithm is executed to perform the clustering task as well as learn the most suitable metric simultaneously. To guarantee the convergence of the algorithm, the Steffensen iteration is introduced into the process of updating cluster centers. The selection of parameter for the algorithm is also discussed. The experimental results on a collection of UCI (University of California, Irvine) datasets demonstrate the effectiveness of the proposed algorithm.

**Key words**: distance metric learning; clustering; fuzzy C-means algorithm; hybrid distance metric; Steffensen iteration method

**摘 要**: 提出了一种基于 DI-FCM(double indices fuzzy C-means)算法框架的无监督距离学习算法——基于混合距离学习的双指数模糊 C 均值算法 HDDI-FCM(double indices fuzzy C-means with hybrid distance).数据集未知距离度量被表示为若干已有距离的线性组合,然后执行HDDI-FCM,在对数据集进行有效聚类的同时进行距离学习.为了保证迭代算法收敛,引入了 Steffensen 迭代法来改进计算簇中心点的迭代公式.讨论了算法中参数的选择.基于 UCI(University of California,Irvine)数据集的实验结果表明该算法是有效的.

**关键词**: 距离学习;聚类;模糊 C 均值算法;混合距离;Steffensen 迭代法

**中图法分类号**: TP181 **文献标识码**: A

* Supported by the National Natural Science Foundation of China under Grant Nos.60773206, 60704047, 90820002 (国家自然科学基金)

Received 2008-07-28; Revised 2008-11-27; Accepted 2009-03-05

在聚类分析过程中,需要根据数据点之间的相似或相异程度,对数据点进行区分和分类.因此,为数据集选择合适的距离度量来评价数据点之间的相似或相异程度,对聚类分析的效果至关重要.在此过程中,选用不同的距离度量,其效果相去甚远.然而在实际研究中,由于研究人员缺乏对数据集的认识,所以很难为数据集选择合适的距离度量.在众多的距离度量中,欧氏距离最为常用.但是,欧氏距离对噪声比较敏感[1].此外,欧氏距离仅适用于特征空间中超球结构的数据集,对超立方体结构、超椭球结构的数据集效果不太理想.因此,在诸如生物工程、计算机视觉等领域,欧氏距离并不是一个很好的选择.

如何为数据集选择合适的距离度量,这一直是困扰学术界的一个难题.近年来,在机器学习领域已经提出了多种方法来学习数据集中未知的距离.根据是否有先验的训练样本提供,距离学习可分为有监督距离学习[2−6]和无监督距离学习[7,8]两类.前者借助于带标记的训练样本集,学习到数据集中的未知距离表示;而后者则没有任何关于数据集标记的先验信息.

对于无监督距离学习来说,其主要思想是在保持数据之间的局部或全局几何特性的前提下,学习隐藏在数据集中的低维流形.目前,大多数方法都是在保证几何特性的前提下把数据集向低维流形投影.本文提出了一种方法.在此方法中,适合于数据集的未知距离度量被表示为若干已有距离的线性组合,然后执行一种基于 DI-FCM(double indices fuzzy C-means)框架的算法.在实现距离学习的同时,也得到了良好的聚类结果.

本文第 1 节提出一种表示数据集距离度量的新方法——基于线性组合的混合距离表示方法,并给出用于更新中心点的 Steffensen 迭代公式.第 2 节回顾基于欧氏距离的双指数模糊 C 均值算法 DI-FCM.在此基础上,第 3 节提出基于混合距离的双指数模糊 C 均值算法 HDDI-FCM(double indices fuzzy C-means with hybrid distance).第 4 节给出距离的 3 种组合形式,通过实验数据说明此方法的有效性.第 5 节给出结论.

## 1 基于线性组合的混合距离表示新方法

距离空间是数学中一个重要的基本概念.距离空间(metric space)是一种拓扑空间,其上的拓扑由指定的距离函数决定.设 $X$ 是一个非空集,$X$ 被称为距离空间,是指在 $X$ 上定义了一个二元实值函数 $d(\boldsymbol{x},\boldsymbol{y})$满足以下 3 个条件:

(i) 非负性:$d(\boldsymbol{x},\boldsymbol{y})>0,\forall \boldsymbol{x}\neq\boldsymbol{y},d(\boldsymbol{x},\boldsymbol{x})=0$.

(ii) 对称性:$d(\boldsymbol{x},\boldsymbol{y})=d(\boldsymbol{y},\boldsymbol{x})$.

(iii) 三角不等式:$d(\boldsymbol{x},\boldsymbol{y})\leqslant d(\boldsymbol{x},\boldsymbol{z})+d(\boldsymbol{z},\boldsymbol{y}),\forall \boldsymbol{z}$.

这里,$d$ 称为 $X$ 上的一个距离,以 $d$ 为距离的距离空间记作$(X,d)$.

**定理**. 如果 $d_k(\boldsymbol{x},\boldsymbol{y})$是一个距离,则其线性组合 $D(\boldsymbol{x},\boldsymbol{y})=\sum_{k=1}^{n}\omega_k d_k(\boldsymbol{x},\boldsymbol{y})$ 也是一个距离.

证明:条件(i)、条件(ii)易证.以下证明 $D(\boldsymbol{x},\boldsymbol{y})$满足三角不等式(iii):

$$D(\boldsymbol{x},\boldsymbol{y})=\sum_{k=1}^{K}\omega_k d_k(\boldsymbol{x},\boldsymbol{y})\leqslant\sum_{k=1}^{K}\omega_k(d_k(\boldsymbol{x},\boldsymbol{z})+d_k(\boldsymbol{z},\boldsymbol{y}))=\sum_{k=1}^{K}\omega_k d_k(\boldsymbol{x},\boldsymbol{z})+\sum_{k=1}^{K}\omega_k d_k(\boldsymbol{z},\boldsymbol{y})=D(\boldsymbol{x},\boldsymbol{z})+D(\boldsymbol{z},\boldsymbol{y}),$$

即 $D(\boldsymbol{x},\boldsymbol{y})\leqslant D(\boldsymbol{x},\boldsymbol{z})+D(\boldsymbol{z},\boldsymbol{y})$.因此,$D(\boldsymbol{x},\boldsymbol{y})$是一个距离.证毕. □

本文通过如下线性组合来表示数据集中的未知距离度量:

$$D(\boldsymbol{x},\boldsymbol{y})=\sum_{k=1}^{K}\omega_k^{\,p}d_k(\boldsymbol{x},\boldsymbol{y}) \tag{1}$$

$$\text{s.t.}\ \sum_{k=1}^{K}\omega_k^{q}=1 \tag{2}$$

其中,$p>q>0$,$D(\boldsymbol{x},\boldsymbol{y})$表示数据点 $x$ 到数据点 $y$ 的距离,$d_k(\boldsymbol{x},\boldsymbol{y})$是其第 $k$ 个距离分量.

假设 $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_n\}$ 为 $s$ 维欧氏空间 $R^s$ 中的数据集,$\boldsymbol{x}_j(j=1,2,\ldots,n)$为特征向量,$\boldsymbol{V}=\{\boldsymbol{v}_1,\boldsymbol{v}_2,\ldots,\boldsymbol{v}_c\}\subset R^s$,$\boldsymbol{v}_i$ $(i=1,2,\ldots,c)$为第 $i$ 个簇的中心点,$\mu_{ij}$ 表示第 $j$ 个样本点属于第 $i$ 个簇的模糊程度,$m$ 是模糊指标,$c$ 为该数据集上簇的数量.通常使用最小平方误差法来估计每个簇的中心点 $\boldsymbol{v}_i$.对于 $\boldsymbol{v}_i$,使下式取得最小值:

$$J(\boldsymbol{X},\boldsymbol{V})=\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}{}^{m}D^2(\boldsymbol{x}_j,\boldsymbol{v}_i)\ .$$

对 $\boldsymbol{v}_i(i=1,2,\ldots,c)$求导,有

$$\frac{\partial J(\boldsymbol{X},\boldsymbol{V})}{\partial \mathbf{v}_i}=\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^{m}\frac{\partial D^2(\boldsymbol{x}_j,\boldsymbol{v}_i)}{\partial \mathbf{v}_i}=0\ ,\ i=1,2,\ldots,c \tag{3}$$

如果 $D(\boldsymbol{x}_j,\boldsymbol{v}_i)$取欧氏距离,求解较为方便,可以直接写出中心点 $\boldsymbol{v}_i$的显式表达式.但是,本文采用多个距离的线性组合表示 $D(\boldsymbol{x}_j,\boldsymbol{v}_i)$,而且距离分量未知,因此,直接写出其显式表达式就很困难了.本文采用迭代法来求解.将公式(1)代入公式(3),得到:

$$\frac{\partial J(\boldsymbol{X},\boldsymbol{V})}{\partial \boldsymbol{v}_i}=2\sum_{j=1}^{n}\mu_{ij}^{m}D(\boldsymbol{x}_j,\boldsymbol{v}_i)\sum_{k=1}^{K}\omega_k^p\frac{\partial d_k(\boldsymbol{x}_j,\boldsymbol{v}_i)}{\partial \boldsymbol{v}_i}=0\ ,\ i=1,2,\ldots,c \tag{4}$$

假设迭代过程产生关于中心点 $\boldsymbol{v}_i$的迭代序列 $\{\boldsymbol{v}_i^{(l)}\}_{l=1}^{\infty}$,$\varphi(\cdot)$为相应的迭代函数,则构造如下定点迭代过程:

$$\boldsymbol{v}_i^{(l+1)}=\varphi(\boldsymbol{v}_i^{(l)})\ ,\ i=1,2,\ldots,c,\ l=1,2,\ldots \tag{5}$$

距离分量 $d_k(\boldsymbol{x}_j,\boldsymbol{v}_i)$的选取是任意的,因此不能保证迭代过程收敛.为了使算法收敛,本文采用 Steffensen 迭代法[9]进行迭代,从而使目标函数收敛到其局部极小值.具体方法如下:

$$\boldsymbol{y}_i^{(l)}=\varphi(\boldsymbol{v}_i^{(l)}) \tag{6}$$

$$\boldsymbol{z}_i^{(l)}=\varphi(\boldsymbol{y}_i^{(l)}) \tag{7}$$

$$\boldsymbol{v}_i^{(l+1)}=\boldsymbol{v}_i^{(l)}-\frac{(\boldsymbol{y}_i^{(l)}-\boldsymbol{v}_i^{(l)})^2}{\boldsymbol{z}_i^{(l)}-2\boldsymbol{y}_i^{(l)}+\boldsymbol{v}_i^{(l)}}\ ,\ i=1,2,\ldots,c \tag{8}$$

## 2 基于欧氏距离的双指数模糊 C 均值算法

Zadeh 提出模糊集的概念之后,模糊聚类的研究成为学术界的一大热点,其中最为引人注目的就是模糊 C 均值算法 FCM(fuzzy C-means)[10,11].当前,研究的热点多集中于固定模糊指标 $m$ 值的算法研究,对于约束条件的研究却是一个空白.在标准 FCM 中,模糊指标 $m$ 取值范围为 $m>1$.本文在此基础上对约束条件的幂指数进行推广,向约束条件中引入幂指数 $r$,从而得到一种新算法,本文称其为双指数模糊 C 均值算法(DI-FCM).DI-FCM 的算法性能由模糊指标 $m$ 和约束条件的幂指数 $r$ 共同决定.其意义在于,在理论上有效地扩展了 $m$ 的取值范围,将模糊指标 $m$ 的取值由原先的 $m>1$ 扩展到 $m>r>0$.

令 $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_n\}$为 $s$ 维空间中的有限数据集,$\boldsymbol{x}_k=\{x_{k1},x_{k2},\ldots,x_{ks}\},k=1,2,\ldots,n$.令 $\boldsymbol{V}=\{\boldsymbol{v}_1,\boldsymbol{v}_2,\ldots,\boldsymbol{v}_n\}\subset R^s$,其中,$\boldsymbol{v}_i$表示第 $i$ 个簇的中心.$m>0$ 为模糊指标,$r>0$,对任意整数 $c$ 有 $2\le c\le n$.定义$\|\boldsymbol{x}_k-\boldsymbol{v}_i\|=\sqrt{\sum_{h=1}^{s}(x_{kh}-v_{ih})^2}$.与 FCM 类似,DI-FCM($m$,$r$)的目标函数定义为

$$J_{m,r}(\boldsymbol{U},\boldsymbol{V})=\sum_{k=1}^{n}\sum_{i=1}^{c}\mu_{ik}^{m}\|\boldsymbol{x}_k-\boldsymbol{v}_i\|^2,\quad m>0 \tag{9}$$

其中,$\boldsymbol{U}=[\mu_{ik}]_{c\times n}$为模糊划分矩阵,其元素$\mu_{ik}$满足如下 3 个约束条件:

$$\mu_{ik}\in[0,1],\ k=1,2,\ldots,n,\ i=1,2,\ldots,c \tag{10}$$

$$\sum_{i=1}^{c}\mu_{ik}^{r}=1\ ,\ m>r>0,\ k=1,2,\ldots,n \tag{11}$$

$$0<\sum_{k=1}^{n}\mu_{ik}<n\ ,\ i=1,2,\ldots,c \tag{12}$$

使用 Lagrange 乘子法,易得 $J_{m,r}$在约束条件(10)~(12)下取局部极小值的必要条件如下:

$$\boldsymbol{v}_i=\frac{\sum_{j=1}^{n}\mu_{ij}^{m}\boldsymbol{x}_j}{\sum_{j=1}^{n}\mu_{ij}^{m}},\ m>0,\ i=1,2,\ldots,c \tag{13}$$

$$\mu_{ik}=\left(\frac{\|\boldsymbol{x}_k-\boldsymbol{v}_i\|^{-\frac{2}{m/r-1}}}{\sum_{i=1}^{c}\|(\boldsymbol{x}_k-\boldsymbol{v}_i)\|^{-\frac{2}{m/r-1}}}\right)^{\frac{1}{r}},\ m>r>0,\ i=1,2,\ldots,c,\ k=1,2,\ldots,n \tag{14}$$

显然,当 $r$=1 时,DI-FCM 算法退化为 FCM 算法;当 $r\neq1$ 时,根据信息论的相关概念,数据集 $X$ 中第 $k$ 个样本的 $r$ 阶$\beta$熵为

$$H_{\beta_r}(x_k)=\frac{1}{1-2^{r-1}}\left(1-\sum_{i=1}^{c}\mu_{ik}^{r}\right),\ r>0\ 且\ r\neq1 \tag{15}$$

显然,当 $r\neq1$ 且满足约束条件(11)时,$H_{\beta_r}(\boldsymbol{x}_k)$ 值为 0,这意味着数据集 $\boldsymbol{X}$ 中样本点 $\boldsymbol{x}_k$ 从属于各个簇的不确定性最小.故 DI-FCM($m$,$r$)算法的实质就是数据集 $\boldsymbol{X}$ 中关于各个样本点 $\boldsymbol{x}_k$ 的模糊隶属度$\mu_{ik}$的 $r$ 阶$\beta$熵为 0 值的情况下,求解目标函数 $J_{m,r}$ 的最小值.

根据文献[10],为了保证算法收敛,目标函数(9)应该是 $\boldsymbol{U}$ 上的凸函数,满足 $m>r>0$.根据已有的研究,通常,$m/r$ 在[1.5,2.5]之间时可以得到较好的聚类效果.

## 3 基于混合距离学习的双指数模糊 C 均值算法

### 3.1 HDDI-FCM算法及推导

对于聚类算法而言,建立合适的准则函数对算法的效果至关重要.从直观上来说,我们总希望类内距离应尽可能地小,而类间距离尽可能地大.在此基础上加以引申,即各个数据点到各自所属簇的中心的距离之和应尽可能地小.结合双指数模糊 C 均值算法和本文提出的基于线性组合的混合距离,本文定义准则函数及约束条件如下:

$$J(\boldsymbol{U},\boldsymbol{V})=\sum_{j=1}^{n}\sum_{i=1}^{c}\mu_{ij}^{m}D^2(\boldsymbol{x}_j,\boldsymbol{v}_i) \tag{16}$$

其中:

$$D(\boldsymbol{x}_j,\boldsymbol{v}_i)=\sum_{k=1}^{K}\omega_k^{p}d_k(\boldsymbol{x}_j,\boldsymbol{v}_i) \tag{17}$$

为混合距离;$\boldsymbol{U}=[\mu_{ij}]_{c\times n}$ 为模糊划分矩阵,其元素$\mu_{ij}$满足:

$$\sum_{i=1}^{c}\mu_{ij}^{r}=1 \tag{18}$$

距离分量 $d_k$ 权重$\omega_k$满足:

$$\sum_{k=1}^{K}\omega_k^{q}=1 \tag{19}$$

运用 Lagrange 乘子法,构造无约束条件的准则函数如下:

$$E=\sum_{j=1}^{n}\sum_{i=1}^{c}\mu_{ij}^{m}D^2(\boldsymbol{x}_j,\boldsymbol{v}_i)+\sum_{j=1}^{n}\gamma_j\left(\sum_{i=1}^{c}\mu_{ij}^{r}-1\right)+\lambda\left(\sum_{k=1}^{K}\omega_k^{q}-1\right) \tag{20}$$

上式取极小值的必要条件为

$$\frac{\partial E}{\partial u_{ij}}=m\mu_{ij}^{m-1}D^2(\boldsymbol{x}_j,\boldsymbol{v}_i)+\gamma_j r\mu_{ij}^{r-1}=0,\quad i=1,2,\ldots,c,\ j=1,2,\ldots,n \tag{21}$$

$$\frac{\partial E}{\partial \gamma_j}=\sum_{i=1}^{c}\mu_{ij}^{r}-1=0,\quad j=1,2,\ldots,n \tag{22}$$

$$\frac{\partial E}{\partial \omega_k}=2p\omega_k^{p-1}\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^{m}D(\boldsymbol{x}_j,\boldsymbol{v}_i)d_k(\boldsymbol{x}_j,\boldsymbol{v}_i)+\lambda q\omega_k^{q-1}=0,\quad k=1,2,\ldots,K \tag{23}$$

$$\frac{\partial E}{\partial \lambda}=\sum_{k=1}^{K}\omega_k^{q}-1=0 \tag{24}$$

解公式(21)~(24)构成的方程组,得到准则函数(16)、(17)在约束条件(18)、(19)下取极小值的必要条件为:

$$\mu_{ij}=\left(\frac{\left(\sum_{k=1}^{K}\omega_k^{p}d_k(\boldsymbol{x}_j,\boldsymbol{v}_i)\right)^{-\frac{2}{m/r-1}}}{\sum_{i=1}^{c}\left(\sum_{k=1}^{K}\omega_k^{p}d_k(\boldsymbol{x}_j,\boldsymbol{v}_i)\right)^{-\frac{2}{m/r-1}}}\right)^{\frac{1}{r}} \tag{25}$$

$$\omega_k=\left(\frac{\left(\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^{m}D(\boldsymbol{x}_j,\boldsymbol{v}_i)d_k(\boldsymbol{x}_j,\boldsymbol{v}_i)\right)^{-\frac{1}{p/q-1}}}{\sum_{k=1}^{K}\left(\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^{m}D(\boldsymbol{x}_j,\boldsymbol{v}_i)d_k(\boldsymbol{x}_j,\boldsymbol{v}_i)\right)^{-\frac{1}{p/q-1}}}\right)^{\frac{1}{q}} \tag{26}$$

基于混合距离学习的双指数模糊 C 均值算法如下:

**HDDI-FCM 算法**.

输入:点集 $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_n\}$,聚类数目 $c$、距离分量数目 $K$.

输出:使目标函数最小化的数据集 $\boldsymbol{X}$ 的 $c$-划分.

方法:

1. 算法初始化:
   1a. 生成了随机数,初始化模糊隶属度矩阵 $\boldsymbol{U}=[\mu_{ij}]_{c\times n}$;
   1b. 根据 $\boldsymbol{U}$,初始化聚类中心 $\boldsymbol{v}_i,i=1,2,\ldots,c$;
   1c. 初始化权值向量 $\{\omega_k\}_{k=1}^{K}$.
2. 重复以下过程,直到相邻两次将计算目标准则函数的差值小于 $10^{-6}$:
   2a. 根据公式(25)计算$\mu_{ij}$;
   2b. 根据迭代公式(6)~(8)计算新的簇中心 $\boldsymbol{v}_i,i=1,2,\ldots,c$;
   2c. 根据公式(26)计算距离分量的权重$\omega_k$;
   2d. 根据公式(16)、(17)计算目标准则函数.

3.1.1 初始化

算法的初始化工作主要包括初始化模糊隶属度矩阵 $\boldsymbol{U}$、聚类中心 $\boldsymbol{v}_i$ 以及混合距离各个分量 $d_k$ 所占的权重$\omega_k$.

模糊隶属度矩阵 $\boldsymbol{U}$ 的初始值通过产生一系列 0-1 之间的随机数获得.在后续的迭代过程中,公式(25)保证约束条件(11)成立.

为了在首次迭代之前得到中心点的初始值,一种方案是随机选择数据集中若干个数据点作为初始的中心点.但是,这种方案在首次迭代过程中会出现个别数据点到其所属簇的中心点(即其自身)的距离为 0 的情况.如果在迭代公式$\varphi(\cdot)$中距离分量 $d_k(\boldsymbol{x}_j,\boldsymbol{v}_i)$出现在分母上,就会出现分母为 0 的情况.为了解决此问题,本文把随机生成的模糊隶属度矩阵 $\boldsymbol{U}$ 进行反模糊化(defuzzy),从而得到数据集的初始划分.对于每个簇,其中心点坐标简单地取此簇中各数据点坐标的平均值.

初始化$\omega_k$比较简单,令 $\omega_1^{(0)}=\omega_2^{(0)}=\ldots=\omega_K^{(0)}$ .根据约束条件(19),有

$$\omega_k^{(0)}=K^{-1/q},\ k=1,2,\ldots,K \tag{27}$$

3.1.2 迭 代

在此阶段,算法依次反复计算模糊隶属度$\mu_{ij}$、聚类中心 $\boldsymbol{v}_i$、距离分量权重$\omega_k$,并重新计算目标函数.当相邻两次目标函数的值小于一个给定的值时,则认为迭代过程结束.

重新计算模糊隶属度$\mu_{ij}$,实际上就是根据数据点 $\boldsymbol{x}_j$ 到各个簇中心点之间的混合距离,重新确定各个数据点从属于各个簇的模糊度.在此过程中,各个数据点从属于每个簇的程度发生了改变,但其变化方向趋于使准则函数(16)、(17)尽可能地小.

在计算中心点 $\boldsymbol{v}_i^{(l+1)}$ 的过程中,算法使用混合距离(17)计算每个数据点到此簇中心点 $\boldsymbol{v}_i^{(l)}$ 最新的距离,结合各个数据点从属于此簇的模糊程度$\mu_{ij}$更新簇的中心点.从本质上讲,这一过程与基于欧氏距离的 $K$ 均值算法是一致的,它们都是求各个数据点坐标的在特定距离空间中的平均值.所不同的是,本算法考虑了各个数据点从属于某个簇的模糊程度$\mu_{ij}$,并采用了新的混合距离计算公式.此外,计算新的中心点 $\boldsymbol{v}_i^{(l+1)}$ 是在上一次迭代得到的中心点 $\boldsymbol{v}_i^{(l)}$ 的基础上进行的.

计算混合距离中每个距离分量 $d_k$ 的权重$\omega_k$ 体现了距离学习的思想.从距离分量权重$\omega_k$ 的更新公式(26)中可以看出,$\omega_k$ 的计算考虑了当前迭代中每个数据点从属于每一个簇的模糊程度,以及各个数据点到各个簇中心点的距离.

Steffensen 迭代法的使用,保证了目标函数在每次迭代后其值都比上一次小,最终收敛于其局部极小值.当相邻两次计算目标准则函数(20)的差值小于 $10^{-6}$ 时,我们就认为迭代过程结束.

### 3.2 参数的选择

与 DI-FCM 类似,为了保证算法收敛,应该满足 $m>r>0$.通常,$m/r$ 在[1.5,2.5]范围内.

下面讨论 $p$ 和 $q$ 的取值.参数 $q$ 的取值与问题域有关,通常情况下,$q$ 取 1.根据公式(26)不难得到:

$$\lim_{p/q\to 1^+}\omega_t\to 1 \tag{28}$$

其中,$t=\arg\min_k\left(\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^m D(\boldsymbol{x}_j,\boldsymbol{v}_i)d_k(\boldsymbol{x}_j,\boldsymbol{v}_i)\right)$.可见,为了使算法能够自适应地选择更合适的距离分量 $d_k$ 以得到更理想的聚类效果,$p$ 应该取略大于 $q$ 的正值.其原因在于,公式:

$$\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^m D(\boldsymbol{x}_j,\boldsymbol{v}_i)d_k(\boldsymbol{x}_j,\boldsymbol{v}_i) \tag{29}$$

体现了划分矩阵 $\boldsymbol{U}=[\mu_{ij}]_{c\times n}$ 对数据集进行划分时,数据点在混合距离 $D$ 和距离分量 $d_k$ 的共同作用下的紧密程度.其值越小,说明距离分量 $d_k$ 越适合于数据集,它对应的权值$\omega_k$ 应该越大.

## 4 实 验

本文在 UCI 数据集上进行实验,所用数据集均从互联网[12]下载得到.数据集的相关特征见表 1.

**Table 1** Information on datasets used in our experiments

**表 1** 实验中使用的数据集信息

| Datasets | Number of instance | Number of features | Number of classes |
|---|---|---|---|
| iris | 150 | 4 | 3 |
| wdbc | 569 | 30 | 2 |
| wine | 178 | 13 | 3 |
| yeast | 1 484 | 8 | 10 |

使用以上数据集,本文选用如下 3 组距离组合进行实验:

组合 1. $d_1(\boldsymbol{x},\boldsymbol{y})=\|\boldsymbol{x}-\boldsymbol{y}\|$, $d_2(\boldsymbol{x},\boldsymbol{y})=(1-\mathrm{e}^{-\beta\|\boldsymbol{x}-\boldsymbol{y}\|^2})^{1/2}$.

混合距离由欧氏距离和由 Wu 和 Yang 提出的新距离[1]叠加而成.在公式(4)的基础上推导可得,第($l$+1)次迭

代的中心点 $\boldsymbol{v}_i^{(l+1)}=(v_{i1}^{(l+1)},v_{i2}^{(l+1)},...,v_{is}^{(l+1)})$ 的计算公式如下:

$$v_{ih}^{(l+1)}=\frac{\sum_{j=1}^{n}\mu_{ij}^{m}D(\boldsymbol{x}_j,\boldsymbol{v}_i^{(l)})x_{jh}\left[\frac{\omega_1^p}{d_1(\boldsymbol{x}_j,\boldsymbol{v}_i^{(l)})}+\beta\frac{\omega_2^p}{d_2(\boldsymbol{x}_j,\boldsymbol{v}_i^{(l)})}\mathrm{e}^{-\beta d_1^2(\boldsymbol{x}_j,\boldsymbol{v}_i^{(l)})}\right]}{\sum_{j=1}^{n}\mu_{ij}^{m}D(\boldsymbol{x}_j,\boldsymbol{v}_i^{(l)})\left[\frac{\omega_1^p}{d_1(\boldsymbol{x}_j,\boldsymbol{v}_i^{(l)})}+\beta\frac{\omega_2^p}{d_2(\boldsymbol{x}_j,\boldsymbol{v}_i^{(l)})}\mathrm{e}^{-\beta d_1^2(\boldsymbol{x}_j,\boldsymbol{v}_i^{(l)})}\right]},\ i=1,2,\ldots,c,\ h=1,2,\ldots,s,\ \bar{\boldsymbol{x}}=\frac{\sum_{j=1}^{n}\boldsymbol{x}_j}{n},\ l=1,2,\ldots \quad (30)$$

其中,参数 $\beta=\left(\frac{\sum_{j=1}^{n}\|\boldsymbol{x}_j-\bar{\boldsymbol{x}}\|^2}{n}\right)^{-1}$.

组合 2. $d_1(\boldsymbol{x},\boldsymbol{y})=\|\boldsymbol{x}-\boldsymbol{y}\|$, $d_2(\boldsymbol{x},\boldsymbol{y})=\sqrt{(\boldsymbol{x}-\boldsymbol{y})\boldsymbol{A}(\boldsymbol{x}-\boldsymbol{y})^T}$, 其中,$\boldsymbol{A}=diag(a_{11},a_{22},...,a_{ss})$,$a_{ii}=1/var_i(i=1,2,...,s)$,$var_i$ 是数据集第 $i$ 维的方差.

混合距离由欧氏距离和统计距离叠加而成.在公式(4)的基础上推导可得,中心点 $\boldsymbol{v}_i^{(l+1)}=(v_{i1}^{(l+1)},v_{i2}^{(l+1)},...,v_{is}^{(l+1)})$ 的计算公式如下:

$$v_{ih}^{(l+1)}=\frac{\sum_{j=1}^{n}\mu_{ij}^{m}D(\boldsymbol{x}_j,\boldsymbol{v}_i^{(l)})x_{jh}\left[\frac{\omega_1^p}{d_1(\boldsymbol{x}_j,\boldsymbol{v}_i^{(l)})}+\frac{\omega_2^p a_{hh}}{d_2(\boldsymbol{x}_j,\boldsymbol{v}_i^{(l)})}\right]}{\sum_{j=1}^{n}\mu_{ij}^{m}D(\boldsymbol{x}_j,\boldsymbol{v}_i^{(l)})\left[\frac{\omega_1^p}{d_1(\boldsymbol{x}_j,\boldsymbol{v}_i^{(l)})}+\frac{\omega_2^p a_{hh}}{d_2(\boldsymbol{x}_j,\boldsymbol{v}_i^{(l)})}\right]},\ i=1,2,\ldots,c,\ h=1,2,\ldots,s,\ l=1,2,\ldots \quad (31)$$

组合 3. $d_k(\boldsymbol{x},\boldsymbol{y})=\left(\sum_{h=1}^{s}|\boldsymbol{x}_h-\boldsymbol{y}_h|^{2k}\right)^{\frac{1}{2k}}$,即 $d_k(\boldsymbol{x},\boldsymbol{y})$取 $2k$-范数,$k=1,2,\ldots$

混合距离由若干项明氏距离叠加而成.对于奇数次项来说,涉及到绝对值,求导不方便.本文使用前若干个偶数次项来合成混合距离.在公式(4)的基础上推导可得,中心点 $\boldsymbol{v}_i^{(l+1)}=(v_{i1}^{(l+1)},v_{i2}^{(l+1)},...,v_{is}^{(l+1)})$ 公式如下:

$$v_{ih}^{(l+1)}=\frac{\sum_{j=1}^{n}\mu_{ij}^{m}D(\boldsymbol{x}_j,\boldsymbol{v}_i^{(l)})x_{jh}\sum_{k=1}^{K}\frac{\omega_k^p}{d_k(\boldsymbol{x}_j,\boldsymbol{v}_i^{(l)})}\left(\frac{\boldsymbol{x}_j,\boldsymbol{v}_i^{(l)}}{d_k(\boldsymbol{x}_j,\boldsymbol{v}_i^{(l)})}\right)^{2k-2}}{\sum_{j=1}^{n}\mu_{ij}^{m}D(\boldsymbol{x}_j,\boldsymbol{v}_i^{(l)})\sum_{k=1}^{K}\frac{\omega_k^p}{d_k(\boldsymbol{x}_j,\boldsymbol{v}_i^{(l)})}\left(\frac{\boldsymbol{x}_j,\boldsymbol{v}_i^{(l)}}{d_k(\boldsymbol{x}_j,\boldsymbol{v}_i^{(l)})}\right)^{2k-2}},\ i=1,2,\ldots,c,\ h=1,2,\ldots,s,\ l=1,2,\ldots \quad (32)$$

**实验 1**. 抗干扰能力.

本实验说明,混合距离可以较为成功地继承距离分量的某些特性.本实验选用距离组合 1 进行实验.在组合 1 中,由于距离分量 $d_2$ 抗干扰能力强[1],故距离组合 1 形成的混合距离在抗干扰能力方面有了明显的提高.

为了说明在新的混合距离下的抗干扰能力,本文在二维空间中随机生成一组数据点,可得其中心点(用"*"给出).加上例外点后,其新的中心点用"×"标出.在欧氏距离下,其中心点位置如图 1(a)所示.改用本文提出的新的混合距离,调节混合距离中各分量所占的比重,用 $\omega_1$ 表示距离分量 $d_1$ 在混合距离中所占比重,用 $\omega_2$ 表示距离分量 $d_2$ 在混合距离中所占比重.中心点位置分别如图 1(b)~图 1(d)所示.

从图中我们可以看出,向原数据集中加入例外点后,随着欧氏距离分量 $d_1$ 在混合距离表达式中所占比重的减小,新的中心点相对原中心点的偏移量逐渐减少.这说明混合距离的抗干扰能力逐渐增强.可见,基于线性组合的混合距离可以较为成功地继承距离分量的某些特性.

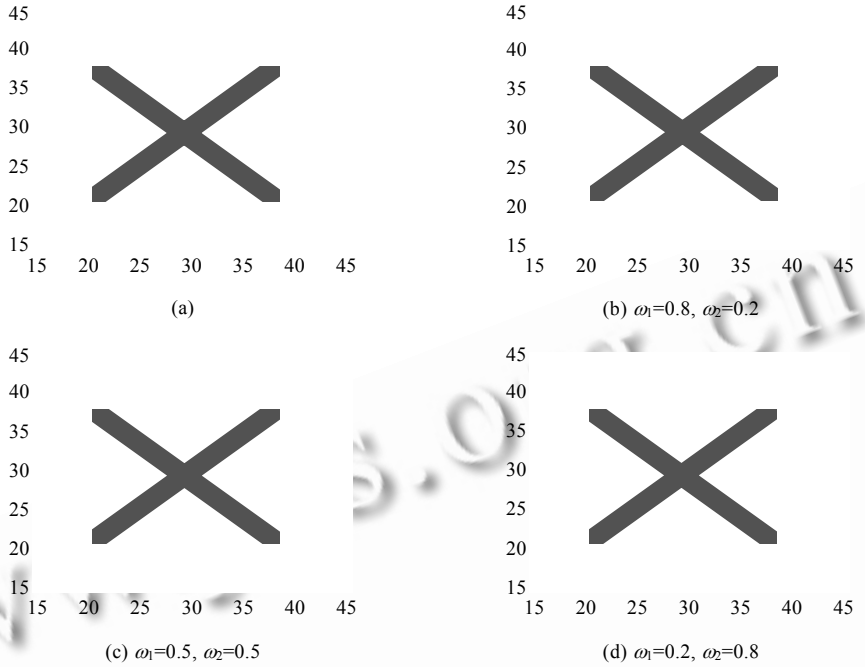
**实验 2**. 基于 *RandIndex* 的聚类性能比较.

本实验基于距离组合 1 和组合 2 研究本文算法的聚类效果,*RandIndex* 指标用来评价聚类结果与外部准则(external criterion)相吻合的程度.对于所给点集 $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_n\}$,假设 $\boldsymbol{U}^{(e)}=\{u_1^{(e)},u_2^{(e)},...,u_R^{(e)}\}$ 和 $\boldsymbol{U}=\{u_1,u_2,\ldots,u_C\}$是数据集 $\boldsymbol{X}$ 上两种不同的划分,对于 $1\leqslant i\neq i'\leqslant R$,$1\leqslant j\neq j'\leqslant C$,有 $\bigcup_{i=1}^{R}u_i^{(e)}=S=\bigcup_{j=1}^{C}u_j$, $u_i^{(e)}\cap u_{i'}^{(e)}=\phi=u_j\cap u_{j'}$.若 $\boldsymbol{U}^{(e)}$ 是外部准则,$\boldsymbol{U}$ 是聚类结果,令 $a$ 为在 $\boldsymbol{U}^{(e)}$和 $\boldsymbol{U}$ 中均出现在同一类中的点对的数量,$b$ 为在 $\boldsymbol{U}^{(e)}$中出现在同一类中,

而在 $\boldsymbol{U}$ 中出现在不同类中的点对的数量,$c$ 为在 $\boldsymbol{U}^{(e)}$中出现在不同类中、而在 $\boldsymbol{U}$ 中出现在同一类中的点对的数量,$d$ 为在 $\boldsymbol{U}^{(e)}$和 $\boldsymbol{U}$ 中均出现在不同类中的点对数量,则 *RandIndex* 通过如下式子进行计算:

$$RandIndex=\frac{a+d}{a+b+c+d} \tag{33}$$

显然,*RandIndex* 值介于[0,1]之间.当两种分类情况完全一致时,*RandIndex* 的值为 1.




(a) (b) $\omega_1$=0.8, $\omega_2$=0.2

(c) $\omega_1$=0.5, $\omega_2$=0.5 (d) $\omega_1$=0.2, $\omega_2$=0.8

Fig.1 Anti-Noise of the algorithm with distance combination 1

图 1 采用距离组合 1 时算法的抗干扰能力

表 2 和表 3 给出了混合距离采用距离组合 1 时,算法 HDDI-FCM 在不同数据集上所得的实验结果.表 2 中,DI-FCM 采用欧氏距离,即距离分量 $d_1$;DI-AFCM 采用参考文献[1]所提出的距离,即距离分量 $d_2$.类似地,表 4 和表 5 为混合距离采用距离组合 2 时所得的实验结果.表 4 中,DI-SFCM 是 DI-FCM 算法改用统计距离(即距离组合 2 中的距离分量 $d_2$)而得的算法.实验中,DI-FCM,DI-AFCM,DI-SFCM 均取 $m$=2.5,$r$=1.1.在本文提出的 HDDI-FCM 算法中,取 $m$=2.5,$r$=1.1,$p$=1.03,$q$=1,得到 *RandIndex* 值见表 2 和表 4.算法收敛时,不同数据集上各距离分量所占权重$\omega_k$($k$=1,2)见表 3 和表 5.

**Table 2** Comparison of the *RandIndex* performance for different algorithms (distance combination 1)

**表 2** 不同算法 *RandIndex* 指标对比(距离组合 1)

| | iris | wdbc | wine | yeast |
|---|---|---|---|---|
| DI-FCM | 0.879 732 | 0.750 377 | 0.710 531 | 0.688 104 |
| DI-AFCM | 0.912 394 | 0.802 730 | 0.728 115 | 0.670 096 |
| HDDI-FCM (combination 1) | 0.912 394 | 0.805 478 | 0.734 273 | 0.691 103 |

**Table 3** Component weights for hybrid distance when HDDI-FCM convergences (distance combination 1)

**表 3** HDDI-FCM 算法收敛时混合距离分量所占权重(距离组合 1)

| | iris | wdbc | wine | yeast |
|---|---|---|---|---|
| $\omega_1$ | $4.42\times10^{-15}$ | 0 | 0 | ≈1 |
| $\omega_2$ | ≈1 | 1 | 1 | $1.55\times10^{-13}$ |

**Table 4** Comparison of the *RandIndex* performance for different algorithms (distance combination 2)

**表 4** 不同算法的 *RandIndex* 指标对比(距离组合 2)

| | iris | wdbc | wine | yeast |
|---|---|---|---|---|
| DI-FCM | 0.879 732 | 0.750 377 | 0.710 531 | 0.688 104 |
| DI-SFCM | 0.836 779 | 0.848 177 | 0.939 821 | 0.541 543 |
| HDDI-FCM (combination 2) | 0.879 732 | 0.848 177 | 0.939 821 | 0.691 672 |

**Table 5** Component weights for hybrid distance metric when HDDI-FCM convergences (distance combination 2)

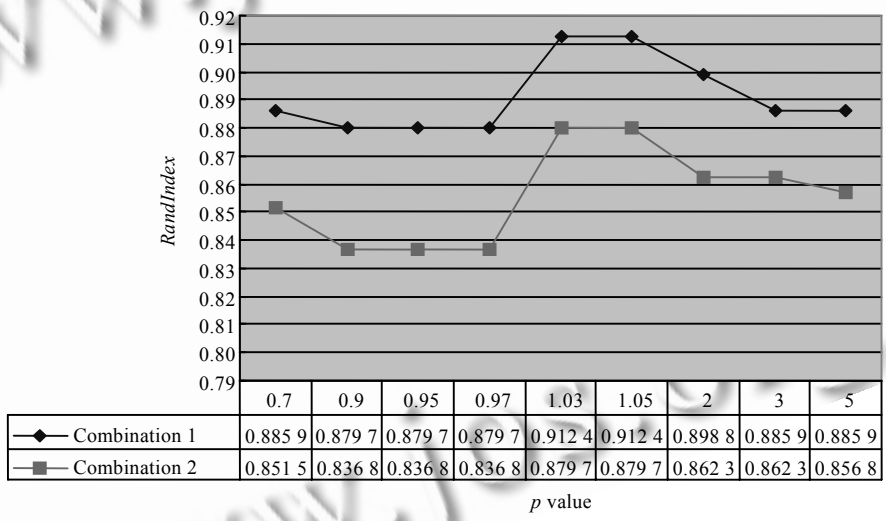
**表 5** HDDI-FCM 算法收敛时混合距离分量所占权重(距离组合 2)

| | iris | wdbc | wine | yeast |
|---|---|---|---|---|
| $\omega_1$ | ≈1 | 0 | 0 | 1 |
| $\omega_2$ | $1.80\times10^{-4}$ | 1 | 1 | 0 |

从表 2 所示实验结果中我们发现,对于 iris,wdbc,wine 这 3 个数据集,采用函数 $d_2$ 作为距离的 DI-AFCM 算法效果要好于采用函数 $d_1$ 作为距离的 DI-FCM 算法.相应地,表 3 中本文算法收敛时,距离分量 $d_2$ 所对应的权重 $\omega_2$ 也接近于 1.对于 yeast 数据集和表 4 与表 5 所示结果,也有类似结果.可见,HDDI-FCM 算法在收敛过程中从已有的各个距离分量中自适应地选择了一个更合适的距离分量.从部分结果中我们也发现,算法在选择更合适的距离分量的同时,还找到了一个更优的局部最优解.可见,HDDI-FCM 算法在进行距离学习的同时,还能进行更有效的聚类.

**实验 3**. 参数 $p$ 的不同取值对算法性能的影响.

本实验研究在固定 $m$,$r$,$q$ 值的情况下,不同 $p$ 值对算法性能的影响.取 $m$=2.5,$r$=1.1,$q$=1.采用 iris 数据集,混合距离各分量分别采用组合 1 和组合 2.算法收敛时,对于不同的 $p$ 值,所得聚类结果 *RandIndex* 值如图 2 所示.




| | 0.7 | 0.9 | 0.95 | 0.97 | 1.03 | 1.05 | 2 | 3 | 5 |
|---|---|---|---|---|---|---|---|---|---|
| Combination 1 | 0.885 9 | 0.879 7 | 0.879 7 | 0.879 7 | 0.912 4 | 0.912 4 | 0.898 8 | 0.885 9 | 0.885 9 |
| Combination 2 | 0.851 5 | 0.836 8 | 0.836 8 | 0.836 8 | 0.879 7 | 0.879 7 | 0.862 3 | 0.862 3 | 0.856 8 |

Fig.2 Influence of different parameter $p$ on clustering results (on iris dataset)

图 2 参数 $p$ 的不同取值对聚类结果的影响(iris 数据集)

我们发现,当 $p\to1^+$时,其聚类效果越好.这一结果与第 3.2 节所作的相关结论一致.其他 3 个数据集也可以得到相同的结论,限于篇幅,不再分别给出图表.

**实验 4**. 基于划分熵的聚类性能分析.

本实验采用的混合距离基于距离组合(3),研究聚类所得结果的模糊性.对于性能良好的模糊聚类算法而言,某个数据点从属于不同簇的各个隶属度中,其最大值应该大大超过其他值,这体现了数据点从属于某一个簇的

确定程度较大.显然,这种体现在隶属度上的不对称性越强,就说明聚类算法的效果越好.如果将模糊隶属度看作概率,那么可以使用划分熵来度量此不对称性.划分熵有如下定义[13]:

$$PE = -\frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{c}\mu_{ij}\log\mu_{ij} \tag{34}$$

这里,$n$ 表示数据集中数据点的总数;$c$ 表示簇的数量;$\mu_{ij}$ 表示数据点 $j$ 从属于簇 $i$ 的模糊程度,其值越小,表示不对称性越强.

分别选用前 1 项、前 3 项、前 5 项来实验,可得划分熵见表 6.

**Table 6** Partition entropy for clustering results

**表 6** 聚类结果的划分熵

| | iris | wdbc | wine | yeast |
|---|---|---|---|---|
| DI-FCM | 0.558 959 | 0.261 955 | 0.531 184 | 2.485 314 |
| The proposed algorithm (3 components) | 0.554 776 | 0.259 849 | 0.526 451 | 2.463 698 |
| The proposed algorithm (5 components) | 0.552 154 | 0.258 558 | 0.525 234 | 2.460 246 |

基于欧氏距离的 DI-FCM 算法相当于取一个距离分量的情况.从以上实验中可以看出,采用明氏距离组合,随着距离分量数目的增加,所得聚类结果的划分熵减少,这说明聚类结果的模糊性减小.

## 5 结 论

为数据集选择合适的距离度量对聚类分析的效果至关重要.传统的 FCM 算法的性能受欧氏距离的限制.本文提出了一种距离表示方法——基于线性组合的混合距离表示方法,数据集未知的距离度量被表示成若干已知距离的线性组合.在已有 DI-FCM 框架的基础上,提出了 HDDI-FCM 算法对其组合系数进行学习,从而得到其最佳的距离表示.为了保证迭代的收敛,本文使用 Steffensen 迭代法来改造中心点计算的迭代公式.还讨论了参数 $p$ 取值对算法性能的影响.多个实验结果表明,本文提出的方法是有效的,即不但可以进行有效的聚类,而且可以进行距离学习.

**References**:


[1] Wu KL, Yang MS. Alternative c-means clustering algorithms. Pattern Recognition, 2002,35(10):2267−2278. [doi: 10.1016/S0031-3203(01)00197-2]

[2] Xing EP, Ng AY, Jordan MI, Russell S. Distance metric learning with application to clustering with side-information. In: Becker S, Thrun S, Obermayer K, eds. Advances in Neural Information Processing Systems 15. Cambridge: MIT Press, 2002. 505−512.

[3] Bar-Hillel A, Hertz T, Shental N, Weinshall D. Learning a mahalanobis metric from equivalence constraints. Journal of Machine Learning Research, 2005,6:937−965.

[4] Bilenko M, Basu S, Mooney RJ. Integrating constraints and metric learning in semi-supervised clustering. In: Greiner R, Schuurmans D, eds. Proc. of the 21st Int'l Machine Learning Conf. New York: ACM Press, 2004. 81−88.

[5] Ceccarelli M, Maratea A. Improving fuzzy clustering of biological data by metric learning with side information. Int'l Journal of Approximate Reasoning, 2008,47(1):45−57. [doi: 10.1016/j.ijar.2007.03.008]

[6] Yang L, Jin R, Sukthankar R, Liu Y. An efficient algorithm for local distance metric learning. In: Proc. of the AAAI. Menlo Park: AAAI Press, 2006. 543−548.

[7] Ye JP, Zhao Z, Liu H. Adaptive distance metric learning for clustering. In: Proc. of the CVPR. Washington: IEEE Computer Society Press, 2007. 1−7.

[8] Wang XZ, Wang YD, Wang LJ. Improving fuzzy *c*-means clustering based on feature-weight learning. Pattern Recognition Letters, 2004,25(10):1123−1132. [doi: 10.1016/j.patrec.2004.03.008]

[9] Atkinson K. An Introduction to Numerical Analysis. 2nd ed., New York: John Wiley & Sons, 1989.

[10] Hathaway RJ, Bezdek JC, Tucker WT. An improved convergence theorem for the fuzzy c-means clustering algorithms. In: Bezdek J, ed. Proc. of the Analysis of Fuzzy Information, Vol.3. Boca Raton: CRC Press, 1987. 123−131.

[11] Gao XB, Xie WX. A study survey on the progress of fuzzy clustering algorithms. Chinese Science Bulletin, 1999,44(21): 2241−2251 (in Chinese with English abstract).

[12] Newman DJ, Hettich S, Blake CL, Merz CJ. UCI Repository of machine learning databases. Irvine: Department of Information and Computer Science, University of California, 1998. http://archive.ics.uci.edu/ml/

[13] Bezdek JC. Pattern Recognition with Fuzzy Objective Function Algorithms. New York: Plenum Press, 1981.

**附中文参考文献**:

[11] 高新波,谢维信.模糊聚类理论发展及应用的研究进展.科学通报,1999,44(21):2241−2251.

**王骏**(1978－),男,江苏苏州人,博士生,讲师,主要研究领域为模式识别,机器学习,数字图像处理,虚拟现实.

**王士同**(1964－),男,教授,博士生导师,主要研究领域为模糊神经网络,模式识别,生物信息学,数字图像处理.