基于模糊 Fisher 准则的半模糊聚类算法

曹苏群[①②]　王士同[①]　陈晓峰[①]　谢振平[①]　邓赵红[①]
[①](江南大学信息学院　无锡　214122)
[②](淮阴工学院机械系　淮安　223001)

摘　要：该文针对线性可分数据提出一种鲁棒的基于模糊 Fisher 准则的半模糊聚类算法 FFC-SFCA。FFC-SFCA 通过模糊化散布矩阵，将模糊理论引入 Fisher 判别方法，通过对模糊 Fisher 准则函数迭代优化实现聚类。FFC-SFCA 的优势在于具有很好的鲁棒性且可以获得可分性好的聚类结果，同时，可以求得最优鉴别矢量和分类阈值。实验证实了 FFC-SFCA 的有效性以及对两个常规聚类算法的优越性。

关键词：Fisher 准则；半模糊聚类；最优鉴别矢量
中图分类号：TP181　　**文献标识码**：A　　**文章编号**：1009-5896(2008)09-2162-04

Fuzzy Fisher Criterion Based Semi-Fuzzy Clustering Algorithm

Cao Su-qun[①②]　Wang Shi-tong[①]　Chen Xiao-feng[①]　Xie Zhen-ping[①]　Deng Zhao-Hong[①]
[①](*School of Information*, *Jiangnan University*, *Wuxi* 214122, *China*)
[②](*Department of Mechanical Engineering*, *Huaiyin Institute of Technology*, *Huaian* 223001, *China*)

Abstract: The robust Fuzzy Fisher Criterion based Semi-Fuzzy Clustering Algorithm (FFC-SFCA) for linearly separable data is presented in this paper. FFC-SFCA incorporates Fisher discrimination method with fuzzy theory using fuzzy scatter matrix. By iteratively optimizing the fuzzy Fisher criterion function, the final clustering results are obtained. FFC-SFCA exhibits its robustness and capability to obtain well separable clustering results. In addition, optimal discriminant vector and threshold of classifier can also be figured out. The experimental results for artificial and real datasets demonstrate its validity and distinctive superiority over the two conventional clustering algorithms.

Key words: Fisher criterion; Semi-fuzzy clustering; Optimal discriminant vector

1　引言

有监督情况下，常常通过 Fisher 线性判别(FLD)[1]进行投影方向上投影点类内和类间散布矩阵的优化运算，使得投影点类间尽量分开同时类内尽量密集。2002 年，Clausi 提出了 KIF(K-means Iterative Fisher)方法[2]，将 Fisher 准则巧妙地应用于无监督聚类，在纹理分割等实验中取得了可分性好的聚类结果。该算法通过 K-means 初始化与迭代 Fisher 线性判别方法组合实现，本质上属于硬划分，虽然执行效率较高，但抗噪性能较差。本文将模糊理论引入 Fisher 判别方法，给出模糊 Fisher 准则函数定义，进而提出了一种半模糊聚类的新算法 FFC-SFCA(Fuzzy Fisher Criterion based Semi-Fuzzy Clustering Algorithm)，与 KIF 算法相比，FFC-SFCA 具有 3 个优点：(1)明确给定了目标函数定义，具有严格的数学理论基础；(2)属于半模糊聚类算法，综合了传统硬聚类收敛速度快和模糊聚类的对初始化不敏感和抗噪性能强的优点；(3)该方法在对数据集实现聚类的同时，不仅可以求得最优鉴别矢量，用于特征提取和降维，而且可以求得分界线进而构造出分类器。

2007-02-05 收到，2007-09-28 改回
2004 年教育部优秀人才支持计划(NCET-04-0496)，模式识别国家重点实验室开放课题，南京大学软件新技术国家重点实验室开放课题，教育部重点科学研究项目(105087)和国防应用基础研究基金项目(A1420061266)资助课题

2　模糊 Fisher 准则

参照 Kuo-lung Wu 等在文献[3]中模糊散布矩阵定义，将模糊理论引入 Fisher 判别方法。

设一集合包含 N 个 d 维样本 $\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N$，模式类别有 c 个，在该样本空间，定义各类样本均值向量记为 $\boldsymbol{m}_i$，模糊类内散布矩阵记为 $\boldsymbol{S}_{fw}$，

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

模糊类间散布矩阵记为 $\boldsymbol{S}_{fb}$，

$$\boldsymbol{S}_{fb}=\sum_{i=1}^{c}\sum_{j=1}^{N}u_{ij}^{m}(\boldsymbol{m}_i-\overline{\boldsymbol{x}})(\boldsymbol{m}_i-\overline{\boldsymbol{x}})^{\mathrm{T}} \tag{2}$$

设 $\boldsymbol{y}=\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{x}$，在该投影空间，定义各类样本均值向量记为 $\widetilde{\boldsymbol{m}}_i$，有

$$\widetilde{\boldsymbol{m}}_i=\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{m}_i \tag{3}$$

定义模糊类内散布矩阵记为 $\widetilde{\boldsymbol{S}}_{fw}$，$\widetilde{\boldsymbol{S}}_{fw}=\sum_{i=1}^{c}\sum_{j=1}^{N}u_{ij}^{m}\cdot(\boldsymbol{y}_j-\widetilde{\boldsymbol{m}}_i)(\boldsymbol{y}_j-\widetilde{\boldsymbol{m}}_i)^{\mathrm{T}}=\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{S}_{fw}\boldsymbol{\omega}$。模糊类间散布矩阵记为 $\tilde{\boldsymbol{S}}_{fb}$，$\tilde{\boldsymbol{S}}_{fb}=\sum_{i=1}^{c}\sum_{j=1}^{N}u_{ij}^{m}(\widetilde{\boldsymbol{m}}_i-\overline{\boldsymbol{y}})(\widetilde{\boldsymbol{m}}_i-\overline{\boldsymbol{y}})^{\mathrm{T}}=\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{S}_{fb}\boldsymbol{\omega}$。

定义模糊 Fisher 准则(Fuzzy Fisher Criterion)函数:

$$J_{\mathrm{FFC}}=\frac{\tilde{\boldsymbol{S}}_{fb}}{\widetilde{\boldsymbol{S}}_{fw}}=\frac{\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{S}_{fb}\boldsymbol{\omega}}{\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{S}_{fw}\boldsymbol{\omega}}\tag{4}$$

对于线性可分数据集，将 J_{FFC} 作为聚类目标函数，当 J_{FFC} 取得极大值时，表明样本点在 $\boldsymbol{\omega}$ 方向上投影类间距离最大且类内距离最小。使用 Lagrange 乘子法求解 J_{FFC} 取得极大值时，$\boldsymbol{\omega}$，$\boldsymbol{m}_i$ 和 u_{ij} 的取值。

定义 Lagrange 函数为

$$L=\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{S}_{fb}\boldsymbol{\omega}-\lambda\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{S}_{fw}\boldsymbol{\omega}+\sum_{j=1}^{N}\lambda_j\left(\sum_{i=1}^{c}u_{ij}-1\right)\tag{5}$$

式中 λ 和 $\lambda_j(j=1,2,\cdots,n)$ 为 Lagrange 乘子。

将 L 分别对 $\boldsymbol{\omega}$，$\boldsymbol{m}_i$ 及 u_{ij} 求偏导数，并令偏导数为零，即

$$\partial L/\partial\boldsymbol{\omega}=0\tag{6}$$

$$\partial L/\partial\boldsymbol{m}_i=0\tag{7}$$

$$\partial L/\partial u_{ij}=0\tag{8}$$

对于式(6)，由于 $\boldsymbol{S}_{fb}$ 和 $\boldsymbol{S}_{fw}$ 均为对称半正定矩阵，当 $\boldsymbol{S}_{fw}$ 非奇异时，可求解得:

$$\boldsymbol{S}_{fw}^{-1}\boldsymbol{S}_{fb}\boldsymbol{\omega}=\lambda\boldsymbol{\omega}\tag{9}$$

解式(9)为求一般矩阵 $\boldsymbol{S}_{fw}^{-1}\boldsymbol{S}_{fb}$ 的本征值问题，λ 即为该矩阵的特征值，而 $\boldsymbol{\omega}$ 为对应的特征向量。

对于式(7)，可以解得

$$\boldsymbol{m}_i=\frac{\sum_{j=1}^{N}u_{ij}^{m}(\boldsymbol{x}_j-\overline{\boldsymbol{x}}/\lambda)}{\sum_{j=1}^{N}u_{ij}^{m}(1-1/\lambda)}\tag{10}$$

对于式(8)，可以解得

$$u_{ij}=\left(\frac{\lambda_j}{m(\lambda\boldsymbol{\omega}^{\mathrm{T}}(\boldsymbol{x}_j-\boldsymbol{m}_i)(\boldsymbol{x}_j-\boldsymbol{m}_i)^{\mathrm{T}}\boldsymbol{\omega}-\boldsymbol{\omega}^{\mathrm{T}}(\boldsymbol{m}_i-\overline{\boldsymbol{x}})(\boldsymbol{m}_i-\overline{\boldsymbol{x}})^{\mathrm{T}}\boldsymbol{\omega})}\right)^{\frac{1}{m-1}}\tag{11}$$

又 $\sum_{k=1}^{c}u_{kj}=1$，$j=1,2,\cdots,N$，所以有

$$\begin{aligned}1&=\sum_{k=1}^{c}u_{kj}\\&=\sum_{k=1}^{c}\left\{\lambda_j\Big/[m(\lambda\boldsymbol{\omega}^{\mathrm{T}}(\boldsymbol{x}_j-\boldsymbol{m}_k)(\boldsymbol{x}_j-\boldsymbol{m}_k)^{\mathrm{T}}\boldsymbol{\omega}-\boldsymbol{\omega}^{\mathrm{T}}(\boldsymbol{m}_k-\overline{\boldsymbol{x}})\\&\quad\cdot(\boldsymbol{m}_k-\overline{\boldsymbol{x}})^{\mathrm{T}}\boldsymbol{\omega})]\right\}^{1/(m-1)}\end{aligned}\tag{12}$$

式(11)和式(12)两式相除，得

$$\begin{aligned}u_{ij}=&\Big(\boldsymbol{\omega}^{\mathrm{T}}(\boldsymbol{x}_j-\boldsymbol{m}_i)(\boldsymbol{x}_j-\boldsymbol{m}_i)^{\mathrm{T}}\boldsymbol{\omega}-(1/\lambda)\boldsymbol{\omega}^{\mathrm{T}}(\boldsymbol{m}_i-\overline{\boldsymbol{x}})(\boldsymbol{m}_i-\overline{\boldsymbol{x}})^{\mathrm{T}}\\&\cdot\boldsymbol{\omega}\Big)^{-\frac{1}{m-1}}\Big/\left[\sum_{k=1}^{c}\Big(\boldsymbol{\omega}^{\mathrm{T}}(\boldsymbol{x}_j-\boldsymbol{m}_k)(\boldsymbol{x}_j-\boldsymbol{m}_k)^{\mathrm{T}}\boldsymbol{\omega}-(1/\lambda)\boldsymbol{\omega}^{\mathrm{T}}\right.\\&\left.\cdot(\boldsymbol{m}_k-\overline{\boldsymbol{x}})(\boldsymbol{m}_k-\overline{\boldsymbol{x}})^{\mathrm{T}}\boldsymbol{\omega}\Big)^{-\frac{1}{m-1}}\right]\end{aligned}\tag{13}$$

在模糊聚类中，通常限定 $u_{ij}\in[0,1]$，因此，对上式给出如下限定条件，若

$$\boldsymbol{\omega}^{\mathrm{T}}(\boldsymbol{x}_j-\boldsymbol{m}_i)(\boldsymbol{x}_j-\boldsymbol{m}_i)^{\mathrm{T}}\boldsymbol{\omega}\le(1/\lambda)\boldsymbol{\omega}^{\mathrm{T}}(\boldsymbol{m}_i-\overline{\boldsymbol{x}})(\boldsymbol{m}_i-\overline{\boldsymbol{x}})^{\mathrm{T}}\boldsymbol{\omega}\tag{14}$$

则 $u_{ij}=1$ 且对所有 $i'\ne i$，有 $u_{i'j}=0$。也就是说，当该条件满足时，样本点 $\boldsymbol{x}_j$ 将完全隶属于第 i 类，即在此范围内采用硬划分。此条件的几何意义为：将样本点、聚类中心以及所有样本中心向 $\boldsymbol{\omega}$ 方向投影，若样本投影点与聚类中心投影点欧氏距离小于或等于聚类中心投影点与所有样本中心投影点间欧氏距离的 $1/\sqrt{\lambda}$，则对该点采用硬划分。可以使用图 1，对此做出直观解释。

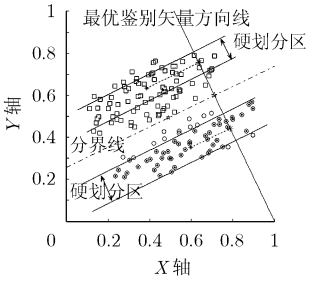

图 1 FFC-SFCA 硬划分区示意图

图 1 中“□”为第 1 类样本点，“○”为第 2 类样本点，样本点中带“.”标记的为满足式(14)而采用硬划分的点，将这些点按虚线方向向最优鉴别矢量方向线投影，根据图示，硬划分区域为垂直最优鉴别矢量方向且对称分布于聚类中心的两带状区域。

3　FFC-SFCA 算法

根据求解结果，FFC-SFCA 算法描述如下：

步骤 1　使用 k-means 算法初始化隶属矩阵 $\boldsymbol{U}$ 以及聚类中心 $\boldsymbol{m}_i$；

步骤 2　使用式(1)、式(2)式分别计算 $\boldsymbol{S}_{fw},\boldsymbol{S}_{fb}$，根据式(9)求得矩阵 $\boldsymbol{S}_{fw}^{-1}\cdot\boldsymbol{S}_{fb}$ 的最大特征值 λ，并取 $\boldsymbol{\omega}$ 为矩阵 $\boldsymbol{S}_{fw}^{-1}\cdot\boldsymbol{S}_{fb}$ 属于 λ 的模为 1 的特征向量；

步骤 3　使用式(4)计算模糊 Fisher 准则函数 J_{FFC}，若它相对上次准则函数数值的改变量小于某个阈值或者迭代次数超过设定次数，则算法停止；

步骤 4　使用式(13)，式(10)分别计算新的隶属矩阵 $\boldsymbol{U}$ 以及聚类中心 $\boldsymbol{m}_i$，返回步骤 2。

4 实验结果及其分析

4.1 实验 1 人工数据集

首先，由两个相邻的椭圆中随机生成若干数据点构成两类二维人工数据集，如图 2 所示。使用 FCM，KIF，FFC-SFCA 3 种算法对该人工数据集进行聚类实验，各算法聚类效果请见图 3–图 5。

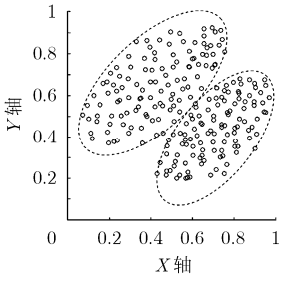

图 2 两类二维人工数据集

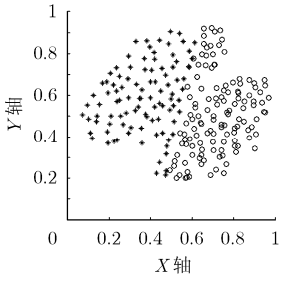

图 3 FCM 聚类效果

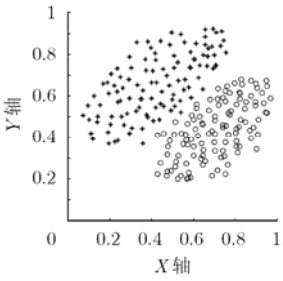

图 4 KIF 聚类效果

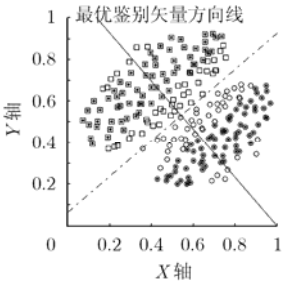

图 5 FFC-SFCA 聚类效果
(加点样本点为硬划分点)

从图中可以看出，KIF 与 FFC-SFCA 聚类结果相近，均好于 FCM。FFC-SFCA 可计算出最优鉴别矢量以及分界阈值，为降维和设计线性分类器提供了很好的依据，图 5 中实线即为最优鉴别矢量方向线，点划线即为分界线。

为进一步比较 KIF 和 FFC-SFCA，通过数据加噪实验，测试 KIF 和 FFC-SFCA 算法的抗噪性能。在图 2 中人工数据集加入一定数量[0,1]间的随机噪声点，分别使用 KIF 和 FFC-SFCA 对其进行聚类，并用著名的约当指数(Rand Index) [4]评价其聚类结果。定义 P_1,P_2 分别为对原数据集 D 和加噪数据集聚类划分结果，通过公式 $\mathrm{Rand}(P_1,P_2)=\dfrac{a+b}{n\times(n-1)/2}$ 来计算这两种划分的一致性。其中 a 表示 D 中任意两个样本 d_i,d_j 在 P_1,P_2 中同属于一类的个数；b 表示 d_i,d_j 都不属于同一类的个数；n 表示数据集 D 的样本个数。Rand Index 的范围为[0,1]，只有在 P_1,P_2 完全一致的情况下，Rand Index 值为 1。Rand Index 统计的值越小，两种划分差距越大，也说明该聚类算法受噪声点影响大。为了保证比较的准确性和公平性，记录两种算法分别进行 100 次实验后得到的 Rand Index 的均值，绘制图 6 曲线。

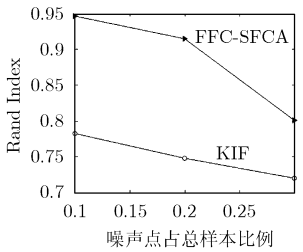

图 6 两种算法针对噪声点的 Rand Index 变化曲线

从图 6 中可以看出，KIF 算法聚类结果受噪声点影响很大，随着噪声点数目的增加，Rand Index 越来越小。FFC-SFCA 算法受噪声点影响则较小，而且在噪声点占样本总数 20%以内时，Rand Index 稳定 90%以上。因此，属模糊聚类的 FFC-SFCA 算法与 KIF 硬聚类算法相比，有显著的抗噪性能。

4.2 实验 2 标准数据集 Iris

用 UCI[5]中的 Iris 数据集来检验该聚类方法的有效性[6]。Iris 数据集包含 3 类共 150 条样本记录，每条记录有 4 个属性。分别用 FCM，KIF 和 FFC-SFCA 进行聚类，将聚类结果与原数据集样本标记对比，统计各算法错分样本数，进而计算其准确率，结果如表 1 所示。

表 1 3 种算法对 Iris 数据集聚类准确率比较

算法	错分样本数	分类准确率(%)
FCM	16	89.3
KIF	5	96.7
FFC-SFCA	2	98.7

从表 1 可以看出，对 Iris 数据集进行聚类分析，FFC-SFCA 在分类准确率上均优于 FCM 和 KIF 算法。

5 结束语

本文提出了一种半模糊聚类算法 FFC-SFCA，该方法通过模糊化散布矩阵，将模糊理论引入 Fisher 判别方法，并定义模糊 Fisher 准则函数，通过对该函数进行迭代优化运算进而实现聚类。实验表明，该方法对线性可分数据集，具有良好的聚类效果且抗噪性能显著，并且在实现聚类的同时，可以求得最优鉴别矢量用于特征提取和降维，也可以求得分界线进而构造分类器。由于 FFC-SFCA 需要求解矩阵特征值和特征向量并且样本点隶属度等需通过迭代公式进行计算，因而在高维大数据量情况下，需要进一步研究算法效率提高的途径。

参 考 文 献

[1] 边肇祺, 张学工. 模式识别. 第二版. 北京: 清华大学出版社, 2000: 87–90

[2] Clausi D A. K-means iterative Fisher(KIF) unsupervised

clustering algorithm applied to image texture segmentation. *Pattern Recognition*, 2002, 35(9): 1959–1972.

[3] Wu Kuo-Lung, Yu Jian, and Yang Miin-Shen. A novel fuzzy clustering algorithm based on a fuzzy scatter matrix with optimality tests. *Pattern Recognition Letters*, 2005, 26(4): 639–652.

[4] Rand W. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 1971, 66(336): 846–850.

[5] Blake C L and Merz C J. UCI repository of machine learning databases, Irvine. CA: University of California, Department of Information and Computer Science, http://www.ics.uci.edu/~mlearn/MLRepository.html, 1998, 7.

[6] 修宇, 王士同, 吴锡生等. 方向相似性聚类方法 DSCM. 计算机研究与发展, 2006, 43(8): 1425–1431.
Xiu Yu, Wang Shi-tong, and Wu Xi-sheng, *et al.*. The directional similarity-based clustering method DSCM. *Journal of Computer Research and Development*, 2006, 43(8): 1425–1431.

曹苏群: 男，1976 年生，讲师，博士生，从事模式识别、机器学习等研究.

王士同: 男，1964 年生，教授，博士生导师，从事人工智能、机器学习等研究.

陈晓峰: 男，1977 年生，博士生，从事人工智能、模式识别的研究.