文章编号: 1001-0920(2010)01-0111-06

改进的半监督模糊聚类算法

高翠芳, 吴小俊, 张松顺
(江南大学 信息工程学院, 江苏 无锡 214122)

摘 要: 针对 Grira 等近期提出的利用点对约束的半监督模糊聚类算法,其约束项与竞争聚类算法(CA)的目标函数之间数量级不一致,造成隶属度调整过度的问题,在重新定义目标函数的基础上提出一种改进算法,约束惩罚函数采用约束点对中两个样本新的联合表达式,使数量级与经典模糊聚类算法一致. 实验结果显示,新算法的约束项与 CA 目标函数之间能很好地协调合作,并能通过对模糊隶属度的适度调整,实现更准确的聚类.

关键词: 半监督模糊聚类; 竞争聚类算法; 点对约束; 惩罚代价函数
中图分类号: TP311 文献标识码: A

An improved semi-supervised fuzzy clustering algorithm

GAO Cui-fang, WU Xiao-jun, ZHANG Song-shun
(School of Information Engineering, Jiangnan University, Wuxi 214122, China. Correspondent: WU Xiao-jun, E-mail: wu_xiaojun@yahoo.com.cn)

Abstract: The semi-supervised fuzzy clustering with pairwise constraints recently proposed by Grira is analyzed. The disagreement on the magnitude order between constraint term and objective function of competitive clustering algorithm(CA) is the main cause for the overadjustment of membership values. Aiming at this problem, an improved algorithm is proposed based on a redefined objective function. Its penalty cost function introduces a new co-expression of two samples in the pairs, which has the same magnitude order as that of the typical fuzzy clustering. Experimental results show that the constraint term of the new algorithm can achieve good agreement and cooperation with the objective function of CA, and can produce more accurate clustering results by moderately enhancing or reducing the ambiguous memberships.

Key words: Semi-supervised fuzzy clustering; CA algorithm; Pairwise constraints; Penalty cost function

1 引 言

半监督聚类与传统的有监督和无监督聚类不同,它突破了传统方法只考虑一种样本类型的局限,能对数据集中的已知和未知样本进行综合研究,具有很多优于传统方法的性能. 这种聚类方法主要适用于数据集中有少量已知样本类属信息的情况. 对于有限的已知信息,半监督聚类所利用的方法大致有两种: 一种是利用已知样本之间的点对约束[1-4]; 另一种是直接利用已知样本的类属信息[5,6]. 前者是将先验信息表示成样本点对的形式,根据两个已知样本可以(或不可以)聚为一类. 引入两种约束[7]: 属于同一类的点对和属于不同类的点对. 然后将点对约束转化成约束惩罚函数加入无监督聚类中,对原目标函数进行适当修改,得到新的半监督聚类的优化模型.

如何充分利用先验信息指导聚类过程,以达到提高聚类性能的目的,是半监督聚类需要解决的关键技术. 点对约束作为一种利用先验信息的有效途径[7],近年来受到很多关注. 围绕基于点对约束的半监督聚类算法的研究也随之展开,产生了诸多不同的半监督聚类算法,如半监督硬聚类(PCKmeans)算法[2],半监督模糊聚类(AFCC)算法[1],半监督谱聚类算法[3]等. 其中 AFCC 算法将点对约束引入竞争模糊聚类(CA)算法[8]中,希望利用半监督方法来进一步提高模糊聚类算法已有的良好性能. 但 AFCC 算法中约束项的重要程度过大,调整后的隶

收稿日期: 2009-02-06; 修回日期: 2009-08-27.
基金项目: 教育部新世纪优秀人才计划项目(NCET-06-0487); 国家自然科学基金项目(60572034, 60973094); 江苏省自然科学基金项目(BK2006081); 江南大学创新团队计划项目(JNIRT0702).
作者简介: 高翠芳(1974-), 女, 石家庄人, 博士生, 从事模式识别、计算智能等研究; 吴小俊(1967-), 男, 江苏丹阳人, 教授, 博士生导师, 从事模式识别、计算机视觉等研究.

属度偏离正常范围很多, 因此其约束项与 CA 目标函数之间不能很好地协调合作.

本文在实验数据的基础上分析了 AFCC 算法, 发现其问题的原因是约束项与 CA 目标函数在数量级上不一致. 因此根据经典模糊聚类原理[9], 本文重新定义了半监督模糊聚类算法的目标函数, 将新约束惩罚函数引入 CA 算法中, 实现了与 CA 目标函数之间很好地协调合作, 得到了一种更合理的半监督模糊聚类算法.

2 利用点对约束的半监督模糊聚类(AFCC)算法

AFCC 算法是在 CA 算法[8] 的基础上扩展得到的. 给定 $X = \{x_1, x_2, \cdots, x_N\} \subset R^L$ 为包含 N 个样本的数据集, CA 算法的目标函数为

$$J_{CA} = \sum_{k=1}^{C}\sum_{i=1}^{N}(u_{ik})^2 d^2(x_i, L_k) - B\sum_{k=1}^{C}\Big[\sum_{i=1}^{N}(u_{ik})\Big]^2, \tag{1}$$

$$\text{s.t.} \sum_{k=1}^{C} u_{ik} = 1,\ i = 1, 2, \cdots, N. \tag{2}$$

其中: C 是聚类数目; u_{ik} 是第 i 个样本属于第 k 类的隶属度; $L_k = (L_{k1}, L_{k2}, \cdots, L_{kL})$ 是第 k 类的聚类中心; $N_k = \sum_{i=1}^{N} u_{ik}$ 是所有样本属于第 k 类的隶属度之和, 即第 k 类的势. CA 算法通过各类之间的势的竞争得到聚类数目. 在 CA 算法中, 参数 B 的选择很重要, 它反映了目标函数中第 2 项(竞争项) 相对于第 1 项(FCM 项) 的重要程度, 同时还要保证两项的数量级相同. 文献[8] 给出了 B 的表达式为

$$B(t) = \frac{G\exp(-t/S)}{\sum_{k=1}^{C}\Big[\sum_{i=1}^{N}(u_{ik})\Big]^2}\Big[\sum_{k=1}^{C}\sum_{i=1}^{N}(u_{ik})^2 d^2(x_i, L_k)\Big], \tag{3}$$

其中 t 是迭代次数. B 随 t 变化, 在每次迭代过程中更新.

AFCC 算法将点对约束的惩罚代价加入 CA 算法的目标函数中, 得到了半监督模糊聚类算法. 根据已知样本是否属于同一类的先验信息, 建立两个数据集合 M 和 C, 集合中的元素是两两样本的点对. 设 M 是属于同一类的点对集, C 是属于不同类的点对集. 若 $(x_i, x_j) \in M$, 则表示已知样本 x_i 和 x_j 属于同一类. 同理, $(x_i, x_j) \in C$ 表示样本 x_i 和 x_j 属于不同类. 点对中的两个样本没有先后顺序, 且满足对称性, 即 $(x_i, x_j) \in M$ 等价于 $(x_j, x_i) \in M$. 文献[1] 给出了 AFCC 算法的目标函数

$$J_{AFCC} = \sum_{k=1}^{C}\sum_{i=1}^{N}(u_{ik})^2 d^2(x_i, L_k) - B\sum_{k=1}^{C}\Big[\sum_{i=1}^{N}(u_{ik})\Big]^2 + A\Big(\sum_{(x_i,x_j)\in M}\sum_{k=1}^{C}\sum_{l=1,l\neq k}^{C} u_{ik}u_{jl} + \sum_{(x_i,x_j)\in C}\sum_{k=1}^{C} u_{ik}u_{jk}\Big), \tag{4}$$

同样满足式(2) 的约束条件. 可以看出, AFCC 算法由 3 部分组成: FCM 项, 竞争项和约束项. 其中第 3 项是点对约束的惩罚代价, 它计算聚类过程中与已知约束相违背的代价, 包括两部分: 1) 集合 M 中同类点对的惩罚代价, 即取两个样本属于不同类的隶属度乘积; 2) 集合 C 中不同类点对的惩罚代价, 即取两个样本属于同一类的隶属度乘积. 文献[1] 给出了目标函数中两个重要参数 A 和 B 的定义, 即

$$A = \frac{N}{M}\frac{\sum_{k=1}^{C}\sum_{i=1}^{N}(u_{ik})^2 d^2(x_i, L_k)}{\sum_{k=1}^{C}\sum_{i=1}^{N}(u_{ik})^2}, \tag{5}$$

$$B(t) = \frac{G\exp(-|t-t_0|/S)}{\sum_{k=1}^{C}\Big[\sum_{i=1}^{N}(u_{ik})\Big]^2}@ \Big[\sum_{k=1}^{C}\sum_{i=1}^{N}(u_{ik})^2 d^2(x_i, L_k) + A\Big(\sum_{(x_i,x_j)\in M}\sum_{k=1}^{C}\sum_{l=1,l\neq k}^{C} u_{ik}u_{jl} + \sum_{(x_i,x_j)\in C}\sum_{k=1}^{C} u_{ik}u_{jk}\Big)\Big]. \tag{6}$$

式(5) 中 M 为约束点对数.

A 是约束项的权重系数, 体现了半监督的重要程度, 希望利用归一化性能指标来控制约束项的重要程度[10]. 当归一化程度不好时, A 较大; 相反, 归一化程度较好时, A 较小. 但式(5) 没有考虑约束项与 FCM 项在数量级上的一致性以及两者之间的协调关系.

3 改进的半监督模糊聚类算法(ISFCA)

3.1 ISFCA 算法的研究动机

将约束惩罚代价引入 CA 算法中, 对 CA 的目标函数进行修改, 这意味着约束项与 CA 目标函数之间应是辅助性关系[11], 只有约束项的重要程度合适时, 才能适当调整隶属度的幅度. 在仿真实验中发现, AFCC 算法的约束项对隶属度的调整幅度很大, 调整后的隶属度超出正常范围很多, 这表明 AFCC 算法的约束项与 CA 目标函数之间不能很好地协调合作. 根据经典模糊聚类原理, 对 AFCC 算法中约束惩罚函数的数量级进行分析.

如图 1 所示, d_{ik} 表示第 i 个样本到第 k 个类中心的距离, u_{ik} 表示第 i 个样本属于第 k 类的隶属度. 经典模糊聚类原理采用了隶属度和距离两种度量的乘积. 这两种度量不仅共同影响聚类结果, 而且共同决定目标函数值的数量级, 缺少任何一种都会使数量

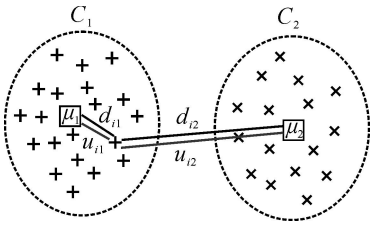

图 1　经典模糊聚类算法 FCM 的原理

级不同. 而 AFCC 算法的约束惩罚函数中只考虑隶属度而忽略了距离, 显然与 CA 目标函数的数量级不一致, 而且式(5) 定义权重系数 A时也没有考虑这一关键问题, 这就是 AFCC 算法中约束项的重要程度过大的原因. 对此, 本文将距离补充到约束惩罚函数中, 在重新定义目标函数的基础上, 提出了改进的半监督模糊聚类算法(ISFCA).

3.2　ISFCA算法描述

本文定义 ISFCA 算法的目标函数为

$$J_{ISFCA} = \sum_{k=1}^{C}\sum_{i=1}^{N}(u_{ik})^2 d^2(x_i, L_k) - B\sum_{k=1}^{C}[\sum_{i=1}^{N}(u_{ik})]^2 + A(\sum_{(x_i,x_j)\in M}\sum_{k=1}^{C}\sum_{l=1,l\neq k}^{C}u_{ik}d(x_i, L_k)u_{jl}d(x_j, L_l) + \sum_{(x_i,x_j)\in C}\sum_{k=1}^{C}u_{ik}d(x_i, L_k)u_{jk}d(x_j, L_k)), \quad (7)$$

满足式(2) 的约束条件. 式(7) 中前面两项是 CA 算法的目标函数, 第 3 项是新的约束惩罚代价函数. 该惩罚函数是隶属度与距离的乘积, 采用约束点对中两个样本联合表达的方式, 数量级相当于隶属度平方乘以距离平方. 其中: 隶属度平方由约束点对中的两个样本各自/ 提供0 一个隶属度; 同样, 距离平方由约束点对中的两个样本各自/ 提供0 一个距离. 这种定义与 CA 算法的目标函数在数量级上是一致的.

通过最小化公式(7) 中的目标函数 J_{ISFCA} 可以得到 ISFCA 算法的聚类中心和隶属度迭代公式. 聚类中心迭代公式为

$$L_{kp} = [\sum_{i=1}^{N}(u_{ik})^2 x_{ip} + \frac{A}{2}(\sum_{(x_i,x_j)\in M}\sum_{l=1,l\neq k}^{C}u_{ik}u_{jl}b_{ik}|x_{jp} - L_{lp}| + \sum_{(x_i,x_j)\in C}b_{ik}b_{jk}u_{ik}u_{jk}(x_{ip} + x_{jp}))]/(\sum_{i=1}^{N}(u_{ik})^2 + A\sum_{(x_i,x_j)\in C}b_{ik}b_{jk}u_{ik}u_{jk}). \quad (8)$$

其中

$$b_{ik} = \begin{cases} 1, & x_{ip} \geqslant L_{kp}; \\ -1, & \text{else}; \end{cases}$$

$$b_{jk} = \begin{cases} 1, & x_{jp} \geqslant L_{kp}; \\ -1, & \text{else}. \end{cases}$$

隶属度迭代公式为

$$u_{rs} = u_{rs}^{CA} + u_{rs}^{Constrains}. \quad (9)$$

其中

$$u_{rs}^{CA} = \frac{1/d^2(x_r, L_s)}{\sum_{k=1}^{C}[1/d^2(x_r, L_k)]} + \frac{B}{d^2(x_r, L_s)}(N_s - N_r), \quad (10)$$

$$u_{rs}^{Constrains} = \frac{A}{2d^2(x_r, L_s)}(\overline{Cv_r} - Cv_{rs}). \quad (11)$$

式(10) 中

$$N_s = \sum_{i=1}^{N}u_{is},\quad N_r = \frac{\sum_{k=1}^{C}[N_k/d^2(x_r, L_k)]}{\sum_{k=1}^{C}[1/d^2(x_r, L_k)]}.$$

式(11) 中

$$\overline{Cv_r} = \sum_{k=1}^{C}[(\sum_{(x_r,x_j)\in M}\sum_{l=1,l\neq k}^{C}d(x_r, L_k)u_{jl}d(x_j, L_l) + \sum_{(x_r,x_j)\in C}d(x_r, L_k)u_{jk}d(x_j, L_k))/d^2(x_r, L_k)]/\sum_{k=1}^{C}[1/d^2(x_r, L_k)],$$

$$Cv_{rs} = \sum_{(x_r,x_j)\in M}\sum_{l=1,l\neq s}^{C}d(x_r, L_s)u_{jl}d(x_j, L_l) + \sum_{(x_r,x_j)\in C}d(x_r, L_s)u_{js}d(x_j, L_s).$$

从式(9) 可以看出, ISFCA 算法的总隶属度由两部分组成: u_{rs}^{CA} 是 CA 算法的隶属度, $u_{rs}^{Constrains}$ 是约束项的隶属度. 其中: Cv_{rs} 是样本 x_r 属于第 s 类的惩罚代价, $\overline{Cv_r}$ 是样本 x_r 属于所有类的平均惩罚代价. 当 $Cv_{rs} > \overline{Cv_r}$ 时, $u_{rs}^{Constrains}$ 为负, 约束惩罚使 x_r 属于 s 类的隶属度减少; 相反, $Cv_{rs} < \overline{Cv_r}$ 时, $u_{rs}^{Constrains}$ 为正, 约束惩罚使 x_r 属于 s 类的隶属度增大. 若样本 x_r 不是集合 M 或C 中的样本, 则 $u_{rs}^{Constrains} = 0$. $u_{rs}^{Constrains}$ 可使总隶属度增大或减少, 即通过叠加约束项的隶属度可以实现对 CA 算法的隶属度进行调整.

由于新约束惩罚函数与 FCM 项的数量级一致, ISFCA 算法中参数 A的取值较简单, 只要取常数值即可. 下式给出了估算 A值的一种方法:

$$A = \frac{u_{rs}^{Constrains} \cdot 2d^2(x_r, L_s)}{\overline{Cv_r} - Cv_{rs}}, \quad (12)$$

其中 $u_{rs}^{Constrains}$ 为期望的隶属度调整幅度. 上式由式(11) 变形得到, 可有助于减少 A取值的盲目性. 另外, 由于 CA 算法中竞争项与 FCM 项的数量级一致, ISFCA 算法中仍采用式(3) 定义的 B

ISFCA 算法的主要步骤如下:

1) 初始化: 给出最大聚类数目 C_{max}(不小于期望的聚类数目), 随机初始化聚类中心, 平均初始化隶属度, 计算各类的隶属度之和(势).

2) 利用式(3) 计算 B值.

3) 利用式(9) 计算隶属度.

4) 计算各类的势 $N_s = \sum_{i=1}^{N} u_{is}(s = 1, 2, \cdots, C)$, 如果某类的势小于阈值, 则删除该类, 更新聚类数目, 归一化隶属度矩阵.

5) 利用式(8) 计算聚类中心.

6) 重复 2) ~ 5), 直至达到迭代终止条件 $|J_{ISFCA}(t) - J_{ISFCA}(t-1)| < \varepsilon$(预定义 $\varepsilon = 10^{-5}$).

ISFCA 算法对初始值 C_{max} 并不敏感, 当 C_{max} 取值大于正确的聚类数目时, 一般会在聚类初期收敛到正确的聚类数. 如果不考虑运算时间的影响, 理论上 C_{max} 可以取很大值.

4 仿真实验

通过在 IRIS 数据集上的实验结果分析了 ISFCA 算法的性能, 并与 AFCC 算法的实验数据进行了比较. IRIS 标准数据集来源于 UCI 机器学习库[12], 共分 3 类, 每类有 50 个样本. 约束点对的选取参照文献[1] 中的方法, 每次迭代前先找出最模糊的分类, 分别以隶属度 0.7 和 0.4 作为该类扩展的内、外边界(这里将位于两种边界的中间部分统称为模糊边界). 选取模糊边界上的 3 对样本, 根据已知类属情况划分到数据集 M 或 C, 得到约束点对. 实验中的参数取为 $G_0 = 0.8$, $A = 0.1$, $C_{max} = 6$.

4.1 隶属度调整幅度分析

ISFCA 算法的主要目的是通过约束惩罚函数的调整作用, 使模糊样本实现正确聚类. 图 2 显示了模糊样本 x_{67} 所属类别的隶属度调整过程.

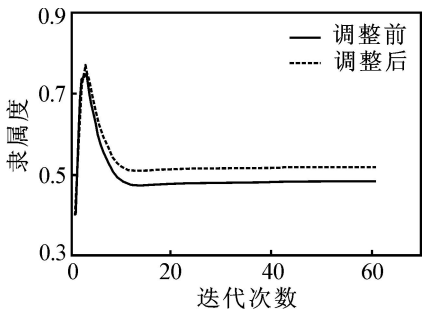

图 2 模糊样本 x_{67} 的隶属度调整过程

图 2 中, 调整前是 CA 算法的隶属度 u_{rs}^{CA}, 调整后是总隶属度 u_{rs}. 聚类过程中对模糊隶属度的适度调整, 验证了 ISFCA 算法在理论上的有效性.

表 1 和表 2 分别给出了 ISFCA 算法和 AFCC 算法中部分模糊隶属度的调整数据, 同时也反映了两种算法的约束项与 CA 目标函数之间的协调程度. 其中约束隶属度 $u_{rs}^{Constrains}$ 作为辅助调整项, 在数值上是有符号的, 正值表示它对该类的隶属度有加强作用, 负值表示对该类的隶属度有减弱作用.

表 1 ISFCA 算法中约束惩罚代价对总隶属度的调整作用

样本点	类别	u_{rs}^{CA}	$u_{rs}^{Constrains}$	u_{rs}
x_{67}	Class1:	0.0023	-0.0002	0.0021
	Class2:	0.4935	0.0091	0.5026
	Class3:	0.5042	-0.0089	0.4953
x_{73}	Class1:	0.0110	0.0006	0.0116
	Class2:	0.4605	0.0131	0.4736
	Class3:	0.5285	-0.0137	0.5148
x_{69}	Class1:	0.0188	-0.0002	0.0186
	Class2:	0.5404	0.0057	0.5461
	Class3:	0.4408	-0.0054	0.4354

表 1 中样本 x_{67} 和 x_{73} 的正确分类应是第 2 类, 而 CA 算法的聚类结果不但模糊而且是错误的. 叠加了约束隶属度之后, 样本 x_{67} 改变了总隶属度的类别属性, 实现了正确分类; 样本 x_{73} 虽然没有改变类别属性, 但属于第 2 类的隶属度增加了, 而且在以后的进一步调整中可以实现正确分类. 又如样本 x_{69} 的正确分类也是第 2 类, 虽然 CA 算法得到了正确结果, 但模糊度却较高, $u_{rs}^{Constrains}$ 的调整作用加大了模糊类别之间的差距, 使样本 x_{69} 属于第 2 类的程度加大了.

表 2 AFCC 算法中约束惩罚代价对总隶属度的调整作用

样本点	类别	u_{rs}^{CA}	$u_{rs}^{Constrains}$	u_{rs}
x_{67}	Class1:	0.0104	-0.0400	-0.0296
	Class2:	0.4860	0.4023	0.8883
	Class3:	0.5036	-0.3623	0.1413
x_{73}	Class1:	0.0143	-0.3252	-0.3109
	Class2:	0.4534	0.9771	1.4305
	Class3:	0.5322	-0.6519	-0.1197
x_{69}	Class1:	0.0201	-0.2511	-0.2310
	Class2:	0.5460	1.3293	1.8753
	Class3:	0.4339	-1.0782	-0.6443

表 2 中的 $u_{rs}^{Constrains}$ 值比表 1 中大得多, 叠加 u_{rs}^{CA} 之后, 总隶属度 u_{rs} 的调整幅度很大, 甚至超出[0, 1] 范围. 实际上在 CA 算法中, 由于竞争项的调整作用, 也会使少数样本的隶属度有微小值(小于 10^{-2}) 超出[0, 1] 范围. 文献[8] 对此进行了简单修正, 使它们重新回到正常范围内. 但 AFCC 算法中的隶属度偏离正常范围太多, 即使修正到[0, 1] 范围内, 也会影响聚类效果. 表 2 的数据表明, AFCC 算法的约束项很难与 CA 目标函数之间很好地协调合作.

4.2 重要参数分析

AFCC 算法中, 权重系数 a 可在一定程度上调整约束项的重要程度, 但约束惩罚函数中缺少距离度量, 因此 AFCC 算法的有效性很大程度上取决于 a 与距离度量的大小对比. 为了研究 a 能否将约束

隶属度控制在合理范围内, 本文将这两个关键量的数值变化进行了直观比较, 如图 3 所示(其中距离度量采用所有样本到其他类中心的距离平方的平均值 $\bar{d}^2$).

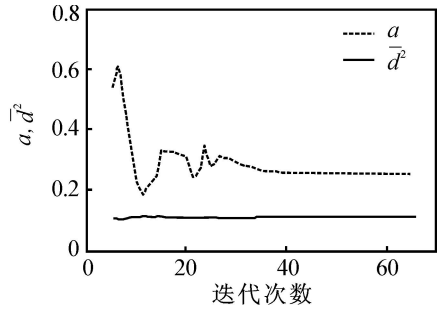

图 3　AFCC 算法中参数 a 与距离度量 d^2 的比较

图 3 表明, 整个迭代过程中 a 的值一直大于 d^2. 迭代初期 a 很大, 此时分类还不明确, 不存在模糊边界上的约束点对. 随后 a 出现波动, 这是约束惩罚项产生的调整作用, 此时 a 体现了权重系数的作用. 图 3 中的 a 值大约是 d^2 的 2 倍, 而模糊样本到其他类中心的距离比总体平均距离 d^2 更小. 显然 AFCC 算法中约束项的重要程度太大.

ISFCA 算法中竞争项和约束项的参数变化范围也会在一定程度上影响算法性能, 实验结果见表 3.

表 3　ISFCA 算法中 G_0 和 A 取不同值时的聚类效果　%

G_0	聚类精度(迭代次数)		
	A< 0.05	A∈[0.05, 0.2]	A> 0.2
< 0.1	)))	)))	)))
0.4	96.7(45)	98(45)	95.3(45)
0.8	96.7(60)	98(61)	95.3(61)
1.2	96.7(73)	98(73)	95.3(72)
> 1.5	)))	)))	)))

G_0 取不同值会影响目标函数的收敛速度, 在[0.1, 1.5] 范围内, G_0 越小, 所需迭代次数则越少, 但 G_0 的变化对聚类精度没有明显影响. 当 G_0 小于 0.1 时, 各类之间的竞争非常弱, 得不到正确聚类数目. G_0 大于 1.5 时, 算法会陷入一个大类, 造成聚类失败. 权重系数 A 的取值范围会影响聚类性能, 当 A 取值在[0.05, 0.2] 时, 有助于提高聚类精度. A 小于 0.05 时, 半监督的作用不明显, ISFCA 算法退化为 CA 算法. 当 A 大于 0.2 时, 隶属度调整幅度过大, 此时 ISFCA 算法的性能退化.

4.3　聚类效果分析

测试了约束点对数目不同时 3 种相关算法的聚类精度(聚类正确率), 如图 4 所示. 约束点对采用随机选择方式, 不同约束点对数的聚类精度取 10 次实验的平均值. 3 种算法中超出正常范围的隶属度, 均按文献[8] 的调整方法使其重新回到[0, 1] 范围.

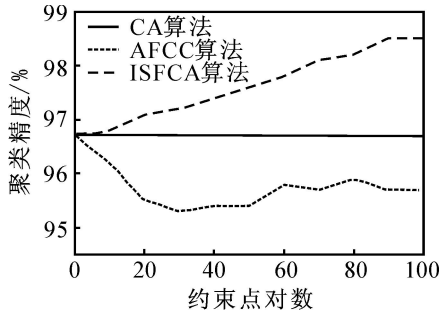

图 4　不同约束点对情况下的聚类精度

图 4 显示, AFCC 算法的聚类精度比 CA 算法降低了, 这是由于隶属度的过度调整影响了聚类效果. 特别是随机选择的约束点对如果不是模糊样本, 则调整后的隶属度会偏离正常范围更多, 很难实现正确聚类. 而 ISFCA 算法采用新的约束惩罚函数, 对隶属度进行适度调整, 增加了半监督模糊聚类算法的合理性, 聚类精度能在 CA 算法的基础上得到进一步提高.

5　结　　论

本文重新定义了基于点对约束的半监督模糊聚类算法的目标函数, 并提出了改进的 ISFCA 聚类算法. 新算法的约束惩罚函数依赖于隶属度和距离两种度量, 与经典模糊聚类算法的数量级一致. 研究发现, ISFCA 算法的约束项与 CA 算法的目标函数之间能很好地协调合作, 并能通过对隶属度的适度调整实现更准确的聚类. 实验结果显示, 新算法是一种更合理的半监督模糊聚类算法. 而且, ISFCA 算法中的约束惩罚函数具有一定的通用性, 如果将其引入其他模糊聚类算法中, 有可能得到不同类型的半监督模糊聚类算法, 具体适用情况有待于进一步研究. 另外, 通过改进 AFCC 算法中权重系数 a 的表达式, 有望从另一个角度实现约束项与 CA 目标函数之间的数量级一致.

参考文献(References)

[1] Grira N, Crucianu M, Boujemaa N. Active semi-supervised fuzzy clustering[J]. Pattern Recognition, 2008, 41(5): 1834-1844.

[2] Basu S, Banerjee A, Mooney R J. Active semi-supervision for pairwise constrained clustering[C]. Proc of the 4th SIAM Int Conf on Data Mining. Florida: SIAM, 2004: 333-344.

[3] Lu Z, Carreira-Perpin n M A. Constrained spectral clustering through affinity propagation[C]. IEEE Conf on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2008: 1-8.

[4] Yan R, Zhang J, Yang J, et al. A discriminative learning framework with pairwise constraints for video

object classification[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2006, 28(4): 5782593.

[5] Pedrycz W, Amato A, Lecce V D, et al. Fuzzy clustering with partial supervision in organization and classification of digital images[J]. IEEE Trans on Fuzzy Systems, 2008, 16(4): 10082102 6.

[6] 孙广玲, 唐降龙. 基于分层高斯混合模型的半监督学习算法[J]. 计算机研究与发展, 2004, 41(1): 1562161.
(Sun G L, Tang X L. A semi2supervised learning algorithm based on a hierarchical GMM[J]. J of Computer Research and Development, 2004, 41(1): 1562161.)

[7] Wagstaff K, Cardie C. Clustering with Instance2level Constraints[C]. Proc of the 17th Int Conf on Machine Learning. San Francisco: Morgan Kaufmann Publishers Inc, 2000: 1103211 10.

[8] Frigui H, Krishnapuram R. Clustering by competitive agglomeration[J]. Pattern Recognition, 1997, 30(7): 1109211119.

[9] Bezdek J C, Ehrlich R, Full W. FCM: The fuzzy C2means clustering algorithm [J]. Computer & Geoscience, 1984, 10(2/3): 1912203.

[10] Grira N, Crucianu M, Boujemaa N. Fuzzy clustering with pairwise constraints for knowledge2driven image categorization[J]. IEE Proc) Vision, Image and Signal Processing, 2006, 153(3): 2992304.

[11] Lange T, Law M H C, Jain A K, et al. Learning with constrained and unlabelled data[C]. IEEE Computer Society Conf on Computer Vision and Pattern Recognition. Washington DC: IEEE Computer Society, 2005, 1: 7312738.

[12] Blake C, Keogh E, Merz C J. UCI repository of machine learning database[EB/OL]. (2008211216). http://archive.ics.uci.edu/ml/datasets.html.

(上接第 109 页)

[11] Yeong2chan C, Bo2sen C. Robust tracking designs for both holonomic and nonholonomic constrained mechanical systems: Adaptive fuzzy approach[J]. IEEE Trans on Fuzzy Systems, 2000, 8(1): 46266.

[12] Yonggon L, Zak S H. Uniformly ultimately bounded fuzzy adaptive tracking controllers for uncertain systems[J]. IEEE Trans on Fuzzy Systems, 2004, 12(6): 7972811.

(上接第 114 页)

[4] 蔡元龙. 模式识别[M]. 西安: 西安电子科技大学, 1992.
(Cai Y L. Mode recognition [M]. Xi. an: Xidian University, 1992.)

[5] 傅惠民. 模糊回归分析和数据融合方法[J]. 中国安全科学学报, 2002, 12(6): 73276.
(Fu H M. Fuzzy regression analysis and data fusion [J]. J of China Safety Science, 2002, 12(6): 73276.)

[6] 谢希权, 谢邦荣, 李伟仁. 机载雷达与红外搜索跟踪装置的航迹融合研究[J]. 系统工程与电子技术, 2002, 24(4): 20222.
(Xie X Q, Xie B R, Li W R. Study on track fusion for airborne radar and infrared searching and track system [J]. Systems Engineering and Electronics, 2002, 24(4): 20222.)

[7] 傅惠民, 张应福, 张少波. 解非线性方程组的一元化方法[J]. 机械强度, 1999, 21(3): 2052207.
(Fu H M, Zhang Y F, Zhang S B. Univariate method for solving nonlinear simultaneous equations[J]. J of Mechanical Strength, 1999, 21(3): 2052207.)

[8] Rong Li R, Vesselin P Jilkov. Survey of maneuvering target tracking, Part): Dynamic models[J]. IEEE Trans on Aerospace and Electronic Systems, 2003, 39(4): 13332 1364.