ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.20, No.11, November 2009, pp.2939−2949
doi: 10.3724/SP.J.1001.2009.03410

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

基于模糊最大散度差判别准则的聚类方法*

高 军[1,2,3+], 王士同[1,3]

[1](江南大学 信息工程学院,江苏 无锡 214122)

[2](盐城工学院 信息工程学院,江苏 盐城 224001)

[3](浙江大学 CAD&CG国家重点实验室,浙江 杭州 310027)

Fuzzy Maximum Scatter Difference Discriminant Criterion Based Clustering Algorithm

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

[1](School of Information Engineering, Jiangnan University, Wuxi 214122, China)

[2](School of Information Engineering, Yancheng Institute of Technology, Yancheng 224001, China)

[3](State Key Laboratory of CAD&CG, Zhejiang University, Hangzhou 310027, China)

+ Corresponding author: E-mail: gjxllin@yahoo.cn

Gao J, Wang ST. Fuzzy maximum scatter difference discriminant criterion based clustering algorithm. *Journal of Software*, 2009,20(11):2939−2949. http://www.jos.org.cn/1000-9825/3410.htm

Abstract: In this paper, a fuzzy scatter difference discrimininant criterion is presented. Based on this criterion, fuzzy clustering algorithm FMSDC (fuzzy maximum scatter difference discriminant criterion based clustering algorithm) is also presented. The proposed algorithm reduces dimensionality while clustering by iterative optimizing procedure. First, it introduces the fuzzy concept into maximum scatter difference discriminant criterion; then the parameter η in the fuzzy criterion is appropriately determined based on specific principles so that the sensibility aroused by parameter η can be decreased to some extent; At last clustering and reducing dimensionality are realized according to fuzzy membership μ_{ik} and optional discriminant vector ω, respectively. Experimental results demonstrate the proposed method FMSDC is not only capable of clustering but also robust and capable of reducing dimensionality.

Key words: fuzzy maximum scatter difference discriminant criterion; discriminant vector; dimensionality reduction; fuzzy clustering; robust

摘 要: 基于最大散度差判别准则提出了一种模糊最大散度差准则,并根据模糊最大散度差准则提出一种聚类方法(fuzzy maximum scatter difference discriminant criterion based clustering algorithm,简称FMSDC).该方法通过迭代

* Supported by the National Natural Science Foundation of China under Grant Nos.60773206, 60903100, 90820002(国家自然科学基金); the National High-Tech Research and Development Plan of China under Grant Nos.2007AA1Z158, 2006AA10Z313 (国家高技术研究发展计划(863)); the National Defense Research Foundation of China under Grant No.A1420461266 (国防应用基础研究基金); the Jiangsu Provincial Innovation Project of Graduate Students of China under Grant No.CX09B-175Z (江苏省普通高校研究生科研创新计划); the Open Project Program of the State Key Laboratory of CAD&CG, Zhejiang University of China under Grant No.A0802 (浙江大学CAD&CG国家重点实验室开放课题)

Received 2008-03-03; Revised 2008-05-05; Accepted 2008-07-09

优化方法实现聚类的同时还可以实现特征降维.该方法首先在最大散度差判别准则中引入模糊概念;然后通过具体原则设定模糊最大散度差判别准则中的参数η,从而在一定程度上降低了由参数η引起的敏感性;最后分别根据模糊隶属度μ_{ik}、最优鉴别矢量ω进行聚类和特征降维.实验结果表明,FMSDC方法不但具有基本的聚类功能,而且具有较好的鲁棒性和较强的特征降维能力.

关键词: 模糊最大散度差判别准则;鉴别矢量;降维;模糊聚类;鲁棒性

中图法分类号: TP18 **文献标识码:** A

聚类作为一种有效的无监督模式识别方法,在许多领域得到了成功的运用[1−7].*C*-均值和模糊*C*-均值(fuzzy *C*-means,简称FCM)[8,9]是两种基本的聚类方法,而FCM方法是在*C*-均值方法上引入了模糊的概念,这就使得聚类的结果更符合现实世界的实际情况,因此在许多问题的求解中得到成功的运用[1−3].绝大部分类似于FCM方法的聚类方法[10−12],都是强调尽可能地实现类内散度最小化,而Kuo-Lung Wu等人提出的FCS方法[13]则不但考虑了类内散度最小化,而且充分考虑了类间散度最大化,因此与以上方法相比有更好的聚类效果.

最大散度差判别准则是一种新的线性判别准则[14,15],与Fisher准则一样都是寻找最优的鉴别矢量,使得各类之间尽可能地分开.该准则使用类间散度减去η倍类内散度作为判别标准,在一定程度上克服了Fisher准则类内散度矩阵奇异性问题,同时,宋枫溪等人证明了在类内散度矩阵非奇异时存在唯一的正实值η_0使得最大散度准则和经典的Fisher准则等效,从这个意义上讲,该判别准则可以看成是Fisher准则的泛化.然而,该准则的判别效果在很大程度上依赖于参数η的选取,同时依据该准则的划分属于硬划分,没有引入模糊概念,在一定程度上没有客观地反映现实世界.

因此,本文提出一种基于模糊最大散度差准则的聚类方法:FMSDC(fuzzy maximum scatter difference discriminant criterion based clustering algorithm).该方法不但继承了原准则的优点,而且具有如下优势:(1) 实现了在最大散度差准则中引入模糊概念,并用于无监督的聚类分析;(2) 根据具体原则设定参数η,提高聚类的稳定性;(3) 根据最优鉴别矢量实现特征降维,提高聚类效率和效果,还可以构造相应的分类器.

本文第 1 节简要介绍最大散度差准则.第 2 节重点阐述本文提出的方法:FMSDC.第 3 节通过实验说明FMSDC 方法的效果.第 4 节总结全文并提出以后要解决的问题.

1 最大散度差判别准则

最大散度差判别准则是在 Fisher 准则的基础上提出来的,其基本目的是寻找一个最优投影方向,实现分类的类内散度最小、类间散度最大,取得较好的分类效果,属于有监督的分类判别准则.

定义 1(类内散度)[16]. 假设有n个样本组成的样本集$D=\{x_1,\ldots,x_n\}$,它们分别属于C个不同的类,其中大小为n_k的样本子集D_k属于第k类,给定分类决策平面的法向量ω,则类内散度为$\omega^T S_W \omega$.其中,

$$S_W=\sum_{k=1}^{C}\sum_{x\in D_k}(x-u_k)(x-u_k)^T \tag{1}$$

称为类内散度矩阵,

$$u_k=\frac{1}{n_k}\sum_{x\in D_k}x(k=1,2,...,C) \tag{2}$$

称为均值.

定义 2(类间散度)[16]. 假设有n个样本组成的样本集$D=\{x_1,\ldots,x_n\}$,它们分别属于C个不同的类,其中大小为n_k的样本子集D_k属于第k类,u_k是第k类样本均值,给定分类决策平面的法向量ω,则类间散度为$\omega^T S_B \omega$.其中,

$$S_B=\sum_{k=1}^{C}n_k(u-u_k)(u-u_k)^T \tag{3}$$

称为类间散度矩阵,

$$u=\frac{1}{n}\sum_{x\in D}x \tag{4}$$

称为样本总体均值.

定义 3(最大散度差准则)[14]. 根据定义 1、定义 2,最大散度差准则定义为

$$\max_{\omega\neq 0}\frac{\omega^T S_B\omega-\eta\omega^T S_W\omega}{\omega^T\omega} \tag{5}$$

在取适当的参数η后,满足式(5)的鉴别矢量ω^*为最优鉴别矢量,这样,在矢量ω^*上的投影可以保证达到类内散度最小、类间散度最大.使用与求解Fisher准则相似的方法可以求出ω^*为$S_B-\eta S_W$最大特征值对应的特征向量,此时可以看出,在最大散度差准则中确实解决了Fisher准则中类内散度矩阵奇异性问题.然而,式(5)也表明参数η对求解最优鉴别矢量ω^*起着至关重要的作用.

2 基于模糊最大散度差判别准则的聚类方法:FMSDC

2.1 模糊最大散度差判别准则

使用与文献[13]相似的方法将模糊概念引入类内散度矩阵、类间散度矩阵,形成模糊类内散度、模糊类间散度并构造新的模糊最大散度差判别准则.

定义 4(模糊类内散度). 假设有n个样本组成的样本集$D=\{x_1,\ldots,x_n\}$,分别属于C个不同的类,其中大小为n_k的样本子集D_k属于第k类,u_k是第k类样本均值,给定分类决策平面的法向量ω,则模糊类内散度为$\omega^T S_{FW}\omega$.其中,

$$S_{FW}=\sum_{k=1}^{C}\sum_{i=1}^{n}\mu_{ik}^m(x_i-u_k)(x_i-u_k)^T \tag{6}$$

称为模糊类内散度矩阵[13].μ_{ik}表示第i个样本属于第k类的隶属度且$\sum_{k=1}^{C}\mu_{ik}=1$,m为模糊指数.

定义 5(模糊类间散度). 假设有n个样本组成的样本集$D=\{x_1,\ldots,x_n\}$,分别属于C个不同的类,其中大小为n_k样本子集D_k属于第k类,u_k是第k类样本均值,给定分类决策平面的法向量ω,则模糊类间散度为$\omega^T S_{FB}\omega$.其中,

$$S_{FB}=\sum_{k=1}^{C}\sum_{i=1}^{n}\mu_{ik}^m(u_k-u)(u_k-u)^T \tag{7}$$

称为模糊类间散度矩阵[13]. $u=\frac{1}{n}\sum_{x\in D}x$ 称为样本总体均值,μ_{ik}的定义同定义 4.

定义 6(模糊最大散度差判别准则). 根据定义的模糊类间散度、模糊类内散度,模糊最大散度差准则定义为

$$J_{FMSDC}=\max_{\omega\neq 0}\frac{\omega^T S_{FB}\omega-\eta\omega^T S_{FW}\omega}{\omega^T\omega} \tag{8}$$

根据拉格朗日乘数法,我们可以求解(8)对应的拉格朗日公式:

$$L=\omega^T S_{FB}\omega-\eta\omega^T S_{FW}\omega-\lambda\omega^T\omega+\sum_{i=1}^{n}\lambda_i\left(\sum_{k=1}^{C}\mu_{ik}-1\right) \tag{9}$$

其中,λ,λ_i为拉格朗日系数.

定理 1. 在模糊最大散度差准则中,式(9)的拉格朗日系数λ为$S_{FB}-\eta S_{FW}$对应的本特征值.

证明:若使式(8)成立,则必须满足:

$$\frac{\partial L}{\partial\omega}=0 \tag{10}$$

则有

$$(S_{FB}-\eta S_{FW})\omega=\lambda\omega \tag{11}$$

所以,定理成立. □

由定理 1 可以看出,本文提出的模糊最大散度差准则在求解最优鉴别矢量过程中也解决了 Fisher 准则类内散度矩阵的奇异问题,这说明该准则较好地继承了原准则的优点.

定理 2. 在模糊最大散度差准则中,式(8)成立的必要条件为

$$\mu_{ik}=\frac{\left(\omega^T(x_i-u_k)(x_i-u_k)^T\omega-\frac{1}{\eta}\omega^T(u_k-u)(u_k-u)^T\omega\right)^{\frac{1}{1-m}}}{\sum_{q=1}^{C}\left(\omega^T(x_i-u_q)(x_i-u_q)^T\omega-\frac{1}{\eta}\omega^T(u_q-u)(u_q-u)^T\omega\right)^{\frac{1}{1-m}}} \tag{12}$$

其中,u 为样本总体均值.

$$u_k=\frac{\sum_{i=1}^{n}\mu_{ik}^m\left(x_i-\frac{1}{\eta}u\right)}{\sum_{i=1}^{n}\mu_{ik}^m\left(1-\frac{1}{\eta}\right)} \tag{13}$$

证明:首先证明式(12)成立.根据式(9),则式(8)成立必须满足:

$$\frac{\partial L}{\partial \mu_{ik}}=0 \tag{14}$$

则有

$$\mu_{ik}=\left(\frac{\lambda_i}{m\omega^T(\eta(x_i-u_k)(x_i-u_k)^T-(u_k-u)(u_k-u)^T)\omega}\right)^{\frac{1}{m-1}} \tag{15}$$

又因为

$$\sum_{q=1}^{C}\mu_{iq}=1,$$

所以有

$$\sum_{q=1}^{C}\left(\frac{\lambda_i}{m\omega^T(\eta(x_i-u_q)(x_i-u_q)^T-(u_q-u)(u_q-u)^T)\omega}\right)^{\frac{1}{m-1}}=1 \tag{16}$$

根据式(15)、式(16)得式(12)成立.

同理,可以证明式(13)成立. □

根据模糊隶属度的要求,$\mu_{ik}\in[0,1]$,规定当式(12)的分子满足条件(17)时,令$\mu_{ik}=1$ 且$\mu_{ik}=0,q\neq k$,则

$$\omega^T(x_i-u_k)(x_i-u_k)^T\omega\le\frac{1}{\eta}\omega^T(u_k-u)(u_k-u)^T\omega \tag{17}$$

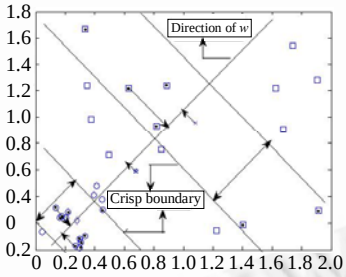


Fig.1 Sketch map of the crisp section

图 1 硬划分示意图

由式(17)可以看出,当某一样本x_i、第k类聚类中心u_k和样本总体均值u分别沿鉴别矢量ω方向投影后,如果样本投影点到聚类中心投影点的距离小于或等于聚类中心投影点到样本总体均值投影点距离的$1/\sqrt{\eta}$ 倍,则样本x_i严格隶属于第k类,即硬划分(如图 1 所示,其中,×和◇分别表示两类样本的聚类中心,☆表示样本总体均值,加"·"的样本所在区域是硬划分区域).而且式(12)、式(13)还表明模糊最大散度差判别准则中的参数η将直接影响到聚类效果的好坏,因此,如何对参数η进行设定对本文的 FMSDC 算法是至关重要的.下面给出的方法能够较好地解决这个问题.

2.2 设定模糊最大散度判别准则中的参数η

参数η在聚类过程中不但影响聚类中心,而且决定着硬划分对应的样本.因此,在一般情况下,如果某样本投影点到某一聚类中心投影点的距离小于或等于所有聚类中心投影点之间最小距离的 1/2,对该样本实行硬划分则较为合理,而且也避免了一个样本被硬划分到多类的情况.由此,可以根据如上分析来设定参数η的值.

定理 3. 当某一样本x_i、第k类聚类中心u_k和样本总体均值u分别沿鉴别矢量ω方向投影后,如果令$\eta=\max\{\eta_1,\eta_2,...,\eta_C\}$,其中,$\eta_k=\frac{N\max_q(\omega^T(u_k-u)(u_k-u)^T\omega)}{\min_{k\neq k^*}(\omega^T(u_{k^*}-u_k)(u_{k^*}-u_k)^T\omega)}$ $(N\ge4)$,则式(17)可以表示为

$$\omega^T(x_i-u_k)(x_i-u_k)^T\omega \le \frac{1}{N}\min_{k^*\ne k}(\omega^T(u_{k^*}-u_k)(u_{k^*}-u_k)^T\omega) \tag{18}$$

证明:因为根据式(17), $\omega^T(x_i-u_k)(x_i-u_k)^T\omega \le \frac{1}{\eta}\omega^T(u_k-u)(u_k-u)^T\omega,$

则根据题设有 $\frac{1}{\eta}\omega^T(u_k-u)(u_k-u)^T\omega \le \frac{1}{\eta_k}\omega^T(u_k-u)(u_k-u)^T\omega,$

则 $\frac{1}{\eta}\omega^T(u_k-u)(u_k-u)^T\omega \le \frac{\min_{k\ne k^*}(\omega^T(u_{k^*}-u_k)(u_{k^*}-u_k)^T\omega)}{N\max_q(\omega^T(u_k-u)(u_k-u)^T\omega)}\omega^T(u_k-u)(u_k-u)^T\omega,$

则 $\frac{1}{\eta}\omega^T(u_k-u)(u_k-u)^T\omega \le \frac{1}{N}\min_{k\ne k^*}(\omega^T(u_{k^*}-u_k)(u_{k^*}-u_k)^T\omega),$

所以有式(18)成立. □

由定理 3 可知,如果按照上述方法对参数η进行设定,可以保证在对某一样本进行硬划分时符合划分的直观含义,即满足该样本投影点到某一聚类中心投影点的距离小于或等于所有聚类中心投影点之间最小距离的$\frac{1}{\sqrt{N}}$ ($N\ge 4$),因此,从以上分析可知,N越大,硬划分的程度越低,反之则模糊程度越高.

正如文献[13]所考虑的一样,在本文的 FMSDC 方法中也要考虑在聚类时,聚类中心的收敛性问题,这样才能保证聚类的效果.

定理 4. 在FMSDC方法中,当μ_{ik}固定且$\eta>1$时,使得u_k是J_{FMSDC}局部最优解的充分必要条件为式(13)成立.

证明:必要性在定理 2 中已得到说明,下面来说明充分性.

将式(8)进行相应的转换:

$$J_{FMSDC}=\max_{\omega\ne 0}\frac{\omega^T S_{FB}\omega-\eta\omega^T S_{FW}\omega}{\omega^T\omega}=\max_{\omega\ne 0}\frac{-\eta\left(\omega^T S_{FW}\omega-\frac{1}{\eta}\omega^T S_{FB}\omega\right)}{\omega^T\omega}=\min_{\omega\ne 0}\frac{\eta\left(\omega^T S_{FW}\omega-\frac{1}{\eta}\omega^T S_{FB}\omega\right)}{\omega^T\omega} \tag{19}$$

则

$$\frac{\partial J_{FMSDC}}{\partial u_k}=\frac{2\eta\left(\omega^T\sum_{i=1}^{n}\mu_{ik}^m\left(\left(1-\frac{1}{\eta}\right)u_k-\left(x_i-\frac{1}{\eta}u\right)\right)\omega\right)}{\omega^T\omega} \tag{20}$$

设$\varphi(u_k)=J_{FMSDC}$,考虑由式(13)得到的拉格朗日$\varphi(u_k)$的Hessian矩阵$H(\varphi(u_k))$.根据式(20)可得:

$$h_{t,k}=\frac{\partial}{\partial u_t}\left[\frac{\partial\varphi(u_k)}{\partial u_k}\right]=\begin{cases}2\sum_{i=1}^{n}\mu_{ik}^m(\eta-1), & \text{当}t=k\text{时}\\ 0, & \text{当}t\ne k\text{时}\end{cases} \tag{21}$$

由式(21)可以看出$H(\varphi(u_k))$是一个对角阵,而根据定理条件$\eta>1$,所以该矩阵是正定矩阵,充分条件成立.这一结果与文献[13]中讨论的结果正好相互对应.

因此,根据定理 1~定理 4 可以得到如下的基于模糊最大散度差准则的聚类方法 FMSDC.

算法(FMSDC). 基于模糊最大散度差判别准则的聚类方法 FMSDC.

Step 1. 给定误差控制量$\varepsilon>0$,并随机产生隶属度矩阵$\boldsymbol{U}^0=(\mu_{ik}^0)_{n\times C}$,设定参数$\eta$,$N$初始值,并设$p=0$.

Step 2. 根据式(13)计算初始聚类中心$u_k^0(k=1,...,C)$.

Step 3. 计算$S_{FB}\eta S_{FW}$ 的最大特征值对应的特征向量ω.

Step 4. 根据定理 3 的前提设定参数η.

Step 5. 根据式(12)、式(18)计算隶属度矩阵$\boldsymbol{U}^{p+1}=(\mu_{ik}^{p+1})_{n\times C}$.

Step 6. 根据式(13)计算聚类中心$u_k^{p+1}(k=1,...,C)$.

Step 7. 如果$|J_{FMSDC}^{p+1}-J_{FMSDC}^{p}|\le\varepsilon$,则输出隶属度矩阵$\boldsymbol{U}^{p+1}$、聚类中心$u_k^{p+1}$、最优鉴别矢量$\omega$;否则,令$p=p+1$,转到 Step 3.

在多类聚类分析中(C>2),由于本文的 FMSDC 方法中引入鉴别矢量,即需将无监督的数据点投影到最优鉴别矢量上去,因此,如果将该方法作为划分聚类方法,则只适用于能够找到使多类样本投影恰好分开的方向线,然而绝大部分数据集不能满足这样的条件,从而限制了 FMSDC 方法作为划分聚类的能力.FMSDC 方法可以通过逐次二分法来实现多类聚类,同时最多可以得到 C–1 个最优的鉴别矢量,可以在一定程度上实现特征降维和构造相应的分类器.通过以上分析,本文的 FMSDC 方法可能更适用于多类的线性问题.

正如前文所述,FCS 方法同时考虑了类内散度、类间散度的问题,而本文的 FMSDC 方法是将模糊最大化散度判别准则引入到无监督的聚类中,通过引入鉴别矢量的方法来实现类内散度最小、类间散度最大,这种技巧不但可能在一定程度上提高聚类的鲁棒性,而且可以在聚类时实现特征降维,也可以根据得到的最优鉴别矢量来构造分类器,这些优势都会在后续的实验中得到验证.从这个意义上讲,FCS 显然在理论上不具备上述优势.

3 实 验

通过以上理论分析,FMSDC方法是将模糊最大散度差准则用于实现聚类,同时还可以完成特征降维的方法.为了说明该方法在处理实际问题时同样具有以上优势,使用FMSDC方法分别测试IRIS数据集[17]、纹理图像数据集、真实基因数据集Yeast galactose data[18]来说明方法的有效性.通过测试IRIS数据集说明方法的基本聚类能力,反映方法的基本功能;通过测试纹理图像数据集(无噪、加噪)来说明方法处理大数据集时的聚类效果和抗噪能力;测试高维基因数据集来表明方法的降维能力.测试的精度使用通用的Rand Index[19]来表示.

3.1 测试IRIS数据集

IRIS数据集是UCI[17]数据集中经典的数据集,经常用来测试聚类的效果.该数据集有 150 个样本组成,分成 3 类(每类有 50 个样本,每个样本有 4 个特征),其中第 2 类和第 3 类有交叉.在本文中使用该数据集来分别测试 FCM、FCS和FMSDC方法,在测试FMSDC方法时,使用划分聚类、二分法分别进行测试,以说明FMSDC方法的基本聚类功能,在测试过程中根据定理 3 和定理 4,选取 η=2,N=4.同时,为了表明参数 N 对硬划分的影响,设定了多种参数 N 值来分别表现划分效果.

从以上结果可以看出,本文的 FMSDC 方法具有如下特点:

(1) 表 1 和表 2 说明 FMSDC 方法具备基本的聚类功能,同时从表 1 还可以看出,FMSDC 方法在测试 IRIS 数据集时的聚类效果比 FCM 和 FCS 两种方法要好,这在一定程度上是由于 FMSDC 方法引进了最优鉴别矢量,这一点还可以分别从使用二分法和划分法测试 FMSDC 得到的效果看出.

Table 1 Rand index of FMSDC, FCM and FCS algorithms for IRIS datasets

表 1 本文的 FMSDC 方法与 FCM、FCS 方法分别测试 IRIS 数据集的 Rand Index 值

Algorithm		Mistaken partition numbers	Accuracy
FCM		16	0.893
FCS		7	0.9417
FMSDC	Partition clustering	6 (ε=1e−6)	0.949 5
	Dichotomy clustering	5 (ε=1e−6)	0.966 7

Table 2 Cluster center of FMSDC (partition clustering) algorithm for IRIS datasets

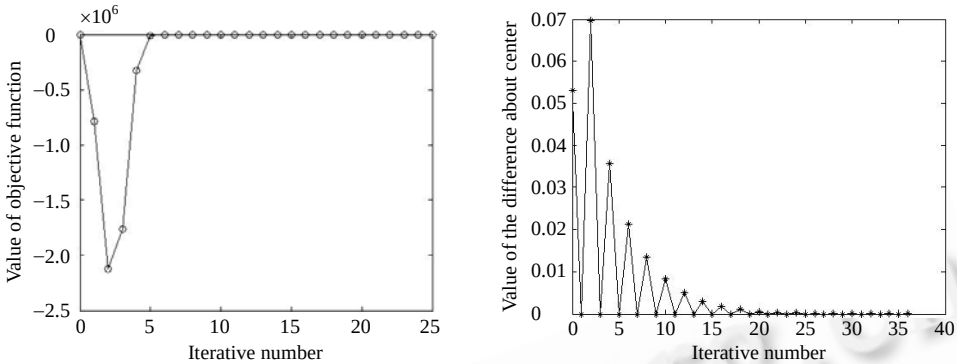
表 2 本文的 FMSDC(划分聚类)方法测试 IRIS 数据集得到的聚类中心

Cluster center			
0.240 22	0.516 29	−0.604 91	−0.845 24
0.509 48	0.248 99	0.300 28	0.131 47
0.740 46	0.397 66	0.701 42	0.747 51

(2) 图 2 表明了 FMSDC 算法和 FCS 算法迭代收敛情况.从该图中可以看出在测试 IRIS 数据集时,本文方法只要经过很少的迭代过程(迭代 6 次)就可以保证算法收敛到局部最优解,而 FCS 算法则要经过 25 次迭代才能收敛,由此可以说明 FMSDC 方法具有收敛速度快、迭代效率高的特点.

(3) 图 3 表明了当参数 N 发生变化时,IRIS 数据集第 3 类样本硬划分的效果.从图中可以看出,当 N 变大时,

样本被硬划分数目降低,即当 N=4 时有 38 个样本被硬划分,当 N=40 时有 27 个样本被硬划分,当 N=400 时只有 9 个样本被硬划分,而当 N =40000 时没有一个样本被硬划分,从而充分说明前面在理论上对参数 N 的分析是合理的,即 N 确实可以控制硬划分的效果.

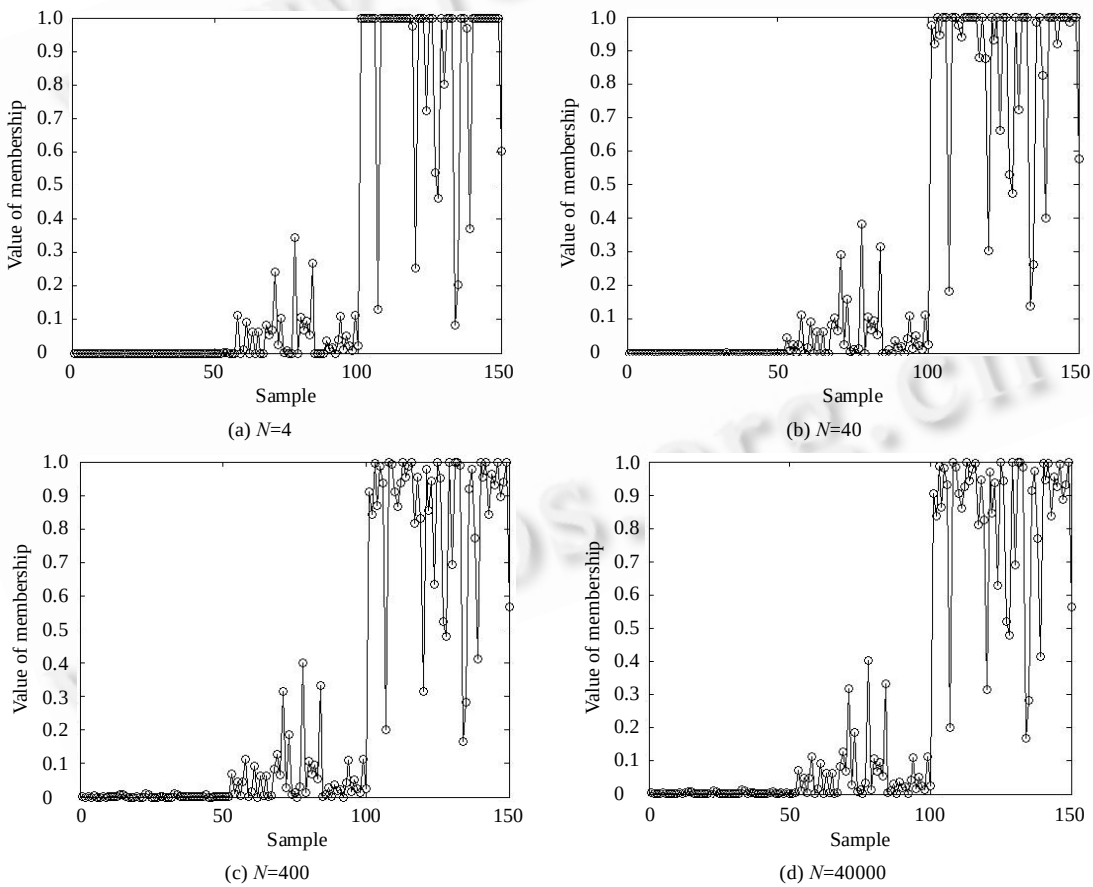


(a) Iterative convergence of FMSDC (b) Iterative convergence of FCS

(a) FMSDC 算法迭代收敛性 (b) FCS 算法迭代收敛性

Fig.2 Iterative convergence of FMSDC and FCS algorithms

图 2 FMSDC 和 FCS 算法迭代收敛性



(a) N=4 (b) N=40

(c) N=400 (d) N=40000

Fig.3 Distribution of the membership degree

图 3 隶属度分布

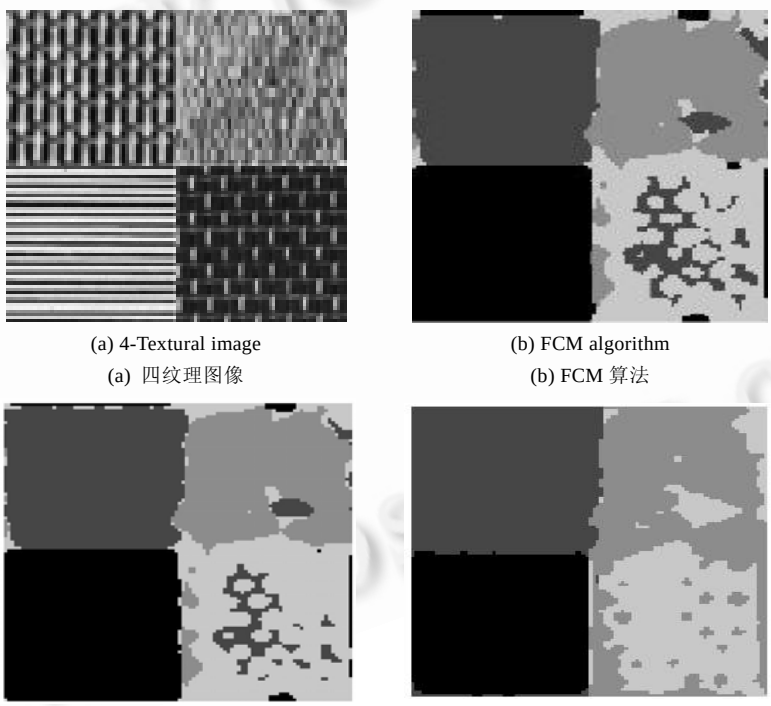
3.2 测试纹理图像

通过使用FMSDC算法对无噪和加噪的纹理图像进行纹理分割来说明算法处理大数据集时的聚类效果和抗噪性能.其中,FMSDC方法使用二分法,参数η=2,N=4.该图像采用Brodatz纹理库[20]中的纹理合成的图像进行纹理分割测试.图 4(a)为合成的四纹理图像,该图像是 100 像素×100 像素,即有 10 000 个样本.为了测试FMSDC算法的抗噪性,首先对合成的纹理图像分别使用FCM、FCS和FMSDC算法进行纹理分割(效果如图 4 所示).为了比较加噪后 4 种算法的抗噪性能,对合成纹理图像加高斯噪音.加高斯噪音图像如图 5(a)所示,分别使用FCM、FCS和FMSDC算法进行加噪纹理分割测试(效果如图 5 所示),其中图 6 是正确的分割.

从图 4~图 6 的结果可以看出,本文的 FMSDC 算法在处理大数据集时具有如下特征:

(1) 从图 4、表 3 可以看出,本文的 FMSDC 方法在处理无噪的纹理图像这种大数据集时的效果比 FCS 方法略好.从这个意义上可以说明,FMSDC 方法作为一种聚类方法在一定程度上可以替代 FCS 方法,同时也说明 FMSDC 方法的聚类效果较强.

(2) 图 5、表 3 说明 FMSDC 方法与其他方法相比具有较强的抗噪性.特别是与 FCS 方法相比更能说明 FMSDC 的优越性.由于 FMSDC 方法引入了最优鉴别矢量,每次二分聚类都能得到一个最优鉴别矢量,这样就可以使样本在沿着这些最优鉴别矢量投影后,投影后的样本点更有利于保证在聚类时尽可能达到类内散度最小、类间散度最大.而这一策略显然是 FCS 算法不具备的.从这个层面上讲,鉴别矢量的引入确实可以提高聚类的效果和鲁棒性.



(a) 4-Textural image (a) 四纹理图像

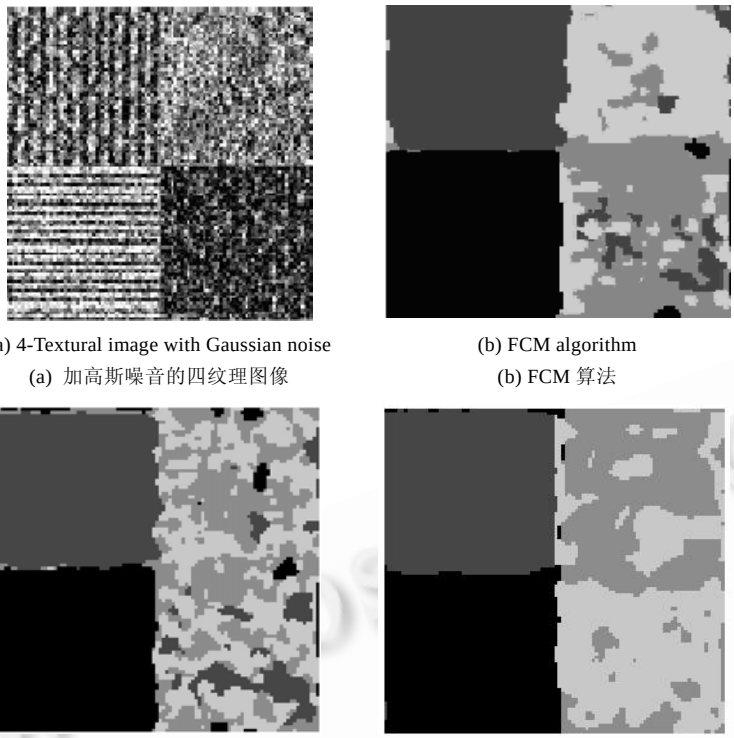
(b) FCM algorithm (b) FCM 算法

(c) FCS algorithm (c) FCS 算法

(d) FMSDC algorithm (d) FMSDC 算法

Fig.4 Segmentation results of four algorithms for 4-textural image

图 4 4 种方法对四纹理图像分割的结果



(a) 4-Textural image with Gaussian noise
(a) 加高斯噪音的四纹理图像
(b) FCM algorithm
(b) FCM 算法
(c) FCS algorithm
(c) FCS 算法
(d) FMSDC algorithm
(d) FMSDC 算法

Fig.5 Segmentation results of four algorithms for Gaussian noisy 4-textural image
图 5 4 种方法对加噪的四纹理图像分割结果

Fig.6 Perfect segmentation results for 4-textural image
图 6 四纹理图像标准分割结果

Table 3 Segmentation accuracies of four algorithms for (noisy) 4-texture image
表 3 4 种方法分割(加噪)四纹理图像的精度

Datasets	Algorithm	Index of Figures	Accuracy
4-Textural image (Fig.4(a))	FCM	Fig.4(b)	0.856
	FCS	Fig.4(c)	0.860 2
	FMSDC (ε=1e−6)	Fig.4(d)	0.863 7
4-Textural image with Gaussian noise (Fig.5(a))	FCM	Fig.5(b)	0.594 2
	FCS	Fig.5(c)	0.709 1
	FMSDC (ε=1e−6)	Fig.5(d)	0.838 8

3.3 测试基因数据集

真实基因数据集Yeast galactose data作为高维数据集[18,21],由 205 个样本组成,分成 4 类,每个样本具有 80 维.本文使用该数据集来测试FMSDC方法(其中 η=2,N=4)的特征降维能力.在使用FMSDC方法进行特征降维时,采用与线性判别函数解决多分类问题相似的方法,将 4 类问题分解成 3 个两类的聚类问题,每聚类 1 次产生 1 个相应的判别矢量,这样可以产生 3 个不同的判别矢量,并对以上的 3 个判别矢量进行正交化,形成一个三维空间,这样就可以将上述八十维的Yeast galactose data数据集在三维空间上进行投影,得到一个新的三维数据集,记为 3-Yeast galactose data.

从表 4 可以看出,FMSDC 方法具有如下特性:

(1) FMSDC 确实具备特征降维能力,这与前面的理论分析是相符的.FMSDC 这一特性是 FCS 不具备的.

(2) FMSDC 方法将八十维的数据集投影成三维数据集时精度没有太大的变化.这就说明 FMSDC 方法在特征降维时能够充分保留原数据集的数据特征信息,具有较强的特征降维能力.

Table 4 Validity of dimensionality reduction

表 4 特征降维的有效性

Datasets	Algorithm	Accuracy
Yeast galactose data	FCM	0.855 67
	FMSDC (ε=1e−6)	0.907 84
3-Yeast galactose data	FCM	0.837 3
	FMSDC (ε=1e−6)	0.878 0

4 总 结

本文提出一种模糊最大散度差判别准则,并将此准则引入到无监督的模糊聚类中,提出一种模糊聚类方法:FMSDC.该方法在通过迭代优化的方法解决聚类问题的同时,还可以得到最优鉴别矢量,实现特征降维,提高聚类效果.在构造模糊最大散度差判别准则过程中,提出了一种根据直观意义设定参数 η 的策略,较好地解决了原最大散度差判别准则对参数 η 敏感性的问题.当然,本文提出的 FMSDC 方法还有需要进一步研究之处,比如如何提高方法本身的处理速度以及方法的核化过程,这些将是我们以后要研究的工作.

References:

[1] Wang XZ, Wang YD, Wang LJ. Improving fuzzy *c*-means clustering based on feature-weight learning. Pattern Recognition Letters, 2004,25(4):1123−1132.

[2] Yu J, Li CX. Novel cluster validity index for FCM algorithm. Journal of Computer Science and Technology, 2006,21(1):137−140.

[3] Yu J. General *C*-means clustering model. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2005,27(8):1197−1211.

[4] Gao G, Wu J. A convergence theorem for the fuzzy subspace clustering (FSC) algorithm. Pattern Recognition, 2008(14): 1939−1947.

[5] Karayiannis NB. An axiomatic approach to soft learning vector quantization and clustering. IEEE Trans. on Neural Networks, 1999, 10(5):1153−1165.

[6] Karayiannis NB. Soft learning vector quantization and clustering algorithms based on ordered weighted aggregation operators. IEEE Trans. on Neural Networks, 2000,11(5):1093−1105.

[7] Karayiannis NB, Bezdek JC. An integrated approach to fuzzy learning vector quantization and fuzzy *c*-means clustering. IEEE Trans. on Fuzzy Systems, 1997,5(4):622−628.

[8] Chung KL, Lin JS. Faster and more robust point symmetry-based *K*-means algorithm. Pattern Recognition, 2007,40(2):410−422.

[9] Sun JG, Liu J, Zhao LY. Clustering algorithm research. Journal of Software, 2008,19(1):48−61 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/19/48.htm

[10] Rouseeuw PJ, Kaufman L, Trauwaert E. Fuzzy clustering using scatter matrices. Computational Statistics & Data Analysis, 1996, 23(4):135−151.

[11] Krishnapuram R, Kim J. Clustering algorithms based on volume criteria. IEEE Trans. on Fuzzy Systems, 2000,8(2):228−236.

[12] Gath I, Geva AB. Unsupervised optimal fuzzy clustering. IEEE Trans. on Pattern Analysis Machine Intelligence, 1989,11(7): 773−781.

[13] Wu KL, Yu J, Yang MS. A novel fuzzy clustering algorithm based on a fuzzy scatter matrix with optimality tests. Pattern Recognition Letters, 2005,26(10):639−652.

[14] Song FX, Cheng K, Yang JY, Liu SH. Maximum scatter difference, large margin linear projection and support vector machines. Acta Automatica Sinica, 2004,30(6):890−896 (in Chinese with English abstract).

[15] Song FX, Zhang D, Yang JY, Gao XM. Adaptive classification algorithm based on maximum scatter difference discriminant criterion. Acta Automatica Sinica, 2006,32(4):541−549 (in Chinese with English abstract).

[16] Bian ZQ, Zhang XG. Pattern Recognition. 2nd ed., Beijing: Tsinghua University Press, 2001 (in Chinese).

[17] Franc V, Hlavac V. Statistical pattern recognition toolbox. 2003. http://cmp.felk.cvut.cz/cmp/software/stprtool

[18] Ideker T, Thorsson V, Ranish JA, Christmas R, Buhler J, Eng JK, Bumgarner RE, Goodlett DR, Aebersold R, Hood L. Integrated genomic and proteomic analyses of a systemically perturbed metabolic network. Science, 2001,292(5):929−934.

[19] Yang CM, Wan BK, Gao XF. Selection of data preprocessing methods and similarity metrics for gene cluster analysis. Progress in Natural Science, 2006,16(6):607−613.

[20] Trygve R. Brodatz textures. 2006. http://www.ux.uis.no/~tranden/brodatz.html

[21] Chung FL, Wang ST, Deng ZH, Shu C, Hu D. Clustering analysis of gene expression data based on semi-supervised clustering algorithm. Soft Computing, 2006,10(5):981−993.

附中文参考文献:

[9] 孙吉贵,刘杰,赵连宇.聚类分析.软件学报,2008,19(1):48−61. http://www.jos.org.cn/1000-9825/19/48.htm

[14] 宋枫溪,程科,杨静宇,刘树海.最大散度差和大间距线性投影与支持向量机.自动化学报,2004,30(6):890−896.

[15] 宋枫溪,张大鹏,杨静宇,高秀梅.基于最大散度差鉴别准则的自适应分类算法.自动化学报,2006,32(4):541−549.

[16] 边肇祺,张学工.模式识别.第 2 版,北京:清华大学出版社,2001.

皋军(1971－),男,江苏阜宁人,博士生,副教授,主要研究领域为模糊系统,人工智能,模式识别,数据挖掘.

王士同(1964－),男,教授,博士生导师,主要研究领域为人工智能,模式识别,医学图像处理,模糊系统,神经网络.