基于模糊最大散度差判别准则的自适应特征提取模糊聚类算法

支晓斌[1],范九伦[2]
(1.西安电子科技大学电子工程学院,陕西西安 710071; 2.西安邮电学院通信与信息工程学院,陕西西安 710121)

摘 要: 指出皋军等人提出的基于模糊最大散度差判别准则(Fuzzy Maximum Scatter Difference Discriminant Criterion,FMSDC)的聚类算法(Fuzzy Maximum Scatter Difference Discriminant Criterion Based Clustering Algorithm,FMSDCA)中聚类中心表达式的推导错误及相关结论的错误,在修改该错误的基础上提出新的基于 FMSDC 的模糊聚类算法:FMSDC-FCS(Fuzzy Compactness and Separation Clustering Algorithm Based on Fuzzy Maximum Scatter Difference Discriminant Criterion).FMSDC-FCS 利用 FMSDC 产生最佳投影矢量,利用模糊紧性分离性(Fuzzy Compactness and Separation,FCS)算法对降维数据聚类,通过交替运行原数据空间中的 FMSDC 和投影空间中的 FCS 来优化投影矢量和聚类结果,最终通过对降维数据的聚类实现对原始数据的聚类.实验结果表明,FMSDC-FCS 总体性能优于原有的 FCS 算法、FMSDCA 算法以及经典的模糊 C-均值算法.

关键词: 模糊聚类; 模糊最大散度差判别准则; 特征提取; 模糊紧性分离性算法
中图分类号: TP181 **文献标识码:** A **文章编号:** 0372-2112 (2011) 06-1358-06

Adaptive Feature Extraction Fuzzy Clustering Algorithm Based on Fuzzy Maximum Scatter Difference Discriminant Criterion

ZHI Xiao-bin[1], FAN Jiu-lun[2]
(1. *School of Electronic Engineering, Xidian University, Xi'an, Shaanxi* 710071, *China*;
2. *School of Communication and Information Engineering, Xi'an Institute of Post and Telecommunications, Xi'an, Shaanxi* 710121, *China*)

Abstract: The derivation mistake of clustering center and the related wrong conclusion in Gao's fuzzy maximum scatter difference discriminant criterion based clustering algorithm (FMSDCA) are pointed out. A new clustering algorithm based on fuzzy maximum scatter difference discriminant criterion (FMSDC), called as fuzzy compactness and separation clustering algorithm based on fuzzy maximum scatter difference discriminant criterion (FMSDC-FCS), is proposed. FMSDC-FCS make use of the FMSDC to generate optimal projection vector and make use of the fuzzy compactness and separation (FCS) algorithm to cluster the reduced-dimensional data set. The projection vector and clustering result are optimized by alternately running FMSDC in the original data space and FCS in the projection space, and the original data is clustered by clustering the reduced-dimensional data. The experimental results demonstrate that the overall performance of FMSDC-FCS surpasses that of original FCS algorithm, FMSDCA and classical fuzzy c-means algorithm.

Key words: fuzzy clustering; fuzzy maximum scatter difference discriminant criterion; feature extraction; fuzzy compact and separation algorithm

1 引言

聚类分析是无监督模式识别中一个基本而重要的方法,模糊 C-均值(Fuzzy C-Means,FCM)聚类算法[1]是最为著名的聚类算法之一,得到学者们的广泛关注[2~5].然而 FCM 是一种基于类内紧致性度量的算法,它只考虑了聚类的类内紧致性而没有考虑类间分离性[6].为了进一步提高 FCM 的性能,Wu 等人提出模糊散布矩阵的概念,在此基础上提出了模糊紧性和分离性(Fuzzy Compactness and Separation,FCS)聚类算法[6].

Fisher 判别分析(Fisher Discriminant Analysis,FDA)是监督模式识别中经典的特征提取方法,数据经 FDA 投影处理后,在维数得到降低的同时,获得了最大类间分离性,在此基础上的分类也就变得更加高效和容易.但

收稿日期:2010-01-26;修回日期:2010-11-11
基金项目:国家自然科学基金(No.60572133);陕西省教育厅研究项目(No.2010JK835)

是在实际应用过程中,FDA 存在着散布矩阵奇异性问题,使得 FDA 的应用受到了很大的限制,鉴于此,众多学者提出了改进算法[7~10].宋枫溪等人提出了最大散度差准则[10],在一定程度上克服了 FDA 的散布矩阵奇异性问题.最近,皋军等人将最大散度差准则"模糊化",提出模糊最大散度差判别准则(Fuzzy Maximum Scatter Difference Discriminant Criterion,FMSDC),并提出一种基于 FMSDC 的聚类算法:FMSDCA(Fuzzy Maximum Scatter Difference Discriminant Criterion Based Clustering Algorithm)[11].FMS-DCA 能够在完成聚类的同时,得到最优投影方向.

本文指出皋军等人提出的 FMSDCA 中聚类中心的表达式是错误的,这种错误会导致如下的后果:(1)由于聚类中心表达式的错误,使得该算法的运行机理是不明确的;(2)导致后续的相关结论(原文定理 4)是错误的;(3)由于理论上的缺陷,FMSDCA 表现出聚类效果不理想,运行时间长的缺点.鉴于此,我们在修正 FMSDCA 错误的基础上,提出新的基于 FMSDC 的模糊聚类算法:FMSDC-FCS(Fuzzy Compactness and Separation Clustering Algorithm Based on Fuzzy Maximum Scatter Difference Discriminant Criterion).FMSDC-FCS 利用 FMSDC 的特征提取特性,交替运行原始数据空间中 FMSDC 和投影空间中的 FCS,通过对投影数据的聚类实现对原始数据的聚类.由于 FMSDC-FCS 中的聚类过程是在降维后的投影空间中进行的,所以 FMSDC-FCS 不仅可以获得优异的分类性能而且可以获得相对较高的执行效率.实验结果验证了 FMSDC-FCS 算法的有效性.

2　FCS 算法和 FMSDCA 算法简介

2.1　FCS 算法

为了能够同时描述聚类的类内紧致性和类间分离性,Wu 等人提出了模糊散布矩阵的概念,并分别定义了模糊总散布矩阵 $\boldsymbol{S}_{\mathrm{FT}}$,模糊类内散布矩阵 $\boldsymbol{S}_{\mathrm{FW}}$和模糊类间散布矩阵 $\boldsymbol{S}_{\mathrm{FB}}$[6]:

$$\boldsymbol{S}_{\mathrm{FT}}=\sum_{j=1}^{c}\sum_{i=1}^{n}u_{ij}^{m}(\boldsymbol{x}_i-\bar{\boldsymbol{x}})(\boldsymbol{x}_i-\bar{\boldsymbol{x}})^{\mathrm{T}} \tag{1}$$

$$\boldsymbol{S}_{\mathrm{FW}}=\sum_{j=1}^{c}\sum_{i=1}^{n}u_{ij}^{m}(\boldsymbol{x}_i-\boldsymbol{v}_j)(\boldsymbol{x}_i-\boldsymbol{v}_j)^{\mathrm{T}} \tag{2}$$

$$\boldsymbol{S}_{\mathrm{FB}}=\sum_{j=1}^{c}\sum_{i=1}^{n}u_{ij}^{m}(\boldsymbol{v}_j-\bar{\boldsymbol{x}})(\boldsymbol{v}_j-\bar{\boldsymbol{x}})^{\mathrm{T}} \tag{3}$$

其中 $\boldsymbol{v}_j=\sum_{i=1}^{n}u_{ij}^{m}\boldsymbol{x}_i\Big/\sum_{i=1}^{n}u_{ij}^{m}$ 称为模糊样本均值,$\bar{\boldsymbol{x}}$ 是整个数据集的均值,u_{ij}是隶属函数,满足 $u_{ij}\in[0,1]$, $\sum_{j=1}^{c}u_{ij}=1$, $m>1$. Wu 等人在模糊散布矩阵概念的基础上提出了 FCS 聚类算法[6].FCS 的目标函数为

$$J_{\mathrm{FCS}}=\sum_{j=1}^{c}\sum_{i=1}^{n}u_{ij}^{m}(\|\boldsymbol{x}_i-\boldsymbol{v}_j\|^2-\eta_j\|\boldsymbol{v}_j-\bar{\boldsymbol{x}}\|^2) \tag{4}$$

其中 $\eta_j\geqslant 0$.定义 Lagrange 函数

$$L_{\mathrm{FCS}}=\sum_{j=1}^{c}\sum_{i=1}^{n}u_{ij}^{m}(\|\boldsymbol{x}_i-\boldsymbol{v}_j\|^2-\eta_j\|\boldsymbol{v}_j-\bar{\boldsymbol{x}}\|^2)+\sum_{i=1}^{n}\alpha_i\sum_{j=1}^{c}(u_{ij}-1) \tag{5}$$

用 L_{FCS}对 v_j 和 u_{ij}分别求偏导数,并令偏导数为零,可得最小化式(4)的必要条件为

$$\boldsymbol{v}_j=\sum_{i=1}^{n}u_{ij}^{m}(\boldsymbol{x}_i-\eta_j\bar{\boldsymbol{x}})\Big/\sum_{i=1}^{n}u_{ij}^{m}(1-\eta_j) \tag{6}$$

$$u_{ij}=\frac{(\|\boldsymbol{x}_i-\boldsymbol{v}_j\|^2-\eta_j\|\boldsymbol{v}_j-\bar{\boldsymbol{x}}\|^2)^{-1/(m-1)}}{\sum_{k=1}^{c}(\|\boldsymbol{x}_i-\boldsymbol{v}_k\|^2-\eta_j\|\boldsymbol{v}_k-\bar{\boldsymbol{x}}\|^2)^{-1/(m-1)}} \tag{7}$$

由式(7)确定的 u_{ij}可能出现负值,为了使 u_{ij}的值在[0,1]区间内,当 u_{ij}为负值时需要对 u_{ij}进行修正,另外参数 η_j 也需要确定;Wu 等人给出了修正 u_{ij}的方法和参数 η_j 的确定方法,具体内容可参阅文献[6].

FCS 通过交替优化式(6)和式(7)最小化 L_{FCS},从而最小化 J_{FCS}.

2.2　FMSDCA 算法

皋军等人在模糊散布矩阵概念的基础上将最大散度差判别准则模糊化,进一步提出了模糊最大散度差判别准则(FMSDC)[11].设投影矢量为 $\boldsymbol{\omega}$,记投影变换为 $y=\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{x}$,在该投影空间,投影后的数据为 $y_i=\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{x}_i$,各类样本均值向量 $\tilde{v}_j=\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{v}_j$.模糊类内散布矩阵 $\tilde{\boldsymbol{S}}_{\mathrm{FW}}$定义为:

$$\tilde{\boldsymbol{S}}_{\mathrm{FW}}=\sum_{j=1}^{c}\sum_{i=1}^{n}u_{ij}^{m}(y_i-\tilde{v}_j)^2=\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{S}_{\mathrm{FW}}\boldsymbol{\omega} \tag{8}$$

模糊类间散布矩阵 $\tilde{\boldsymbol{S}}_{\mathrm{FB}}$定义为:

$$\tilde{\boldsymbol{S}}_{\mathrm{FB}}=\sum_{j=1}^{c}\sum_{i=1}^{n}u_{ij}^{m}(\tilde{v}_j-\bar{y})^2=\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{S}_{\mathrm{FB}}\boldsymbol{\omega} \tag{9}$$

定义模糊最大散度差判别准则函数:

$$J_{\mathrm{FMSDC}}=\max_{\boldsymbol{\omega}\neq 0}\frac{\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{S}_{\mathrm{FB}}\boldsymbol{\omega}-\eta\,\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{S}_{\mathrm{FW}}\boldsymbol{\omega}}{\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{\omega}} \tag{10}$$

皋军等人以上述定义的准则函数为目标函数,提出基于 FMSDC 聚类算法:FMSDCA.定义 Lagrange 函数:

$$L_{\mathrm{FMSDC}}=\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{S}_{\mathrm{FB}}\boldsymbol{\omega}-\eta\,\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{S}_{\mathrm{FW}}\boldsymbol{\omega}-\lambda\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{\omega}+\sum_{i=1}^{n}\lambda_i(\sum_{j=1}^{c}u_{ij}-1) \tag{11}$$

将 L_{FMSDC}分别对 $\boldsymbol{\omega}$, $\boldsymbol{v}_j$ 和 u_{ij}求偏导数,并令其为零,得 L_{FMSDC}取极大值必须满足的方程:

$$(\boldsymbol{S}_{\mathrm{FB}}-\eta\,\boldsymbol{S}_{\mathrm{FW}})\boldsymbol{\omega}=\lambda\boldsymbol{\omega} \tag{12}$$

$$\boldsymbol{v}_j=\sum_{i=1}^{n}u_{ij}^{m}(\boldsymbol{x}_i-\frac{1}{\eta}\bar{\boldsymbol{x}})\Big/\sum_{i=1}^{n}u_{ij}^{m}(1-\frac{1}{\eta}) \tag{13}$$

$$u_{ij}=\frac{[\boldsymbol{\omega}^{\mathrm{T}}(\boldsymbol{x}_i-\boldsymbol{v}_j)(\boldsymbol{x}_i-\boldsymbol{v}_j)^{\mathrm{T}}\boldsymbol{\omega}-\frac{1}{\eta}\boldsymbol{\omega}^{\mathrm{T}}(\boldsymbol{v}_j-\bar{\boldsymbol{x}})(\boldsymbol{v}_j-\bar{\boldsymbol{x}})^{\mathrm{T}}\boldsymbol{\omega}]^{-1/(m-1)}}{\sum_{k=1}^{c}[\boldsymbol{\omega}^{\mathrm{T}}(\boldsymbol{x}_i-\boldsymbol{v}_k)(\boldsymbol{x}_i-\boldsymbol{v}_k)^{\mathrm{T}}\boldsymbol{\omega}-\frac{1}{\eta}\boldsymbol{\omega}^{\mathrm{T}}(\boldsymbol{v}_k-\bar{\boldsymbol{x}})(\boldsymbol{v}_k-\bar{\boldsymbol{x}})^{\mathrm{T}}\boldsymbol{\omega}]^{-1/(m-1)}} \tag{14}$$

由式(14)确定的 u_{ij}可能出现负值,为了使 u_{ij}的值在[0,1]区间内,对 u_{ij}做如下的修正:

如果 $\boldsymbol{\omega}^{\mathrm{T}}(\boldsymbol{x}_i-\boldsymbol{v}_j)(\boldsymbol{x}_i-\boldsymbol{v}_j)^{\mathrm{T}}\boldsymbol{\omega}<\frac{1}{\eta}\boldsymbol{\omega}^{\mathrm{T}}(\boldsymbol{v}_j-\bar{\boldsymbol{x}})(\boldsymbol{v}_j-\bar{\boldsymbol{x}})^{\mathrm{T}}\boldsymbol{\omega}$

则 $u_{ij}=1$,且对其它的 $j'\neq j$, $u_{ij'}=0$ (15)

FMSDCA 通过交替优化式(12),式(13)和式(14)最大化 L_{FMSDC}.皋军等人给出了确定参数 η 的方法,即 $\eta=\max\{\eta_1,\eta_2,\cdots,\eta_c\}$,其中

$$\eta_k=\frac{N\max\limits_{q}\{\boldsymbol{\omega}^{\mathrm{T}}(\boldsymbol{v}_q-\bar{\boldsymbol{x}})(\boldsymbol{v}_q-\bar{\boldsymbol{x}})^{\mathrm{T}}\boldsymbol{\omega}\}}{\min\limits_{k^*\neq k}\{\boldsymbol{\omega}^{\mathrm{T}}(\boldsymbol{v}_{k^*}-\boldsymbol{v}_k)(\boldsymbol{v}_{k^*}-\boldsymbol{v}_k)^{\mathrm{T}}\boldsymbol{\omega}\}}(N\geqslant 4)$$

尽管 FMSDCA 表现出一定的分类性能,但是我们必须指出的是用式(11)对 $\boldsymbol{v}_j$ 求偏导数不能导出如式(13)中各类中心 $\boldsymbol{v}_j$ 的表达式.事实上,

$$\begin{aligned}L_{\mathrm{FMSDC}}=&\boldsymbol{\omega}^{\mathrm{T}}\Big(\sum_{j=1}^{c}\sum_{i=1}^{n}u_{ij}^{m}(\boldsymbol{v}_j-\bar{\boldsymbol{x}})(\boldsymbol{v}_j-\bar{\boldsymbol{x}})^{\mathrm{T}}\Big)\boldsymbol{\omega}\\&-\eta\,\boldsymbol{\omega}^{\mathrm{T}}\Big(\sum_{j=1}^{c}\sum_{i=1}^{n}u_{ij}^{m}(\boldsymbol{x}_i-\boldsymbol{v}_j)(\boldsymbol{x}_i-\boldsymbol{v}_j)^{\mathrm{T}}\Big)\boldsymbol{\omega}\\&-\lambda\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{\omega}+\sum_{i=1}^{n}\lambda_i\Big(\sum_{j=1}^{c}u_{ij}-1\Big)\end{aligned}$$

用 L_{FMSDC}对 $\boldsymbol{v}_j$ 求偏导数,并令其等于零可得方程

$$2\boldsymbol{\omega}^{\mathrm{T}}\Big(\sum_{i=1}^{n}u_{ij}^{m}(\boldsymbol{v}_j-\bar{\boldsymbol{x}})+\eta\sum_{i=1}^{n}u_{ij}^{m}(\boldsymbol{x}_i-\boldsymbol{v}_j)\Big)\boldsymbol{\omega}=\boldsymbol{0}$$

由于 $\boldsymbol{\omega}$ 是非零向量,经整理,上式等价于

$$\boldsymbol{\omega}^{\mathrm{T}}\sum_{i=1}^{n}u_{ij}^{m}(\boldsymbol{v}_j-\bar{\boldsymbol{x}}+\eta(\boldsymbol{x}_i-\boldsymbol{v}_j))=0 \tag{16}$$

式(16)是以 $\boldsymbol{v}_j$ 为解向量的单个线性方程,从该式无法解出唯一的 $\boldsymbol{v}_j$.观察式(16)可知,文献[11]的错误在于直接在上式中将 $\boldsymbol{\omega}^{\mathrm{T}}$约去,得到方程 $\sum_{i=1}^{n}u_{ij}^{m}(\boldsymbol{v}_j-\bar{\boldsymbol{x}}+\eta(\boldsymbol{x}_i-\boldsymbol{v}_j))=0$,进而解出 $\boldsymbol{v}_j$ 的表达式(13).这种错误导致 FMSDCA 的运行机理是不明确的,另外,从整个 FMSDCA 算法流程来看,FMSDCA 的聚类过程是在原数据空间完成,并没有利用投影空间,因此算法是很耗时的.实验中,FMSDCA 表现出一定的分类能力,其原因在于当 $\sum_{i=1}^{n}u_{ij}^{m}(\boldsymbol{v}_j-\bar{\boldsymbol{x}}+\eta(\boldsymbol{x}_i-\boldsymbol{v}_j))=0$ 时,$\boldsymbol{\omega}^{\mathrm{T}}\sum_{i=1}^{n}u_{ij}^{m}\boldsymbol{v}_j-\boldsymbol{\omega}^{\mathrm{T}}\sum_{i=1}^{n}u_{ij}^{m}(\bar{\boldsymbol{x}}-\eta(\boldsymbol{x}_i-\boldsymbol{v}_j))$也必然为零.即 FMSDCA 人为地将最优解的搜索区域缩小,可能只会得到一个近似最优解,而不是最优解.

3 FMSDC-FCS 算法

鉴于文献[11]的错误,本文重新考虑基于 FMSDC 的聚类算法.由 2.2 节的分析可知,我们无法在原数据空间求解得到使得 L_{FMSDC}最大化的聚类中心 $\boldsymbol{v}_j$,取而代之,我们可以在投影空间中求得使得 L_{FMSDC}最大化的聚类中心 $\tilde{v}_j$,这一结论由下面的定理 1 给出.

定理 1 L_{FMSDC}取极大值必须满足的必要条件为

$$\tilde{v}_j=\sum_{i=1}^{n}u_{ij}^{m}\Big(y_i-\frac{1}{\eta}\bar{y}\Big)\Big/\sum_{i=1}^{n}u_{ij}^{m}\Big(1-\frac{1}{\eta}\Big) \tag{17}$$

这里 $\bar{y}=\boldsymbol{\omega}^{\mathrm{T}}\bar{\boldsymbol{x}}$, $y_i=\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{x}_i$.

证明 事实上,由 $\tilde{\boldsymbol{S}}_{\mathrm{FW}}$和 $\tilde{\boldsymbol{S}}_{\mathrm{FB}}$的定义和式(11)可知,

$$\begin{aligned}L_{\mathrm{FMSDC}}=&\sum_{j=1}^{c}\sum_{i=1}^{n}u_{ij}^{m}(\tilde{v}_j-\bar{y})^2-\eta(y_i-\tilde{v}_j)^2)\\&-\lambda\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{\omega}+\sum_{i=1}^{n}\lambda_i\Big(\sum_{j=1}^{c}u_{ij}-1\Big)\end{aligned} \tag{18}$$

用式(18)对 $\tilde{v}_j$ 求导数并令导数为零,可得 $\sum_{i=1}^{n}u_{ij}^{m}(\tilde{v}_j-\bar{y}+\eta(y_i-\tilde{v}_j))=0$,解出聚类中心 $\tilde{v}_j$,即为式(17),证毕.

在文献[11]中,给出的结论"在 FMSDCA 方法中,当 u_{ij}固定且 $\eta>1$ 时,使得 $\boldsymbol{v}_j$ 是 J_{FMSDC}局部最优解的充分必要条件为式(13)"(参见文献[11]定理 4).我们必须指出:由于其式(13)的导出方式是错误的,即式(13)不是目标函数取极大值的必要条件,则更不可能是充分条件,所以这一结论是错误的.

根据本文给出的新的聚类中心的表达式,我们可以给出如下的聚类中心收敛性的结论.

定理 2 当投影方向 $\boldsymbol{\omega}$ 和隶属度 u_{ij}固定且 $\eta>1$ 时,使得 $\tilde{v}_j$ 是 J_{FMSDC}局部最优解的充分必要条件为式(17)成立.

证明 必要性已由定理 1 给出,下面说明充分性.当投影方向 $\boldsymbol{\omega}$ 和隶属度 u_{ij}固定时,由式(10)和 $\tilde{\boldsymbol{S}}_{\mathrm{FB}}$以及 $\tilde{\boldsymbol{S}}_{\mathrm{FW}}$的定义可知,

$$J_{\mathrm{FMSDC}}=\max_{\boldsymbol{\omega}\neq 0}\frac{\sum_{j=1}^{c}\sum_{i=1}^{n}u_{ij}^{m}((\tilde{v}_j-\bar{y})^2-\eta(y_i-\tilde{v}_j)^2)}{\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{\omega}}$$

令 $\varphi(\tilde{v}_j)=\sum_{j=1}^{c}\sum_{i=1}^{n}u_{ij}^{m}((\tilde{v}_j-\bar{y})^2-\eta(y_i-\tilde{v}_j)^2)$,则 J_{FMSDC}等价于 $\max\varphi(\tilde{v}_j)$, $\varphi(\tilde{v}_j)$对 $\tilde{v}_j$ 的二阶导数为 $2\sum_{i=1}^{n}u_{ij}^{m}(1-\eta)$,二阶混合导数为 0,则当 $\eta>1$ 时,$\varphi(\tilde{v}_j)$在由式(17)导出的 $\tilde{v}_j$ 处的 Hessian 矩阵是负定的,充分性成立,证毕.

通过对聚类中心表达式的修正,我们可以用交替迭代计算式(12)、式(17)和式(14)最大化 L_{FMSDC},从而最大化 J_{FMSDC}.这样就得到一个以 FMSDC 为目标函数的新的聚类算法,对于这一新算法的实质,我们可以通过下面的推导来获得进一步的认识.

由式(18)可知,当投影矢量 $\boldsymbol{\omega}$ 取定时,L_{FMSDC}等价于

$$\sum_{j=1}^{c}\sum_{i=1}^{n}u_{ij}^{m}((\tilde{v}_j-\bar{y})^2-\eta(y_i-\tilde{v}_j)^2)+\sum_{i=1}^{n}\lambda_i\Big(\sum_{j=1}^{c}u_{ij}-1\Big)$$

令 $\lambda_i' = -\frac{\lambda_i}{\eta}$,上式变为

$$-\eta\Big(\sum_{j=1}^{c}\sum_{i=1}^{n}u_{ij}^{m}\Big((y_i-\tilde{v}_j)^2-\frac{1}{\eta}(\tilde{v}_j-\bar{y})^2\Big)+\sum_{i=1}^{n}\lambda_i'\Big(\sum_{j=1}^{c}u_{ij}-1\Big)\Big)$$

令

$$L'_{\mathrm{FMSDC}}=\sum_{j=1}^{c}\sum_{i=1}^{n}u_{ij}^{m}\Big((y_i-\tilde{v}_j)^2-\frac{1}{\eta}(\tilde{v}_j-\bar{y})^2\Big)+\sum_{i=1}^{n}\lambda_i'\Big(\sum_{j=1}^{c}u_{ij}-1\Big)\tag{19}$$

则当 $\eta>0$ 时,最大化 L_{FMSDC}等价于最小化 L'_{FMSDC}.对比式(19)与式(5)可知,L'_{FMSDC}与 FCS 的目标函数 L_{FCS}是一致的,只是 FCS 算法中的参数 η_j 被这里的$\frac{1}{\eta}$取代.另外式(14)经过整理可重新表示为

$$u_{ij}=\frac{\big((y_i-\tilde{v}_j)^2-\frac{1}{\eta}(\tilde{v}_j-\bar{y})^2\big)^{-1/(m-1)}}{\sum_{k=1}^{c}\big((y_i-\tilde{v}_k)^2-\frac{1}{\eta}(\tilde{v}_k-\bar{y})^2\big)^{-1/(m-1)}}\tag{20}$$

对比式(17)、式(20)和式(6)、式(7)可知,当 FCS 算法的参数 η_j 被$\frac{1}{\eta}$取代时,式(17)与式(6)一致,式(20)和式(7)一致.由此可知,当投影方向 $\boldsymbol{\omega}$ 取定时,交替计算式(14)和式(17)就是执行投影空间中 FCS 聚类算法.

综上分析有如下结论:本文提出的新算法的实质就是以 FMSDC 为目标函数,在原数据空间中寻找最优投影方向 $\boldsymbol{\omega}$ 和投影空间中的 FCS 算法交替运行,二者相互优化的过程,记新算法为 FMSDC-FCS.因为新算法的聚类过程是在经过寻优的投影空间中进行的,所以与 FMSDCA 相比,新算法可获得更优的聚类效果和更高的执行效率,第 4 节的实验结果验证了这一点.在 FMSDC-FCS 算法在运行过程中,计算初始投影方向 $\boldsymbol{\omega}$ 需要初始隶属度 u_{ij},我们选择由 FCM 给出.这样就得到完整的 FMSDC-FCS 聚类算法.

FMSDC-FCS 算法

(1)给定 $\varepsilon>0$,设定初始参数 η_0 和 N,随机初始化目标函数 J_{FMSDC} 并运行 FCM 算法给出初始隶属度 u_{ij}和各类模糊样本均值 $\boldsymbol{v}_j$;

(2)求出 $\boldsymbol{S}_{\mathrm{FB}}-\eta\boldsymbol{S}_{\mathrm{FW}}$的最大特征值 λ,并取相应的特征向量 $\boldsymbol{\omega}$ 为最优投影矢量;

(3)用 $y=\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{x}$ 对数据投影;

(4)在投影空间中计算式(20)更新隶属度 u_{ij}';计算式(17)更新聚类中心 $\tilde{v}_j'$;

(5)利用式(10)更新目标函数 J'_{FMSDC},如果 $|J'_{\mathrm{FMSDC}}-J_{\mathrm{FMSDC}}|<\varepsilon$,则停止;否则,在原数据空间更新各类模糊样本均值 $\boldsymbol{v}_j'$ 并返回到(2).

4　实验结果及分析

本节通过用 FCM,FCS,FMSDCA 和 FMSDC-FCS 四个算法分别对 1 个人造数据和 3 个真实数据进行对比仿真实验,以验证本文提出的 FMSDC-FCS 算法的有效性.用准确率,迭代次数和运行时间三个指标综合衡量聚类算法的优劣.用 A,N 和 T 分别表示算法的准确率,迭代次数和运行时间,其中准确率定义为 $\mathrm{ac}=\sum_{j=1}^{c}n_j/n$(这里 n_j 表示第 j 类正确分类的数据点个数).在实验中,为避免算法陷入局部最优,算法在随机初始化后运行 100 次取最优分类结果,并取相应的迭代次数和 CPU 运行时间.最大迭代次数设为 200,停止阈值设为 10^{-6}.4 个算法中的参数 m 都取为 2,FCS 算法中的参数 η_j 取为 0.5,与文献[11]中的参数选取方式相同,FMSDCA 和 FMSDC-FCS 算法中的参数 η_0 都取为 2,L 都取为 4.

实验 1　人造数据

图 1 中的数据来自文献[12],该数据是 4 类线性可分数据,但是由于数据的复杂性,如果不对数据进行特征提取,传统中心聚类算法很难将其完全正确分类.图 1 是 4 个算法对该数据的聚类结果.其中,用"◦"、"+"、"∗"和"×"分别表示 4 类数据,"◦"表示聚类中心.图 1(*a*)、(*b*)和(*c*)分别是 FCM,FCS 和 FMSDCA 的聚类结果,图 1(*d*)、(*e*)和(*f*)分别是 FMSDC-FCS 迭代过程中迭代次数为 1 次,4 次和 7 次时聚类结果.由图 1 可知,FCM,FCS 和 FMSDCA 都不能对该数据完全正确分类,而 FMSDC-FCS 算法经过 7 次迭代就能自适应地找到最优投影方向并对该数据正确分类,这表明 FMSDC-FCS 算法具有良好的收敛性和特征提取能力;由表 1 可知,FMSDC-FCS 的运行时间只比 FCM 略多一些,但少于 FCS 和 FMSDCA,这表明 FMSDC – FCS 算法具有相对较高的执行效率.

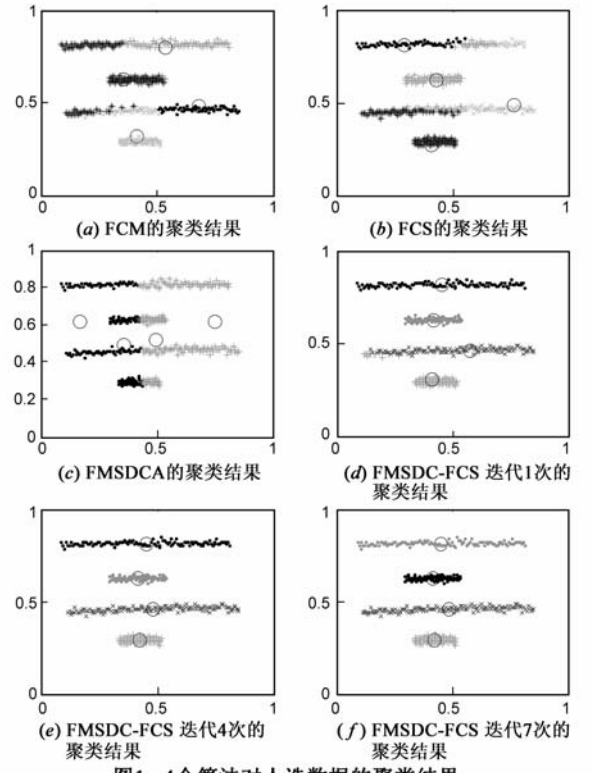

图1　4个算法对人造数据的聚类结果

表 1 4 个算法对人造数据的聚类结果

	A	N	T(s)
FCM	0.7500	48	0.656000
FCS	0.7383	65	1.594000
FMSDCA	0.3418	75	16.000000
FMSDC-FCS	1,0000	7	0.703000

实验 2 真实数据

本节选取的 3 个标准数据 Iris, Breast Cancer 和 Wine 来自于 UCI repository of machine learning databases[13], 数据特性如表 2 所示. 聚类结果如表 3、表 4 和表 5 所示.

表 2 数据描述

	样本数	维数	类数
Iris	150	4	3
Breast Cancer	683	10	2
Wine	178	13	3

表 3 4 个算法对 Iris 数据聚类结果

	A	N	T(s)
FCM	0.8933	24	0.047000
FCS	0.8933	28	0.188000
FMSDCA	0.9600	79	3.125000
FMSDC-FCS	0.9733	26	0.171000

表 4 4 个算法对 Breast-cancer 数据聚类结果

	A	N	T(s)
FCM	0.5652	24	0.188000
FCS	0.6047	21	0.266000
FMSDCA	0.9605	24	2.391000
FMSDC-FCS	0.9693	35	0.609000

表 5 4 个算法对 Wine 数据聚类结果

	A	N	T(s)
FCM	0.6854	47	0.110000
FCS	0.6742	68	0.406000
FMSDCA	0.6742	119	5.828000
FMSDC-FCS	0.7022	144	0.922000

由表 3 可知, FMSDC-FCS 对 Iris 的聚类精度最高, 要明显优于 FCM 和 FCS. FFC-SFCA 也达到较高的聚类精度, 但是 FMSDCA 的用时远远多于 FMSDC-FCS.

由表 4 可知, FMSDCA 和 FMSDC-FCS 对 Breast Cancer 的聚类精度要明显高于 FCM 和 FCS, 相对而言, FMSDC-FCS 的聚类精度更高; 从运行时间来看, FMSDC-FCS 的运行时间远远少于 FMSDCA 的运行时间.

由表 5 可知, 4 个算法对 Wine 数据的聚类精度都不是很高, 相对而言, 本文提出的 FMSDC-FCS 的聚类精度是最高的, 优于其它 3 个算法. 这说明 FMSDC-FCS 算法对真实数据具有更好的适应性, 另外, FMSDC-FCS 的运行时间远远少于 FMSDCA 的运行时间.

5 结论

本文指出了皋军等人提出的 FMSDCA 聚类算法中的若干推导错误, 在分析该算法错误根源的基础上提出新的基于 FMSDC 的聚类算法: FMSDC-FCS. 理论分析和实验结果表明, 本文提出的 FMSDC-FCS 不仅具有良好的特征提取能力, 而且具有相对较高的执行效率, 是一种切实可行的聚类算法. 本文提出的 FMSDC-FCS 是基于两类 FMSDC 的且适合线性可分数据的聚类问题, 将 FMSDC-FCS 推广为多类, 并且核化 FMSDC-FCS 聚类算法是本文下一步的工作.

参考文献

[1] Bezdek J C. Pattern Recognition with Fuzzy Objective Function Algorithms[M]. New York: Plenum Press, 1981. 95 – 107.

[2] Yu J. Optimality test for generalized FCM and its application to parameter selection[J]. IEEE Transactions on Fuzzy Systems, 2005, 13(1): 164 – 176.

[3] Yang M S, Wu K L, et al. Alpha-cut implemented fuzzy clustering algorithms and switching regressions[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B, 2008, 38(3): 588 – 603.

[4] 范九伦, 吴成茂. FCM 算法中隶属度的新解释及其应用[J]. 电子学报, 2004, 32(2): 350 – 352.
Fan J L, Wu C M. The new explanation of membership degree in FCM and its applications[J]. Acta Electronica Sinica, 2004, 32(2): 350 – 352. (in Chinese)

[5] 于剑, 程乾生. 关于 FCM 算法中的权重指数 m 的一点注记[J]. 电子学报, 2004, 32(3): 478 – 480.
Yu J, Cheng Q S. A note on the weighting exponentmin FCM algorithm[J]. Acta Electronica Sinica, 2004, 32(3): 478 – 480. (in Chinese)

[6] Wu K L, Yu J, Yang M S. A novel fuzzy clustering algorithm based on a fuzzy scatter matrix with optimality test[J]. Pattern Recognition Letters, 2005, 26(5): 639 – 652.

[7] Belhumeur P N, Hespanha J P, Kriegman D J. Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1997, 19(7): 711 – 720.

[8] Yu H, Yang J. A direct LDA algorithm for high dimensional data-with application to face recognition[J]. Pattern Recognition, 2001, 34(10): 2067 – 2070.

[9] Ye J P. Characterization of a family of algorithms for generalized discriminant analysis on undersampled problems[J]. Journal of Machine Learning Research, 2005, 6(4): 483 – 502.

[10] 宋枫溪, 程科, 杨静宇, 刘树海. 最大散度差和大间距投影与支持向量机[J]. 自动化学报, 2004, 30(6): 890 – 896.
Song F X, Cheng K, Yang J Y, Liu S H. Maximum scatter dif-

ference,large margin linear projection and support vector machines[J].Acta Automatica Sinica,2004,30(6):890 – 896.(in Chinese)

[11] 皋军,王士同.基于模糊最大散度差判别准则的聚类方法[J].软件学报,2009,20(11):2939 – 2949.
Gao J,Wang S T.Clustering algorithm based on fuzzy maximum scatter difference discriminant criterion[J].Journal of Software,2009,20(11):2939 – 2949.(in Chinese)

[12] 王玲,薄列峰,焦李成.密度敏感的谱聚类[J].电子学报,2007,35(8):1577 – 1581.
Wang L,Bo L F,Jiao L C.Density-sensitive spectral clustering[J].Acta Electronica Sinica,2007,35(8):1577 – 1581.(in Chinese)

[13] Blake C L,Merz C J.UCI repository of machine learning databases[OL].http://mlearn.ics.uci.edu/MLRepository.html,1998 – 07.

作者简介

支晓斌　男,1976 年生,内蒙古巴彦淖尔人.西安电子科技大学博士研究生,西安邮电学院理学院讲师.研究方向:模式识别,模糊集理论及其应用.
E-mail:xbzhi@163.com

范九伦　男,1964 年生,河南温县人.博士后,教授,博士生导师.现为西安邮电学院通信与信息工程学院院长,信息安全研究中心主任.研究方向:模糊集理论、模糊信息处理、模式识别与图像处理、信息安全等.在国内外刊物上发表学术论文 150 余篇.