中图法分类号: TP391.4 **文献标志码**: A **文章编号**: 1006-8961(2011)05-0784-08
论文索引信息:蔡加欣, 杨丰, 冯国灿. 改进退化的半监督模糊聚类应用于 MR 图像分割 [J]. 中国图象图形学报,2011,16(5): 784-791

改进退化的半监督模糊聚类应用于MR图像分割

蔡加欣[1)], 杨丰[1)], 冯国灿[2)]
[1)](南方医科大学生物医学工程学院, 广州 510515) [2)](中山大学数学与计算科学学院, 广州 510275)

摘 要: 半监督聚类利用少量标记样本的辅助信息来引导对大量无标记数据的分割。Pedrycz 提出的半监督 FCM(sFCM)算法应用标记样本的类别归属信息来辅助聚类,其在标记点过于稀少时会退化为无监督 FCM 算法且收敛较慢,难以应用于多数实际问题。在半监督 FCM 的基础上提出一种改进退化的半监督 FCM 算法(dsFCM),通过在 sFCM 迭代过程中设置监督成分的比重,来加大标记样本点对聚类中心的影响力,在聚类精度、速度和鲁棒性上均比半监督 FCM 有所提高,解决了标记点稀疏时的退化问题,在医学图像分割上取得了良好应用。
关键词: 聚类分析; 半监督学习; 图像分割; FCM; 种子聚类

Degeneracy-improved semi-supervised fuzzy clustering with application in MR image segmentation

Cai Jiaxin[1)], Yang Feng[1)], Feng Guocan[2)]
[1)](*School of Biomedical Engineering, Southern Medical University, Guangzhou* 510515 *China*)
[2)](*School of Mathematics and Computational Science, Sun Yet-Sen University, Guangzhou* 510275 *China*)

Abstract: Traditional clustering algorithms are always viewed as unsupervised methods for data grouping to extract information of interest from unlabeled data, while semi-supervised clustering employs limited amount of labeled data to aid the unsupervised grouping of mass unlabeled data. Pedrycz provided a semi-supervised Fuzzy C-Means algorithm (sFCM) to incorporate supervised information of labeled data as an additive part of objective function in the Fuzzy C-Means algorithm (FCM). This paper proposes a novel algorithm called Degeneracy-Improved Semi-Supervised Fuzzy C-Means algorithm (dsFCM) to fundamentally overcome the critical disadvantages of Pedrycz's sFCM algorithm, i. e., degeneracy to the classical FCM algorithm and slow convergence, particularly when applied in actual data set in which the amount of labeled points is far fewer than that of unlabeled points. Experimental results on UCI benchmark data and IBSR brain MR image data demonstrate that dsFCM algorithm can outperform sFCM algorithm in accuracy, speed and robustness. Moreover, it shows that dsFCM algorithm avoids the problems of slow convergence and degeneracy to classical FCM algorithm when applied to real world data clustering with exiguous labeled data, and presents its effectiveness for the application in interactive segmentation of medical images with a small amount of labeled data points given by user.
Keywords: clustering; semi-supervised learning; image segmentation; FCM; seed clustering

收稿日期:2010-02-10;**修回日期**:2010-05-31
基金项目:国家自然科学基金项目(60672115)。
第一作者简介:蔡加欣(1988—),男。南方医科大学生物医学工程专业硕士研究生,主要研究方向为机器学习和图像分析。E-mail:pphotail@ yahoo. com. cn。
通讯作者:杨丰, E-mail:yangf@ smu. edu. cn。

0　引　言

聚类分析是数据挖掘和模式识别中的一个基本问题,其目的是根据样本的密度分布对数据集进行划分,进而提取出感兴趣的信息。传统的聚类分析是一种无监督学习,在聚类过程中并没有利用数据集的标记信息。半监督聚类则利用少量标记样本的监督信息来辅助对大量无标记数据的划分。这些可利用的信息包括标记样本点的类别归属信息、距离或相似性信息,以及标记点之间的成对约束。根据辅助信息的不同,现有的半监督聚类方法大致可以分为基于标记属别信息的半监督聚类、基于成对约束信息的半监督聚类,以及基于距离矩阵学习的半监督聚类。

基于标记属别信息的半监督聚类的思想是在优先保证少量标记样本聚类误差最小的前提下,尽量使得全部数据的聚类误差最小[1]。Bensaid 等人提出一种部分监督 FCM 算法[2],在模糊 C 均值算法(FCM)的迭代过程中,利用标记样本的属别信息来控制聚类中心的生成,但在应用于图像分割时往往收敛速度过慢,且参数鲁棒性较差。Pedrycz 提出基于误差准则优化的半监督 FCM 算法,在 FCM 的目标函数中添加一项标记样本聚类错误的成本项,通过调整该成本项在目标函数中所占比重,来实现利用有标记样本控制对无标记样本聚类[3]。但其在标记样本点稀疏时会退化为传统的无监督 FCM,无法有效利用标记点的监督信息,难以应用于图像分割等大样本数据聚类中。随后人们相继将 Pedrycz 提出的半监督 FCM 应用于遥感图像聚类[4-5]、特征选择[6]和图像分类[7]中,并对其目标函数进行了改进[8-10],但都没有解决在标记点稀疏时大样本数据聚类的退化问题。Basu 等人提出 Seed-Kmean 算法[11],用标记样本点来作为 Kmeans 算法的初始聚类中心,可以看作基于标记属别信息的半监督聚类方法的一个特例。一般而言,基于标记属别信息的半监督聚类利用标记样本对各个类别的先验信息进行辅助聚类。若所有标记样本的聚类归属都已确定,则称为硬假设条件下的半监督聚类(种子聚类),并将这些标记样本点称为对应类别的种子,否则归为软假设条件下的半监督聚类。

医学辅助诊断和医学信息半自动处理都属于典型的半监督学习过程,医生根据需要挑选出少量典型的病历信息作为标记数据,来指导计算机对大量无标记数据进行处理。最近,如何将少量标记数据的类别归属先验和约束先验用于图像分割中的问题得到了人们的关注[12-13]。本文的主要工作是针对 Pedrycz 的半监督 FCM 算法在标记点稀疏时的退化问题,提出一种改进退化的半监督 FCM 算法,成功用于脑部 MR 图像分割,具有更好的分割速度和鲁棒性。本文的主要思想是通过在半监督 FCM 算法的迭代过程中设置标记点权重,加大标记点的中心聚合力,来克服无标记点聚类无法利用到稀疏标记点监督信息的问题,在改进聚类效果的同时也提高了收敛速度。

1　相关工作

Pedrycz 提出一种利用标记样本类别归属信息来控制无监督数据模糊聚类的方法,本文称为半监督 FCM 算法(sFCM)。sFCM 算法的目标函数是

$$\min_{U,V} J = \sum_{k=1}^{n}\sum_{i=1}^{c} u_{ik}^2 d_{ik}^2 + \alpha \sum_{k=1}^{n}\sum_{i=1}^{c} (u_{ik} - f_{ik} b_k)^2 d_{ik}^2 \quad (1)$$

式中,$\boldsymbol{X} = \boldsymbol{X}^{\mathrm{L}} \cup \boldsymbol{X}^{\mathrm{U}} = \{x_1, x_2, \cdots, x_n\}$ 是 $\mathbf{R}^P$ 上的 n 个 P 维样本数据(包括标记样本集 $\boldsymbol{X}^{\mathrm{L}}$ 和无标记样本集 $\boldsymbol{X}^{\mathrm{U}}$),$c$ 是聚类数,$0 < c < n$,$\boldsymbol{U} = [u_{ik}]_{c\times n}$ 是模糊隶属度矩阵,$\boldsymbol{V} = \{v_i\}$ 是聚类中心集合。d_{ik} 表示样本点 x_k 与聚类中心 v_i 的距离。

b_k 是一个布尔变量,指示样本点 x_k 是否为标记数据。

$$b_k = \begin{cases} 1 & x_k \in \boldsymbol{X}^{\mathrm{L}} \\ 0 & x_k \in \boldsymbol{X}^{\mathrm{U}} \end{cases} \quad (2)$$

f_{ik} 代表标记样本点 x_k 与聚类 i 之间的先验隶属度,通常由用户设定。平衡指数 α 是调整目标函数中监督成分与无监督成分之间比例的参数,按经验可取为无标记数据与标记数据间的数量比值 $\alpha_0 = |\boldsymbol{X}^{\mathrm{U}}|/|\boldsymbol{X}^{\mathrm{L}}|$。显然,当 $b_k = 0, k = 1,2,\cdots,n$,或者 $\alpha = 0$ 时,半监督 FCM 算法退化为经典的 FCM 算法。

选择以下约束条件:

$$1)\ 0 < u_{ik} < 1,\quad i = 1,2,\cdots,c;\ k = 1,2,\cdots,n$$

$$2)\ \sum_{i=1}^{c} u_{ik} = 1,\quad k = 1,2,\cdots,n \quad (3)$$

利用 Lagrange 乘子法,求得 u_{ik} 和 v_i 的迭代方程

$$u_{ik} = \frac{1}{1+\alpha}\left[\frac{1+\alpha\left(1-b_k\sum_{j=1}^{c}f_{jk}\right)}{\sum_{j=1}^{c}\frac{d_{ik}^2}{d_{jk}^2}} + \alpha f_{ik}b_k\right] \tag{4}$$

$$v_i = \frac{\sum_{k=1}^{n}u_{ik}^2 x_k}{\sum_{k=1}^{n}u_{ik}^2} \tag{5}$$

Bensaid 等人提出的部分监督 FCM 算法(pFCM)[2],是在 FCM 算法的聚类中心迭代过程中调整标记样本的权重来传播监督信息的影响,其迭代过程如下

$$u_{ik}^{\mathrm{L}} = f_{ik} \tag{6}$$

$$u_{ik}^{\mathrm{U}} = \frac{1}{\sum_{j=1}^{c}\frac{d_{ik}^2}{d_{jk}^2}} = u_{ik}^{\mathrm{FCM}} \tag{7}$$

$$v_i = \frac{\sum_{x_k\in\boldsymbol{X}^{\mathrm{L}}} w_k\,(u_{ik}^{\mathrm{L}})^2 x_k + \sum_{x_k\in\boldsymbol{X}^{\mathrm{U}}}(u_{ik}^{\mathrm{U}})^2 x_k}{\sum_{x_k\in\boldsymbol{X}^{\mathrm{L}}} w_k\,(u_{ik}^{\mathrm{L}})^2 + \sum_{x_k\in\boldsymbol{X}^{\mathrm{U}}}(u_{ik}^{\mathrm{U}})^2} \tag{8}$$

u_{ik}^{L} 表示标记样本点 x_i^{L} 的模糊隶属度,等于先验隶属度 f_{ik}。u_{ik}^{U} 是无标记样本点 x_i^{U} 的模糊隶属度,等于 x_i^{U} 在无监督 FCM 算法迭代过程中的模糊隶属度 u_{ik}^{FCM}。w_k 是在聚类中心计算中标记样本点的权重。

2 改进退化的半监督 FCM 算法

由于标记的获取成本原因,在大多数实际数据集中,无标记样本点的数量往往远远超过标记样本点,即 $|\boldsymbol{X}^{\mathrm{U}}| \gg |\boldsymbol{X}^{\mathrm{L}}|$。在这种微量标记数据集中,半监督 FCM 将会退化为经典 FCM,无标记样本点的隶属度更新与 FCM 相同,无法有效利用标记样本的监督信息。在标记点稀疏的情况下,仅在目标函数中设置监督成分比重,无法避免标记点引导作用被忽略的结果。为了克服 sFCM 在标记点稀疏时的退化问题,考虑在迭代过程中调整监督成分的比重,借鉴文献[2]在聚类中心迭代公式中设置标记样本点的权重来调整监督信息对聚类中心影响力的方法,将 sFCM 的聚类中心迭代公式改为

$$v_i' = \frac{\sum_{x_k\in\boldsymbol{X}^{\mathrm{L}}}(\alpha u_{ik}^{\mathrm{L}})^2 x_k + \sum_{x_k\in\boldsymbol{X}^{\mathrm{U}}}(u_{ik}^{\mathrm{U}})^2 x_k}{\sum_{x_k\in\boldsymbol{X}^{\mathrm{L}}}(\alpha u_{ik}^{\mathrm{L}})^2 + \sum_{x_k\in\boldsymbol{X}^{\mathrm{U}}}(u_{ik}^{\mathrm{U}})^2} \tag{9}$$

若设置 f_{ik} 满足 $0\leqslant f_{ik}\leqslant 1$,$\sum_{i=1}^{c}f_{ik}=1$,则 sFCM 中标记点和无标记点的隶属度迭代公式分别可以写成

$$u_{ik}^{\mathrm{L}} = \frac{1}{1+\alpha}\left[\frac{1}{\sum_{j=1}^{c}\frac{d_{ik}^2}{d_{jk}^2}} + \alpha f_{ik}\right] = \frac{1}{1+\alpha}u_{ik}^{\mathrm{FCM}} + \frac{\alpha}{1+\alpha}f_{ik} \tag{10}$$

$$u_{ik}^{\mathrm{U}} = \frac{1}{\sum_{j=1}^{c}\frac{d_{ik}^2}{d_{jk}^2}} = u_{ik}^{\mathrm{FCM}} \tag{11}$$

按照式(9)对聚类中心迭代过程的修改,等价于保持原来的聚类中心迭代方法不变,而在隶属度的迭代公式中按下式重新设置标记样本点和无标记样本点隶属度的权重

$$u_{ik}^{\mathrm{L}'} = u_{ik}^{\mathrm{L}} \tag{12}$$

$$u_{ik}^{\mathrm{U}'} = \frac{1}{\alpha}u_{ik}^{\mathrm{U}} = \frac{1}{\alpha}u_{ik}^{\mathrm{FCM}} \tag{13}$$

在式(13)中,α 用 $1+\alpha$ 代替,改进的隶属度和聚类中心迭代方程将写成与 sFCM 较为接近的形式

$$u_{ik}' = \frac{1}{1+\alpha}\left(\frac{1}{\sum_{j=1}^{c}\frac{d_{ik}^2}{d_{jk}^2}} + \alpha f_{ik}b_k\right) \tag{14}$$

$$v'_i = \frac{\sum_{k=1}^{n}u_{ik}^2 x_k}{\sum_{k=1}^{n}u_{ik}^2} \tag{15}$$

改进退化的半监督 FCM 算法(dsFCM)的具体步骤如下:

1)初始化。对于给定的含有标记数据和无标记数据的初始样本点集合,设置聚类数 c,根据样本分布特性选择合适的距离测度,设置平衡指数 α、最大迭代次数 *Loop* 以及迭代终止阈值 ε 的值,根据先验知识确定指示向量 $\boldsymbol{b}=\{b_k\}$ 以及标记数据的初始隶属度矩阵 $\boldsymbol{F}=[f_{ik}]$。

2)计算样本点 x_k 与聚类中心 v_i 的距离 d_{ik}。

3)根据式(14)更新隶属度矩阵 $\boldsymbol{U}=[u_{ik}]$。

4)根据式(15)更新聚类中心集合 $\boldsymbol{V}=\{v_i\}$。

5)当 $\|\boldsymbol{V}^{(t)}-\boldsymbol{V}^{(t-1)}\|<\varepsilon$ 或迭代次数 $t>Loop$ 时终止迭代,并根据此时的隶属度矩阵 $\boldsymbol{U}=[u_{ik}]$ 计算各个样本点的聚类归属。否则回到步骤 2)。

其中,先验隶属度 f_{ik} 的初始化按照以下方式完成:

在软假设情况下,假设未知标记样本点的属类,则标记样本点的初始隶属度 f_{ik} 由样本点 x_k 与全部数据的初始聚类中心 $v_i^{(0)}$ 之间的模糊隶属度确定,即

$$\text{dsFCM}: f_{ik} = \frac{1}{\sum_{l=1}^{c}\frac{d_{ik}^{(0)\,2}}{d_{lk}^{(0)\,2}}} \tag{16}$$

在硬假设情况下,标记样本点的属类信息已知,标记数据作为种子点,可采用下式初始化 f_{ik}

$$\text{hdsFCM}: f_{ik} = \begin{cases} 1 & x_k \in \boldsymbol{X}^{\text{L}} \\ 0 & x_k \in \boldsymbol{X}^{\text{U}} \end{cases} \tag{17}$$

在此情形下,dsFCM 算法称为硬修正退化半监督 FCM 算法(hdsFCM)。显然,hdsFCM 是 dsFCM 算法的推广。

与 sFCM 算法相比,dsFCM 放松了对隶属度 $\sum_{i=1}^{c} u_{ik} = 1$ 的约束条件。对于标记样本点,隶属度为

$$u_{ik}^{\text{L}} = \frac{1}{1+\alpha}u_{ik}^{\text{FCM}} + \frac{\alpha}{1+\alpha}f_{ik} \tag{18}$$

对于无标记样本点,则隶属度为

$$u_{ik}^{\text{U}} = \frac{u_{ik}^{\text{FCM}}}{1+\alpha} \tag{19}$$

设置 $\alpha \geqslant \alpha_0$,通过加大标记样本点隶属度对聚类中心的影响力,达到利用少量标记样本的监督信息来辅助对大量无标记数据快速划分的目的。

与 pFCM 算法相比,dsFCM 的标记样本点隶属度计算仍然受无标记数据部分影响,而在 pFCM 中标记点隶属度是由用户给定而始终不变的。在相同条件下,dsFCM 的聚类速度比 pFCM 更快,且参数鲁棒性更好。同样的,两种方法都没有明确的准则目标函数。

3　实验结果与分析

3.1　小样本数据集聚类实验

在 Intel CeleronM1.40 GHz,248 MB Memory 的硬件环境和 Window XP Home,MATLAB7.0 的软件环境下,先选择 UCI(University of California Irvine)机器学习数据库中的 Iris 数据集[14]进行小样本数据集聚类实验。Iris 数据集共有 150 个 4 维数据样本,分为 Setosa、Versicolour 和 Virginica 3 类,每一类各有 50 个样本点。分别使用 FCM、sFCM、dsFCM 进行聚类。FCM、sFCM、dsFCM 均使用欧氏距离测度,并选择相同的初始聚类中心,sFCM 和 dsFCM 采用相同的 8 个标记样本点,标记点初始隶属度按软假设情况设置,平衡指数 $\alpha = 1\,000$,迭代终止误差 $\varepsilon = 0.001$,最大迭代次数 $Loop = 1\,000$。定义聚类精度为

$$AC = \frac{\sum_{k=1}^{n}\delta(y_k, \hat{y}_k)}{n} \tag{20}$$

式中,y_k 是样本点 x_k 的实际聚类属别,$\hat{y}_k$ 是使用聚类算法估计的聚类属别,n 是全部样本点的总数。

3.1.1　Iris 数据集上 sFCM 和 dsFCM 的聚类性能对比

比较 FCM、sFCM 和 dsFCM 3 种算法在 Iris 数据集上的聚类精度、算法迭代次数(IT)和 CPU 运行时间(CPU run time)。实验结果如表 1 所示,sFCM 在聚类精度上较 FCM 有改善,但运算开销也比 FCM 更大,而 dsFCM 无论在聚类精度和速度上都比 FCM 和 sFCM 有较大提高。

表 1　FCM,sFCM 和 dsFCM 在 Iris 数据集上的聚类效果比较

Tab.1　Performation comparision of FCM, sFCM and dsFCM on Iris dataset

	AC/%	IT	CPU 运行时间/s
FCM	89.33	15	0.172
sFCM	98.00	36	0.594
dsFCM	99.33	2	0.031

3.1.2　平衡指数 α 的取值对聚类性能的影响

平衡指数 α 的取值对 sFCM 和 dsFCM 聚类性能的影响如图 1—3 所示。sFCM 和 dsFCM 都采用相同的 8 个标记样本点,其他参数设置不变。在图 1 中,当 $\alpha = \alpha_0 = 17.5$ 时,sFCM 和 dsFCM 的聚类精度已达到最优,当 α 继续增大时,聚类精度保持不变。在图 2、3 中,当 α 的取值增大时,sFCM 的迭代次数

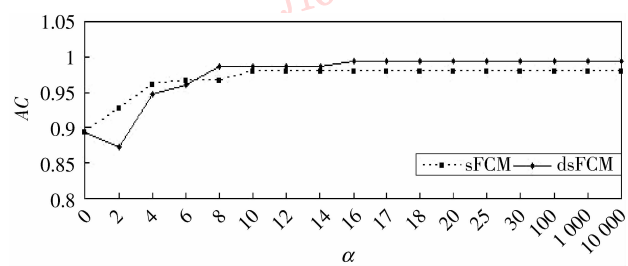

图 1　平衡指数的取值对 sFCM 和 dsFCM 聚类精度的影响

Fig.1　Relation between alpha and clustering AC

和计算时间随之增大,而 dsFCM 却随之减小。实验结果表明,在多数情况下,dsFCM 的聚类精度和速度都优于 sFCM。

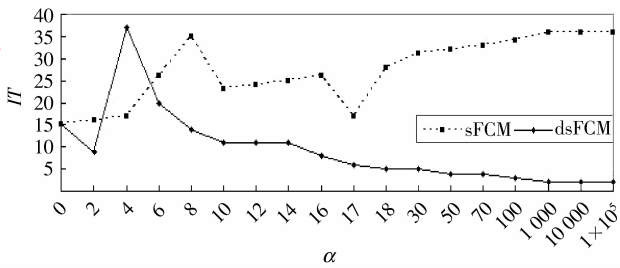

图 2 平衡指数的取值对 sFCM 和 dsFCM 迭代次数的影响

Fig. 2 Relation between alpha and iterative times

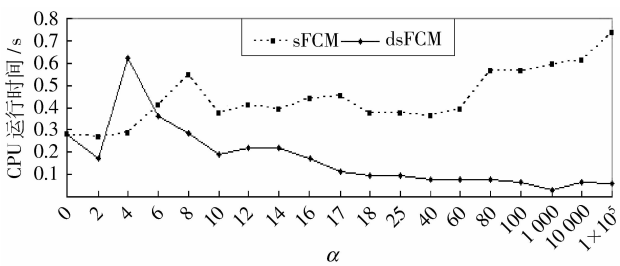

图 3 平衡指数的取值对 sFCM 和 dsFCM 计算时间的影响

Fig. 3 Relation between alpha and CPU run time

3.1.3 标记点数量变化对聚类精度的影响

图 4 中 sFCM 和 dsFCM 的平衡指数 α 取为 1 000,其他参数设置不变。实验结果如图 4 所示,当标记样本点数量增大到一定程度时,sFCM 和 dsFCM 的聚类精度都有一定程度的下降趋势,但 dsFCM 对标记样本点数量变化的敏感度比 sFCM 更低。

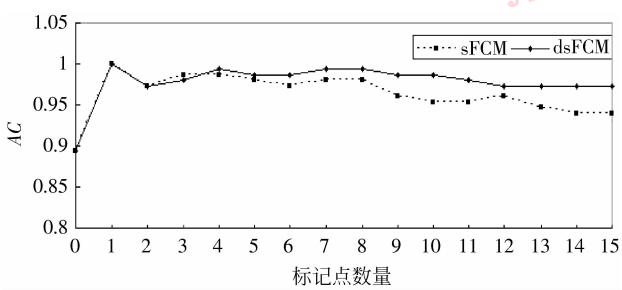

图 4 标记样本点的数量对 sFCM 和 dsFCM 聚类精度的影响

Fig. 4 Relation between number of labeled points and clustering *AC*

3.2 医学图像分割实验

选择 IBSR(internet brain segmentation repository)人脑 MR 图像库[15]作为测试数据,分别使用 FCM、sFCM、dsFCM 和 hdsFCM 4 种方法进行分割,并与专家分割结果进行比较。4 种方法均使用灰度距离测度,且初始聚类中心相同。sFCM、dsFCM 和 hdsFCM 采用同样的标记点。FCM 算法的迭代终止误差 ε 取为 0.001,sFCM、dsFCM 和 hdsFCM 的 ε 取为 10^{-16}。最大迭代次数 *Loop* 取 10 000。平衡指数 α 取值为无标记样本点与标记样本点的数量比值,即 $\alpha = \alpha_0$。

分别对无噪声图像、高斯噪声图像以及椒盐噪声图像进行分割,使用 F 测度(F-measure)、重叠率(overlap rate)、分割精确率(*AC*)等准则来量化评价分割结果。其中,F 测度按如下方法定义:白质真阳性(*TP*)代表白质区域中被划分到白质类的像素点区域面积;白质假阳性(*FP*)为非白质区域被错分为白质聚类的像素点区域面积。白质假阴性(*FN*)表示在白质区域中被错分为非白质聚类的像素点区域面积。白质的精确度定义为

$$P = \frac{TP}{TP + FP}$$

敏感度定义为

$$R = \frac{TP}{TP + FN}$$

则白质聚类的 *F* 测度为

$$F = \frac{2PR}{P + R}$$

白质重叠率表示分割图像白质区域和专家分割图像白质区域之间的交集与并集面积的比值。分割精确率表示以专家分割图为参考下的所有聚类正确的像素面积与全图像素面积的比值。显然,若 *F* 测度越大,重叠率越高或分割精确率越大,则分割效果越好。

3.2.1 无噪声图像分割实验

比较 FCM、sFCM、dsFCM 和 hdsFCM 在无噪声图像上的分割性能,实验结果如图 5 和表 2 所示。图 5(a)是一幅 256×256 的标准脑部 MR 图像,在原图像上取 13 个标记点;图 5(b)是其专家手工分割结果;图 5(c)—(f)分别是使用 FCM、sFCM、dsFCM 和 hdsFCM 进行分割的结果。表 2 给出了 4 种方法分割效果的定量评价。在分割时间上,dsFCM 最快,hdsFCM 其次,而 sFCM 未收敛。在分割精确率上,hsFCM 最高,dsFCM 其次,sFCM 则退化为 FCM,两者的分割结果相同。hdsFCM 对脑脊液成分进行了良好分割。

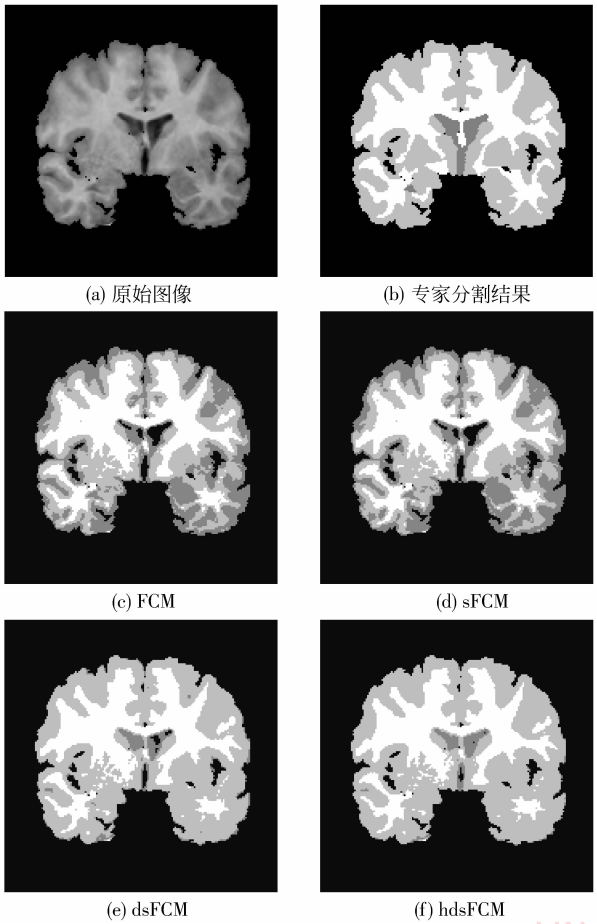

图 5　无噪声脑部 MR 图像分割结果

Fig. 5　Segmentation results on noise-free image

表 2　无噪声脑部 MR 图像分割结果比较

Tab. 2　Performation comparision of algorithms on noise-free image

			FCM/sFCM	dsFCM	hdsFCM
F 测度/%	白质	P	**93.43**	91.84	92.86
		R	78.60	**82.29**	80.50
		F	85.38	**86.80**	86.24
	灰质	P	79.70	**88.03**	86.94
		R	51.70	91.63	**94.70**
		F	62.72	89.80	**90.65**
	脊液	P	6.53	50.00	**74.05**
		R	53.11	63.40	**65.55**
		F	11.63	55.91	**69.54**
重叠率/%	白质		74.49	**76.68**	75.81
	灰质		45.69	81.49	**82.91**
	脊液		6.18	38.80	**53.31**
AC/%			93.07	97.66	**97.88**
分割速度	迭代次数		44/10 000	7	7
	CPU 运行时间/s		5.89/2 074.73	1.54	1.85

3.2.2　高斯噪声图像分割实验

比较 FCM、sFCM、dsFCM 和 hdsFCM 在高斯噪声图像上的分割性能,实验结果如图 6 和表 3 所示。其中图 6(a)是含高斯噪声(均值为 0,方差为 0.02)的 MR 图像,大小为 256×256,共取 13 个标记点;图 6(b)是专家分割结果;图 6(c)—(f) 分别是使用 FCM、sFCM、dsFCM 和 hdsFCM 分割结果。在分割时间上,hdsFCM 最快,dsFCM 其次。在分割精确率上,hdsFCM 与 dsFCM 较高,均达到 91% 以上。sFCM 退化为 FCM,两者的分割精度均为 71.36%。实验结果表明,dsFCM 和 hdsFCM 受噪声的影响比 FCM 和 sFCM 小,具有更好的鲁棒性。

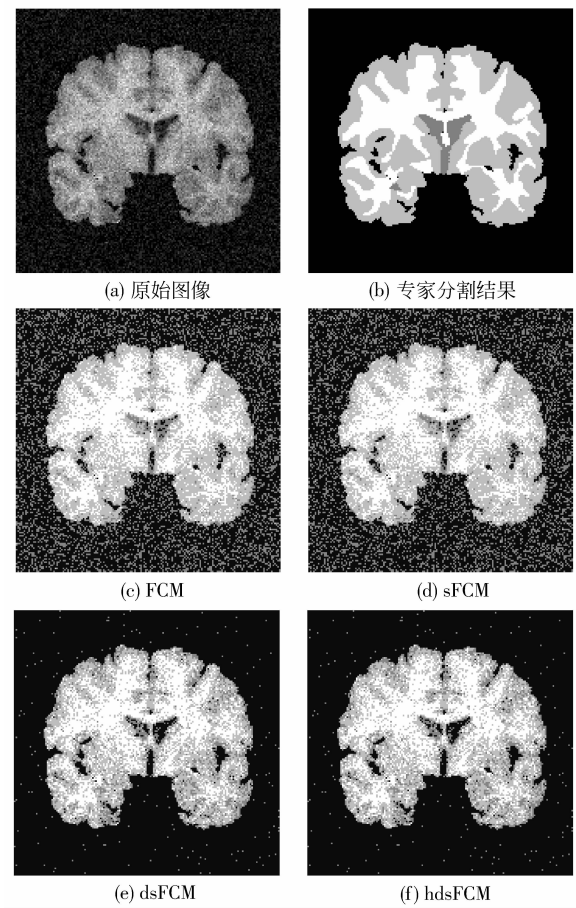

图 6　高斯噪声脑部 MR 图像分割结果

Fig. 6　Segmentation results on Gaussian noise image

表 3　高斯噪声脑部 MR 图像分割结果比较

Tab. 3　Performation comparision of algorithms on Gaussian noise image

	FCM	sFCM	dsFCM	hdsFCM
AC/%	71.36	71.36	91.36	**91.96**
IT	35	128	6	6
CPU 运行时间/s	4.985 0	18.281 0	1.547 0	**1.453 0**

3.2.3 椒盐噪声图像分割实验

比较 FCM、sFCM、dsFCM 和 hdsFCM 在椒盐噪声图像上的分割性能，实验结果如图 7 和表 4 所示。其中图 7(a)为含椒盐噪声(均值为 0，方差 0.02)的 MR 图像，大小为 256 × 256，共取 13 个标记点；图 7(b)是专家分割结果；图 7(c)—(f)分别是 FCM、sFCM、dsFCM 和 hdsFCM 分割的结果。FCM 和 sFCM 受噪声影响较大，对白质和灰质的分割效果较差，没有区分出白质和噪声；而 dsFCM 和 hdsFCM 的分割效果受噪声影响较小。

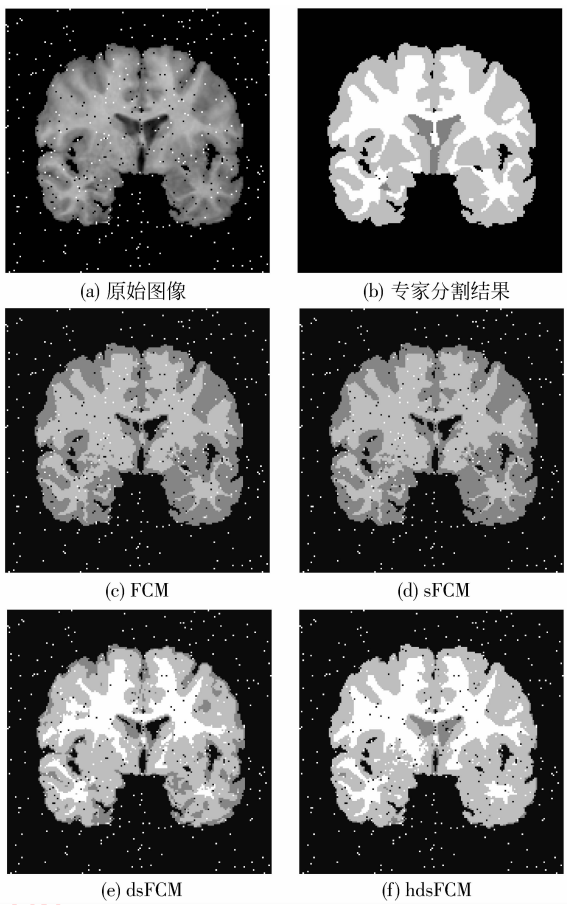

图 7 椒盐噪声脑部 MR 图像分割结果

Fig. 7 Segmentation results on salt and pepper noise image

表 4 椒盐噪声脑部 MR 图像分割结果比较

Tab. 4 Performation comparision of algorithms on salt and pepper noise image

	FCM	sFCM	dsFCM	hdsFCM
AC/%	83.59	83.59	93.15	**96.64**
IT	57	110	7	**6**
CPU 运行时间/s	8.219 0	15.534 0	1.625 0	**1.517 0**

3.2.4 与文献[2]方法的分割性能比较

比较本文使用的改进后的半监督 FCM 算法与文献[2]中提出的 pFCM 算法在医学图像上的分割性能，实验结果如表 5 所示。dsFCM 的参数设置保持不变，pFCM 采用与 dsFCM 相同的标记点、初始聚类中心、距离测度、迭代终止误差和最大迭代次数，标记点先验隶属度均按照软假设情况取值，pFCM 的权重参数 w_k 设置为 $(\alpha-1)^2$。实验结果表明，相同条件下，dsFCM 在保持较高分割精度的同时，有更好的参数鲁棒性和更快的运算速度。

表 5 dsFCM 与 pFCM 分割性能比较

Tab. 5 Performation comparision of dsFCM and pFCM

α	分割精度(*AC*)/%					
	无噪声图像		高斯噪声图像		椒盐噪声图像	
	dsFCM	pFCM	dsFCM	pFCM	dsFCM	pFCM
3 000 ~ 1×10^5	97.66	97.85	91.36	91.89	93.15	14.51

α	迭代次数					
	无噪声图像		高斯噪声图像		椒盐噪声图像	
	dsFCM	pFCM	dsFCM	pFCM	dsFCM	pFCM
3 000	7	161	6	56	8	10 000
4 000	7	10 000	6	56	8	133
5 000	7	161	6	10 000	7	131
α_0	7	164	6	58	7	129
6 000	7	10 000	6	57	7	134
7 000	7	164	6	59	7	10 000
8 000	6	159	6	10 000	7	129
9 000	6	165	5	10 000	6	124
1×10^4	6	162	5	10 000	7	127
5×10^4	5	163	5	58	5	10 000
1×10^5	5	163	5	10 000	5	134

4 结 论

大多数实际应用数据集，如图像、因特网以及文本等数据，都是微量标记数据集，即无标记数据数量远大于标记数据数量。半监督 FCM 算法应用标记数据来辅助聚类，效果高于无监督 FCM 算法。但在微量标记数据集上半监督 FCM 会退化为 FCM 且收敛较慢，无法有效利用标记点的监督信息。本文在半监督 FCM 的基础上提出一种改进退化的半监督 FCM 算法，解决了微量标记数据集上的退化问题，并应用于医学图像分割，由于利用了稀疏标记点的辅助信息，其分割性能较半监督 FCM 有所改进。实验结果表明，改进退化的半监督 FCM 算法的精度和

速度相对更优,鲁棒性更好,能够在实际数据集上应用,实用价值更高。此外本文还提出了半监督聚类中的软种子假设,证实了先验标记点在类别信息缺失的情况下仍对数据集的聚类具有辅助监督作用。继续深入研究软假设情况下的半监督聚类算法是我们下一步工作的内容。

参考文献(References)

[1] Pedrycz W. Algorithms of fuzzy clustering with partial supervision [J]. Pattern Recognition Letters, 1985,3(1): 13-20.

[2] Bensaid A, Bezdek J C, Clarke L P. Partially supervised clustering for image segmentation [J]. Pattern Recognition, 1996, 29(5): 859-871.

[3] Pedrycz W. Fuzzy clustering with partial supervision[J]. IEEE Transactions on Systems, Man, and Cybernetics, 1997, 27(5): 787-795.

[4] Luo Jiancheng, Zhou Chenghu, Yang Yan. Fuzzy clustering method with partial supervision and its application on remotely sensed classification [J]. Remote Sensing Technology and Application,1999,14(4):37-43.[骆剑承,周成虎,杨艳.具有部分监督的遥感影像模糊聚类方法研究与应用[J].遥感技术与应用,1999,14(4):37-43.]

[5] Qiu Lei, Li Guohui, Dai Kexue. Semi-supervised improved fuzzy C-means clustering to remote-sensing image [J]. Application Research of Computers, 2006,23(7):252-253.[邱磊,李国辉,代科学. 遥感图像的半监督的改进 FCM 算法[J].计算机应用研究,2006,23(7):252-253.]

[6] Marcelloni F. Feature selection based on a modified fuzzy C-means algorithm with supervision [J]. Information Sciences, 2003,15(5):201-226.

[7] Pedrycz W, Amato A, Lecce V D, et al. Fuzzy clustering with partial supervision in organization and classification of digital images[J]. IEEE Transactions on Fuzzy Systems, 2008,16(4): 1008-1026.

[8] Stutz C, Runkler T A. Classification and prediction of road traffic using application-specific fuzzy clustering[J]. IEEE Transactions on Fuzzy Systems, 2002,10 (3): 297-308.

[9] Bouchachia A, Pedrycz W. Data clustering with partial supervision[J]. Data Mining and Knowledge Discovery, 2006, 12(1):47-78.

[10] Li Chunfang,Pang Yajing,Qian Lipu, et al. Objective function of semi-supervised FCM clustering algotithm [J]. Computer Engineering and Applications,2009,45(14):128-132.[李春芳,庞雅静,钱丽璞,等. 半监督 FCM 聚类算法目标函数研究[J].计算机工程与应用,2009,45(14):128-132.]

[11] Basu S, Banerjee A, Mooney R J. Semi-supervised clustering by seeding[C] // Proceedings of the 19th International Conference on Machine Learning. San Francisco, CA, USA: Morgan Kaufmann, 2002:19-26.

[12] Duchenne O, Audibert J, Keriven R, et al. Segmentation by transduction[C]//2008 IEEE Conference on Computer Vision and Pattern Recognition. Washington, DC, USA: IEEE Computer Society Press, 2008:1-8.

[13] Yu S X, Shi J. Segmentation given partial grouping constraints [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2004, 26(2): 173-183.

[14] Fisher R A. Center for Machine Learning and Intelligent Systems [EB/OL]. UCI Machine Learning Repository. (1988-07-01) [2009-04-04]. http://archive.ics.uci.edu/ml/datasets/iris.

[15] Verne S. Center for Morphometric Analysis. Internet Brain Segmentation Repository [EB/OL]. (1996-04)[2009-03-05]. http://www.cma.mgh.harvard.edu/ibsr.