DOI:10.11992/tis.201607019
网络出版地址:http://kns.cnki.net/kcms/detail/23.1538.tp.20170407.1734.004.html

基于混合距离学习的鲁棒的模糊 C 均值聚类算法

卞则康,王士同
(江南大学 数字媒体学院,江苏 无锡 214122)

摘 要:距离度量对模糊聚类算法 FCM 的聚类结果有关键性的影响。实际应用中存在这样一种场景,聚类的数据集中存在着一定量的带标签的成对约束集合的辅助信息。为了充分利用这些辅助信息,首先提出了一种基于混合距离学习方法,它能利用这样的辅助信息来学习出数据集合的距离度量公式。然后,提出了一种基于混合距离学习的鲁棒的模糊 C 均值聚类算法(HR-FCM 算法),它是一种半监督的聚类算法。算法 HR-FCM 既保留了 GIFP-FCM (Generalized FCM algorithm with improved fuzzy partitions)算法的鲁棒性等性能,也因为所采用更为合适的距离度量而具有更好的聚类性能。实验结果证明了所提算法的有效性。

关键词:距离度量;FCM 聚类算法;成对约束;辅助信息;混合距离;半监督;GIFP-FCM;鲁棒性
中图分类号:TP181 **文献标志码**:A **文章编号**:1673-4785(2017)04-0450-09

中文引用格式:卞则康,王士同.基于混合距离学习的鲁棒的模糊 C 均值聚类算法[J]. 智能系统学报, 2017, 12(4): 450-458.
英文引用格式:**BIAN Zekang, WANG Shitong. Robust FCM clustering algorithm based on hybrid-distance learning [J]. CAAI transactions on intelligent systems**, 2017, 12(4): 450-458.

Robust FCM clustering algorithm based on hybrid-distance learning

BIAN Zekang, WANG Shitong
(School of Digital Media, Jiangnan University, Wuxi 214122, China)

Abstract: The distance metric plays a vital role in the fuzzy C-means clustering algorithm. In actual applications, there is a practical scenario in which the clustered data have a certain amount of side information, such as pairwise constraints with labels. To sufficiently utilize this side information, first, we propose a learning method based on hybrid distance, in which side information can be utilized to attain a distance metric formula for the data set. Next, we propose a robust fuzzy C-means clustering algorithm (HR-FCM algorithm) based on hybrid-distance learning, which is semi-supervised. The HR-FCM inherits the robustness of the GIFP-FCM (generalized FCM algorithm with improved fuzzy partitions) and has better clustering performance due to the more appropriate distance metric. The experimental results confirm the effectiveness of the proposed algorithm.

Keywords: distance metric; FCM clustering algorithm; pairwise constraints; side information; hybrid distance; semi-supervised; GIFP-FCM; robustness

聚类分析作为一种重要的数据处理技术已经被广泛地应用到各种领域,如模式识别、数据挖掘等。在聚类分析中,需要根据数据点之间的相似或相异程度,对数据点进行区分和分类。因此对于不同的数据集,选择合适的距离度量方式对算法的聚类性能有重要的影响[1]。欧式距离是较为常用的距离度量方式,但其具有以下不足:1)采用欧式距离的方法通常是假设所有变量都是不相关的,并且数据所有维度的方差都为 1,所有变量的协方差为 0[2];2)欧式距离仅仅适用于特征空间中的超球结构,对于其他结构的数据集不太理想;3)欧式距离对噪声比较敏感,聚类结果容易受到噪声的干扰[3]。因此,欧式距离在实际应用中受到了限制。

针对这些问题,近年来提出了多种距离学习的方法,根据在距离学习过程中是否有先验的训练样本,距离学习可以分为有监督距离学习[4-6]和无监督距离学习[7-8]。在有监督距离学习的方法中,需

收稿日期:2016-07-23. 网络出版日期:2017-04-07.
基金项目:国家自然科学基金项目(61272210).
通信作者:卞则康. E-mail:bianzekang@163.com.

要借助数据集的辅助信息进行距离学习,其中辅助信息通常以约束对的形式来表示[9]。由数据集辅助信息学习得到的距离函数,可以有效地反映数据集的自身特点,对数据集具有很好的适用性。

在之前的研究中,人们提出了许多利用辅助信息进行距离学习的算法。比如,将距离学习转化为凸优化问题的方法[4]、相关成分分析法[5]、区分成分分析法等[10]。然而这些方法大多数将目标函数假设在马氏距离的框架下,本质上来说,针对马氏距离学习得到的新距离是欧式距离的线性变换,仍然有欧式距离的缺点。在含有辅助信息的数据集中,欧式距离的聚类性能和鲁棒性不理想。

因此,本文提出了一种基于混合距离学习的鲁棒模糊 C 均值聚类算法(HR-FCM)。在此算法中,数据集的未知距离被表示成若干候选距离的线性组合,在候选的距离度量中加入了非线性的距离度量。与其他有监督的聚类算法[11-12]不同的是,HR-FCM 利用数据集本身含有的少数的辅助信息进行混合距离的学习,相对于欧式距离没有考虑到数据集本身的特征,利用数据集的辅助信息学习得到的混合距离融合了数据集的一些特征,提高了提高算法的聚类性能和鲁棒性。

1　混合距离学习

1.1　混合距离

由于数据集结构特征不同,为了合理地计算不同数据集之间的距离,在距离学习中引入权重已经成为一种常用的方法。本文定义数据集中的混合距离度量的线性组合如下:

$$D(\boldsymbol{x},\boldsymbol{y}) = \sum_{i=1}^{p} \omega_i d_i(\boldsymbol{x},\boldsymbol{y})$$

$$\text{s.t.} \sum_{i=1}^{p} \omega_i = 1, 0 \leqslant \omega_i \leqslant 1, i = 1,2,\cdots,p \quad (1)$$

由文献[13]可证式(1)中 $D(\boldsymbol{x},\boldsymbol{y})$ 是一个距离函数。下面将介绍距离学习的过程。

本文的数据集分为两个部分:1)训练集,它是以约束对形式存在的辅助信息;2)用来聚类的数据集。本文将所有的训练样本集合表示为 $D = \{(\boldsymbol{x}_a^k, \boldsymbol{x}_b^k, y_k), k = 1,2,\cdots,n_p\}$,其中 n_p 为成对约束的对数。每一对约束对都是一个包含 3 个元素的元组 $(\boldsymbol{x}_a^k, \boldsymbol{x}_b^k, y_k)$,其中 $\boldsymbol{x}_a^k$ 和 $\boldsymbol{x}_b^k$ 为 d 维向量的样本点,y_k 是点对 $\boldsymbol{x}_a^k$ 和 $\boldsymbol{x}_b^k$ 之间关系的类标。当 $\boldsymbol{x}_a^k$ 和 $\boldsymbol{x}_b^k$ 属于同一类样本时,y_k 为正;相反,y_k 为负。使用 $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N\}$ 来表示 D 中所有的训练样本点,其中 N 表示样本点的个数。

在距离学习中,借鉴文献[2]的思想,利用最大边界的框架,优化目标函数:

$$\min_{\omega_i,\beta} J = \frac{1}{2}\sum_{i=1}^{p}\omega_i^2 - C\sum_{k=1}^{n} y_k\Big(\sum_{i=1}^{p}\omega_i d_i(\boldsymbol{x}_a^k, \boldsymbol{x}_b^k) - \beta\Big)$$

$$\text{s.t.} \quad y_k\Big(\sum_{i=1}^{p}\omega_i d_i(\boldsymbol{x}_a^k, \boldsymbol{x}_b^k) - \beta\Big) > 0$$

$$\sum_{i=1}^{p}\omega_i = 1, 0 \leqslant \omega_i \leqslant 1, i = 1,2,\cdots,p \quad (2)$$

式中:$d_i(\boldsymbol{x}_a^k, \boldsymbol{x}_b^k)$ 表示第 k 对约束对的第 i 个距离分量,为了便于表示,在之后的介绍中用 d_i^k 来代替 $d_i(\boldsymbol{x}_a^k, \boldsymbol{x}_b^k)$。$y_k$ 为该对样本点的对应类标,C 为惩罚因子,β 为阈值。

使用拉格朗日乘子法优化式(2),其拉格朗日函数为

$$L = \frac{1}{2}\sum_{i=1}^{p}\omega_i^2 - C\sum_{k=1}^{n_p} y_k\Big(\sum_{i=1}^{p}\omega_i d_i(\boldsymbol{x}_a^k, \boldsymbol{x}_b^k) - \beta\Big) + \lambda\Big(1 - \sum_{i=1}^{p}\omega_i\Big) + \sum_{i=1}^{p}\varphi_i\omega_i \quad (3)$$

式中:φ_i 和 λ 为拉格朗日乘子。则式(3)的 KKT 条件为

$$\begin{cases} \dfrac{\partial L}{\partial \omega_i} = 0 \\ \varphi_i \geqslant 0 \\ \varphi_i\omega_i = 0 \end{cases} \quad (4)$$

显然由式(4)无法求得 ω_i,因此先舍弃 ω_i 非负的条件,则可重新构建新的拉格朗日函数,如式(5)所示:

$$L = \frac{1}{2}\sum_{i=1}^{p}\omega_i^2 - C\sum_{k=1}^{n_p} y_k\Big(\sum_{i=1}^{p}\omega_i d_i(\boldsymbol{x}_a^k, \boldsymbol{x}_b^k) - \beta\Big) + \lambda\Big(1 - \sum_{i=1}^{p}\omega_i\Big) \quad (5)$$

可以求得

$$\omega_i = \frac{1}{p} + C\sum_{k=1}^{n_p} y_k d_i^k - \frac{C}{p}\sum_{j=1}^{p}\sum_{k=1}^{n_p} y_k d_j^k \quad (6)$$

由式(6)可以看出,即使在成功的优化过程下 ω_i 也可能出现负值,由前文看出,在考虑 ω_i 为负的条件下,无法用拉格朗日函数求解。因此,在受到加权中心模糊聚类算法[14]的启发,可以将 ω_i 改写为式(7)的形式:

$$\omega_i = \begin{cases} 0, & i \in p^- \\ \dfrac{1}{|p^+|} + C\sum_{k=1}^{n_p} y_k d_i^k - \dfrac{C}{|p^+|}\sum_{j=1}^{p}\sum_{k=1}^{n_p} y_k d_j^k, & i \in p^+ \end{cases} \quad (7)$$

式中:p^+ 表示所有使 ω_i 取正值的 i 的集合,p^- 表示无法使 ω_i 取正值的 i 的集合,使用 $|p^+|$ 和 $|p^-|$ 来分别表示集合 p^+ 和 p^- 的大小。

对于阈值 β,使用梯度下降的方法进行求解,通过求偏导,得到 β 的梯度如下:

$$\nabla_{\beta} J = C \sum_{k=1}^{n_p} y_k \tag{8}$$

为了满足约束条件: $y_k\left(\sum_{i=1}^{p} \omega_i d_i(\boldsymbol{x}_a^k, \boldsymbol{x}_b^k) - \beta\right) > 0$,参考 ω_i 的表示方式,则符合约束条件的 y_k 的集合: $n_p^+ = \left\{y_k \in D : y_k\left(\sum_{i=1}^{p} \omega_i d_i(\boldsymbol{x}_a^k, \boldsymbol{x}_b^k) - \beta\right) > 0\right\}$,重新定义了 β 的梯度公式:

$$\nabla_{\beta} J = C \sum_{k=1}^{|n_p^+|} y_k \tag{9}$$

式中: $|n_p^+|$ 表示集合 $n_p{}^+$ 的大小。使用梯度下降的方法,求解 $\beta' = \beta - \gamma \nabla_{\beta} J$,其中, γ 表示为梯度下降的学习速率,设置 $\gamma = \frac{1}{t}$。

由于集合 $n_p{}^+$ 不断改变,则等式进一步修改为如下形式:

$$\omega_i = \begin{cases} 0, & i \in p^- \\ \dfrac{1}{|p^+|} + CE_i, & i \in p^+ \end{cases} \tag{10}$$

式中:

$$E_i = \sum_{k=1}^{n_p} y_k d_i^k - \frac{1}{|p^+|} \sum_{j=p^+} \sum_{k=n_p^+}^{n_p} y_k d_j^k \tag{11}$$

具体的算法描述如下:

求解集合 p^+ 和 p^- 的算法,算法 1 如下:

1)初始化 $p^+ = \varnothing, p_0^- = \{1, 2, \cdots, p\}, h = 0$;

2) $h = h + 1, p_h^+ = p_{h-1}^+ + \{i\}, p_h^- = p_{h-1}^- - \{i\}$,其中 $i = \arg\max_{i \in p_{h-1}^-} \{E_i\}$;

3)通过式(10)计算 ω_g 并判断其是否大于 0。其中 $g = \arg\max_{i \in p_h^+} \{E_i\}$。如果 $\omega_g > 0$,则返回 2),否则设置 $p_h^+ = p_{h-1}^+, p_h^- = p_{h-1}^-$ 并终止。

求解 ω 具体算法,算法 2 步骤如下:

输入 数据矩阵 $\boldsymbol{X} \in \boldsymbol{R}^{d \times N}$,惩罚因子 C,成对约束 $(\boldsymbol{x}_a^k, \boldsymbol{x}_b^k, y_k)$,其中 $y_k = \{+1, -1\}$;

输出 距离权值 ω,阈值 β。

步骤:

1)初始化: $\omega = \omega^{(0)} = \frac{1}{p}, \beta = \beta^{(0)}$ (在初始权值下取距离的最大值作为 β 的初值)。

2)计算距离矩阵: $\boldsymbol{D}(i, k)$,

3)设置迭代步数: $t = 1$,

4)循环,直至收敛:

①更新学习率: $\gamma = 1/t, t = t + 1$

②更新训练子集:

$$n_p^+ = \left\{y_k \in D : y_k\left(\sum_{i=1}^{p} \omega_i d_i(x_a^k, x_b^k) - \beta > 0\right)\right\}$$

③计算梯度:

$$\nabla_{\beta} J = C \sum_{k \in n_p^+} y_k$$

④更新阈值: $\beta' = \beta - \gamma \nabla_{\beta} J$;

⑤更新集合 p^+ 和 p^-,使用算法 1;

⑥更新 ω。

至此,通过对训练集的距离学习,得到的权值 ω_i,从而得到新的距离函数。通过数据集本身构成的辅助信息学习得到的混合距离,对数据集自身的适应性更高,更有利于聚类效果的改善。

1.2 时间复杂度分析

这个部分主要讨论所提算法的时间复杂度,HR-FCM 算法的时间复杂度主要讨论的是混合距离学习的时间复杂度。总的来说,混合距离学习的最大时间复杂度为 $O(N^2 dp)$,其中 N 表示训练数据集中样本的个数, d 表示样本的维度, p 表示候选距离的个数。算法的主要时间消耗在求解距离矩阵 $\boldsymbol{D}$ 中,时间复杂度为 $O(Ndp)$。在迭代循环中,每一步都有一个线性的时间复杂度,为 $O(\max(N, n_p))$。

2 基于混合距离学习的鲁棒的 FCM 算法

模糊 C 均值聚类算法(FCM),它是一种基于目标函数的聚类算法,是迄今为止应用最广泛、理论最为完善的聚类算法。传统的 FCM 聚类算法使用欧式距离作为距离度量函数导致其聚类性能和鲁棒性较差。

针对传统 FCM 算法的缺点,近年来研究者们提出了一些改进的 FCM 算法,例如:基于改进的模糊划分的模糊 C 均值聚类算法(IFP-FCM)[15] 和基于改进的模糊划分的泛化的模糊 C 均值聚类算法(GIFP-FCM)[16]。IFP-FCM 算法是由 Höppner 和 Klawonn 提出的一种改进的 FCM 聚类算法。IFP-FCM 算法通过对每个数据增加一个隶属约束函数,以降低算法对噪声的敏感性,增加了算法的鲁棒性。但是此算法仍然沿用的是传统的欧式距离作为距离度量,受到 IFP-FCM 算法的启发,朱林等提出了 GIFP-FCM 算法。

在此启发下,本文提出了一种基于混合距离学习的鲁棒的 FCM 聚类算法,算法描述如下:

假设给定一个样本集合 $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n\}$,其中 n 是样本的个数,每一个样本是 d 维,预设聚类中心的集合为 $\boldsymbol{V} = \{\boldsymbol{v}_i, 1 \leqslant i \leqslant c\}$,其中 c 表示类别数。令 u_{ij} 表示第 j 个样本隶属于第 i 类的程度。则隶属

矩阵为 $\boldsymbol{U}=\{u_{ij}\mid 1\leqslant i\leqslant c,1\leqslant j\leqslant n\}$。对于每一个样本 $\boldsymbol{x}_j$,通过构造一个新的隶属约束函数 $f(u_{ij})=\sum_{i=1}^{c}u_{ij}(1-u_{ij}^{m-1})$,同时为每一个样本增加一个惩罚项 a_j 以提高算法的鲁棒性,那么得到新的目标函数为

$$J=\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^{m}d_p^2(\boldsymbol{x}_j,\boldsymbol{v}_i)+\sum_{j=1}^{n}a_j\sum_{i=1}^{c}u_{ij}(1-u_{ij}^{m-1})\ \text{s.t.}\ m>1$$
$$\sum_{i=1}^{c}u_{ij}=1,\quad 0\leqslant u_{ij}\leqslant 1,\quad 1\leqslant i\leqslant c,1\leqslant j\leqslant n \tag{12}$$

式中:$a_j=\alpha\times\min_{1\leqslant s\leqslant c}d_p^2(\boldsymbol{x}_j,\boldsymbol{v}_s)$,$0\leqslant\alpha\leqslant 1$,$\alpha$ 为抗噪参数。

使用拉格朗日乘数法对式(12)进行优化,得到新的聚类中心和隶属函数如式(13)和式(14):

$$\boldsymbol{v}_i=\frac{\sum_{j=1}^{n}u_{ij}^{m}\boldsymbol{x}_j}{\sum_{j=1}^{n}u_{ij}^{m}} \tag{13}$$

$$u_{ij}=\frac{1}{\sum_{k=1}^{c}\left(\dfrac{d_p^2(\boldsymbol{x}_j,\boldsymbol{v}_i)-\alpha\times\min\limits_{1\leqslant s\leqslant c}d_p^2(\boldsymbol{x}_j,\boldsymbol{v}_s)}{d_p^2(\boldsymbol{x}_j,\boldsymbol{v}_k)-\alpha\times\min\limits_{1\leqslant s\leqslant c}d_p^2(\boldsymbol{x}_j,\boldsymbol{v}_s)}\right)^{\frac{1}{m-1}}} \tag{14}$$

$$d_p(\boldsymbol{x}_j,\boldsymbol{v}_i)=\left(\sum_{k=1}^{n}|\boldsymbol{x}_j-\boldsymbol{v}_i|^p\right)\frac{1}{p}\quad\text{s.t.}\quad 1<p<2 \tag{15}$$

式(15)是表示样本与类中心的距离度量公式,当$p=2$ 时,式(15)就是传统的欧氏距离。

本文提出的 HR-FCM 算法,加入了距离学习的过程,通过距离学习出来的距离度量比传统的欧式距离更佳适合具有辅助信息的数据集,增加了算法的聚类性能和鲁棒性。因此,用新的混合距离 $\boldsymbol{D}$ 替换式(13)和式(14)中的距离度量 d_p,得到新的聚类中心公式(16)和隶属度计算公式(17);

$$v_i=\frac{\sum_{j=1}^{n}u_{ij}^{m}\boldsymbol{x}_j}{\sum_{j=1}^{n}u_{ij}^{m}} \tag{16}$$

$$u_{ij}=\frac{1}{\sum_{k=1}^{c}\left(\dfrac{\boldsymbol{D}^2(\boldsymbol{x}_j,\boldsymbol{v}_i)-\alpha\times\min\limits_{1\leqslant s\leqslant c}\boldsymbol{D}^2(\boldsymbol{x}_j,\boldsymbol{v}_s)}{\boldsymbol{D}^2(\boldsymbol{x}_j,\boldsymbol{v}_k)-\alpha\times\min\limits_{1\leqslant s\leqslant c}\boldsymbol{D}^2(\boldsymbol{x}_j,\boldsymbol{v}_s)}\right)^{\frac{1}{m-1}}} \tag{17}$$

式中距离度量 $\boldsymbol{D}$ 的定义如式(18):

$$\boldsymbol{D}(\boldsymbol{x},\boldsymbol{y})=\sum_{i=1}^{p}\omega_i d_i(\boldsymbol{x},\boldsymbol{y})$$
$$\text{s.t.}\quad\sum_{i=1}^{p}\omega_i=1,0\leqslant\omega_i\leqslant 1 \tag{18}$$

式中:ω_i 是通过距离学习得到的权值。

算法 3　HR-FCM 算法

输入　数据矩阵 $\boldsymbol{X}\in\boldsymbol{R}^{d\times N}$,权值向量 $\boldsymbol{\omega}$,聚类数目 c,阈值 ε,模糊指数 m,抗噪参数 α,最大迭代次数 T;

输出　最终的隶属矩阵 $\boldsymbol{U}$。

步骤:

1)初始化:$u_{ij}=u_{ij}^1$,$t=1$;

2)使用式(16)计算新的聚类中心 $\boldsymbol{v}_i^{t+1}$;

3)使用式(17)计算新的隶属矩阵 $\boldsymbol{u}_{ij}^{t+1}$;

4)如果 $\|\boldsymbol{U}^{t+1}-\boldsymbol{U}^t\|<\varepsilon$ 或者 $t>T$,输出最终的隶属矩阵,否则 $t=t+1$ 返回 2)。

HR-FCM 算法通过使用距离学习得到的新的混合距离代替传统 FCM 算法中的欧式距离,进一步增加了算法的抗噪性能。再者,通过数据集本身的辅助信息进行距离的学习的得到的混合距离,比原有的欧式距离更加适合数据集,提高了算法的适用性。HR-FCM 算法与传统的 FCM 算法相比,具有更佳的聚类性能和鲁棒性。

3　实验研究和分析

本章通过实验检测本文提出的 HR-FCM 算法的聚类性能和鲁棒性能。本章的实验主要分为两个部分:1)将本文提出的 HR-FCM 算法与现有的基于欧氏距离的聚类算法作比较,如:FCM、K-means 和 K-medoids,检测算法的聚类性能;2)主要是检测算法的鲁棒性能,通过对实验数据加入不同程度的随机噪声,并与 FCM 算法和 GIFP-FCM 算法作比较。

3.1　实验设置和实验数据

本文的实验参数设置如下:阈值 $\varepsilon=10^{-5}$,最大迭代次数 $T=300$,模糊指数 $T=300$,$m\in\{1.5,2,3,4\}$,$\alpha\in\{0.5,0.7,0.9,0.99\}$。为了实验结果的公平,重复每次聚类过程 20,实验结果取均值。

实验中对于候选距离的选取,选择了基于欧式距离的含有方差的距离分量 $d_1(\boldsymbol{x},\boldsymbol{y})$,非线性的距离分量 $d_3(\boldsymbol{x},\boldsymbol{y})$,曼哈顿距离分量 $d_2(\boldsymbol{x},\boldsymbol{y})$。由这 3 种距离分量线性组合后的混合距离 $\boldsymbol{D}(\boldsymbol{x},\boldsymbol{y})$ 是一个非线性的距离函数。本文预设的 3 个距离度量如式(19)所示:

$$\begin{cases}d_1(\boldsymbol{x},\boldsymbol{y})=(\boldsymbol{x}-\boldsymbol{y})^{\mathrm{T}}\dfrac{1}{\sigma^2}(\boldsymbol{x}-\boldsymbol{y})\\ d_2(\boldsymbol{x},\boldsymbol{y})=\sum\limits_{i=1}^{d}\dfrac{|\boldsymbol{x}_i-\boldsymbol{y}_i|}{\sigma^2}\\ d_3(\boldsymbol{x},\boldsymbol{y})=1-\exp\left(-\dfrac{-3\|\boldsymbol{x}-\boldsymbol{y}\|^2}{\sigma^2}\right)\end{cases} \tag{19}$$

本文选取的实验数据集均来自 UCI 数据集,数据集细节如表 1。由于 UCI 数据集中没有约束对形式的辅助信息,需要选取数据集中的一部分带标签的数据集构成约束对作为训练集。其中,拥有相同的类标的样本点构成正约束对,不同的类标的构成负约束对,选取相同数目的正负约束对进行距离学习。对于本文中的数据集,前 6 个取 10%的数据集构成训练集,最后两个取 1%的数据集。

在抗噪声实验中,在数据集中随机加入 10%和 20%的高斯白噪声(SNR = 40 db 或者 30 db),分别计算本文提出的 HR-FCM 算法、传统 FCM 聚类算法和 GIFP-FCM 算法的聚类性能。

表 1 数据集信息

Table 1 Description of data sets

数据集	样本数	特征数	类别数
breast	683	10	2
wdbc	569	30	2
vowel	528	10	10
sonar	208	60	2
wine	178	13	3
vehicle	846	18	4
led	3 200	24	10
waveform	5 000	21	3

3.2 评价方法

为了评估算法的聚类效果,本文采用了一些标准的评价方法,包括归一化互信息(NMI)[17]和芮氏指数(RI)[18-19],这些将用来评价 HR-FCM 算法与 FCM 的聚类效果。

$$\mathrm{NMI}(X,Y)=\frac{I(X,Y)}{\sqrt{H(X)\cdot H(Y)}} \tag{20}$$

$$\mathrm{RI}(X,Y)=\frac{a+b}{n\times(n-1)/2} \tag{21}$$

式中:X 定义了已知标签的原始数据,Y 定义了对未知标签的原始数据的聚类结果,$I(X,Y)$定义了 X 和 Y 之间的互信息,$H(X)$和 $H(Y)$分别代表了 X 和 Y 的熵,a 定义了 X 和 Y 中任意两个具有相同类标签并且属于同一个样本的数目,b 定义了 X 和 Y 中任意两个具有不同标签并且属于不同类的样本的个数,n 表示原始样本的个数。显而易见,NMI 和 RI 的值都是介于 0~1 的,NMI 和 RI 的值越大,表示 X 和 Y 之间的相似度就越高,即算法的效果越好。

3.3 实验结果和分析

在第 1 部分的实验中,为了检测算法的聚类性能,设置 HR-FCM 算法中的抗噪参数 $\alpha=0$,比较使用了混合距离的 HR-FCM 算法与使用欧氏距离的 FCM 算法和其他常用的基于欧氏距离的聚类算法,检测混合距离对聚类性能的影响。选取上述 7 个数据集作为本次实验的实验数据集,每组数据集运行 20 次,实验结果选取 RI 和 NMI 值的均值,实验结果如图 1 和图 2 所示。

第 1 部分的实验结果表明:对于小数据集,HR-FCM 算法的聚类性能不仅比传统的基于欧氏距离的 FCM 聚类算法要好,也比基于欧氏距离的 K-mean 和 K-medoids 的聚类算法性能好。对于大样本数据集,由于样本数对于 FCM 聚类算法的影响比 K-means 和 K-medoids 的大,此时,距离度量对于聚类性能的影响力下降,因此,K-means 和 K-medoids 算法的聚类性能较佳,但是与传统的 FCM 聚类算法相比,本文提出的基于混合聚类的 HR-FCM 算法具有较好的聚类性能,如 waveform 数据集。

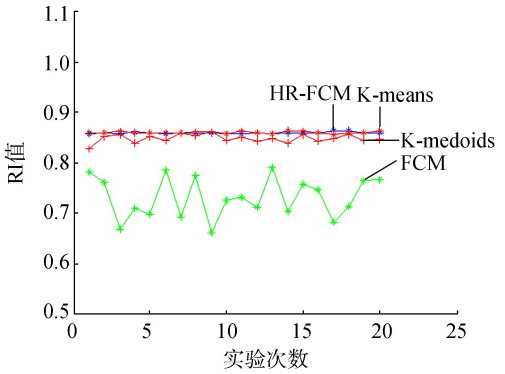
从表 2 的实验中还可以得到,对于高维的数据集,4 种算法聚类性能都有一定的下降。由于数据维度的增加,单个样本的信息增加,在这些信息中存在着不同类型的信息,本文提出的混合距离度量函数相比传统的欧氏距离能够充分地度量出样本之间的距离,学习出样本之间的有效信息,提高算法的聚类性能。对于数据集 wdbc 和 sonar,表 2 的实验结果显示了混合距离的有效性。

表 2 聚类算法性能

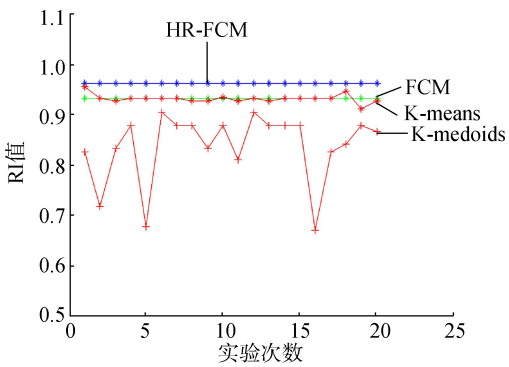
Table 2 The performance of clustering algorithm

算法	breast		wdbc		sonar		vehicle		waveform	
	RI	NMI	RI	NMI	RI	NMI	RI	NMI	RI	NMI
HR-FCM	0.886 9	0.672 0	0.851 1	0.594 2	0.508 8	0.017 1	0.661 6	0.170 9	0.664 4	0.349 0
FCM	0.519 9	0.004 5	0.750 4	0.467 2	0.503 6	0.009 1	0.650 6	0.180 2	0.660 0	0.320 9
K-means	0.520 6	0.004 7	0.750 4	0.467 2	0.503 2	0.643 0	0.652 3	0.186 8	0.667 3	0.362 2
K-medoids	0.520 6	0.004 7	0.770 7	0.499 9	0.502 2	0.007 1	0.650 9	0.186 6	0.670 1	0.368 1

为了检测算法在多类样本数据集中的聚类性能,实验中也使用了多类样本数据集 vowel。由于多类样本中,随着类别的增加,对于样本之间的距离度量的难度增加,因此算法的聚类性能受到一定的影响,如图 1 和图 2 中,对于少类别数据集 wine,各个算法的 RI 和 NMI 值较高,对于多类别数据集 vowel,各个算法的指标出现较大幅度的下降。图 1 和图 2 的实验结果也表明,对于多样本数据集,本文提出的 HR-FCM 算法的聚类性能与其他 3 种算法的聚类性能都出现了一定的下降,但是本文提出的算法较之其他 3 种算法的稳定性较高。



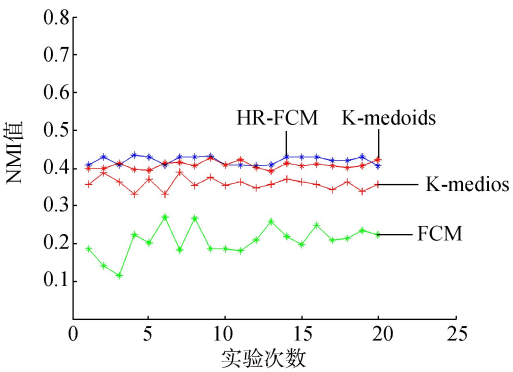
(a) vowel



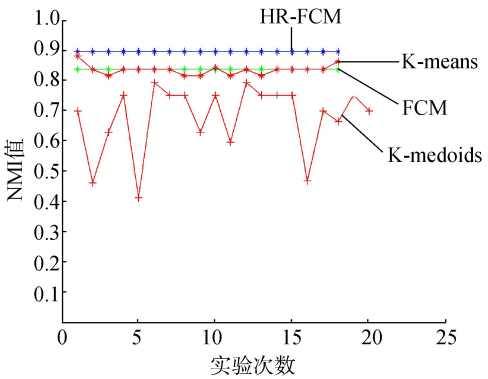
(b) wine

图 1　各数据集的 RI 值

Fig.1　The RI on the different data sets



(a) vowel



(b) wine

图 2　各数据集的 NMI 值

Fig.2　The NMI on the different data sets

在第 2 部分的实验中,为了检测本文算法的鲁棒性能,设置了在两个多类别样本数据集上的对比实验,这两个数据集分别是小样本量的低维数据集 vowel 和大样本量的高维数据集 led。本次实验主要是为了检测 HR-FCM 算法与传统的 FCM 算法和 GIFP-FCM 算法在聚类性能和鲁棒性能上面的差别,因此本次实验,比较了 3 种算法在不同模糊指数的情况下的聚类性能,通过改变噪声的添加比例,比较算法的鲁棒性能。为了进一步检测 HR-FCM 算法受抗噪参数的影响,本次实验设置不同的参数取值,通过比较聚类指标的变化显示算法的鲁棒性的变化。

从表 3 和表 4 的结果中容易看出在相同的模糊指数的情况下,HR-FCM 算法的聚类性能和鲁棒性大多强于传统的 FCM 算法。在模糊指数一定的情况下,随着抗噪参数 α 的增加,HR-FCM 算法的鲁棒性越来越强。在加入噪声后,算法的聚类性能收到了一定的影响,算法的聚类性能下降,FCM 聚类算法的聚类性能下降较多。从表 4 的实验结果中容易看出,在实验数据集为大数据集的情况下,本文提出的 HR-FCM 算法的聚类性能和鲁棒性强于传统的 FCM 算法。由表 3 和表 4 的结果也可以看出本文提出的 HR-FCM 算法的聚类结果要优于 GIFP-FCM 算法。由于本文使用混合距离代替传统的欧氏距离,因此本文的 HR-FCM 算法的鲁棒性能强于 GIFP-FCM 算法。

综上所述,对于含有辅助信息的数据集,本文提出的 HR-FCM 算法由于采用混合距离,使得算法具有较好的适应性,比传统的使用欧氏距离作为距离度量的 FCM 算法和 GIFP-FCM 算法具有更好的聚类性能和鲁棒性。

表 3 噪声实验结果

Table 3 The results on the data sets with or without random noises

数据集 vowel(SNR = 40 dB)			原始数据		加入 10%的随机噪声		加入 20%的随机噪声	
			RI	NMI	RI	NMI	RI	NMI
$m=1.5$	FCM		0.868 5	0.443 7	0.868 7	0.442 8	0.866 1	0.427 6
	HR-FCM	$\alpha=0.5$	0.862 4	0.429 4	0.861 2	0.421 1	0.861 2	0.408 5
		$\alpha=0.7$	0.864 8	0.438 4	0.862 4	0.412 8	0.860 8	0.412 1
		$\alpha=0.9$	0.863 2	0.423 4	0.860 3	0.412 8	0.860 9	0.411 4
		$\alpha=0.99$	0.863 3	0.424 3	0.859 5	0.415 7	0.859 2	0.411 5
	GIFP-FCM	$\alpha=0.5$	0.853 7	0.389 8	0.857 7	0.394 0	0.857 3	0.394 1
		$\alpha=0.7$	0.857 8	0.396 1	0.858 4	0.399 2	0.860 0	0.396 7
		$\alpha=0.9$	0.854 1	0.391 2	0.857 8	0.395 6	0.861 3	0.394 7
		$\alpha=0.99$	0.847 9	0.346 2	0.863 8	0.418 5	0.858 8	0.376 3
$m=2$	FCM		0.736 9	0.304 9	0.727 0	0.286 5	0.722 6	0.292 0
	HR-FCM	$\alpha=0.5$	0.857 9	0.410 4	0.855 8	0.399 3	0.856 4	0.399 8
		$\alpha=0.7$	0.858 9	0.418 2	0.858 7	0.414 5	0.860 1	0.409 0
		$\alpha=0.9$	0.861 9	0.430 4	0.862 0	0.420 9	0.861 3	0.416 5
		$\alpha=0.99$	0.861 4	0.428 3	0.861 0	0.413 3	0.860 9	0.417 7
	GIFP-FCM	$\alpha=0.5$	0.855 7	0.380 9	0.858 8	0.389 1	0.860 3	0.381 9
		$\alpha=0.7$	0.855 9	0.399 7	0.854 2	0.379 7	0.855 9	0.398 3
		$\alpha=0.9$	0.855 2	0.377 8	0.855 4	0.383 8	0.855 9	0.403 0
		$\alpha=0.99$	0.861 0	0.421 1	0.859 9	0.400 0	0.853 3	0.350 4
$m=3$	FCM		0.740 2	0.304 6	0.747 8	0.316 2	0.727 8	0.291 6
	HR-FCM	$\alpha=0.5$	0.856 0	0.401 1	0.856 2	0.396 7	0.853 0	0.383 7
		$\alpha=0.7$	0.855 1	0.393 7	0.855 7	0.394 8	0.852 9	0.395 5
		$\alpha=0.9$	0.856 0	0.411 7	0.855 8	0.406 4	0.853 9	0.393 3
		$\alpha=0.99$	0.860 9	0.425 1	0.858 7	0.408 2	0.859 5	0.417 6
	GIFP-FCM	$\alpha=0.5$	0.773 3	0.298 7	0.775 4	0.269 3	0.772 8	0.300 5
		$\alpha=0.7$	0.860 6	0.383 7	0.854 9	0.380 0	0.848 7	0.353 9
		$\alpha=0.9$	0.860 1	0.399 9	0.851 9	0.388 2	0.858 8	0.407 9
		$\alpha=0.99$	0.842 9	0.348 0	0.854 6	0.393 3	0.859 5	0.388 5
$m=4$	FCM		0.750 9	0.325 9	0.757 6	0.322 9	0.749 8	0.309 5
	HR-FCM	$\alpha=0.5$	0.853 8	0.387 4	0.855 0	0.386 8	0.854 4	0.384 2
		$\alpha=0.7$	0.851 9	0.384 3	0.852 7	0.383 3	0.850 4	0.370 2
		$\alpha=0.9$	0.853 4	0.410 5	0.849 8	0.384 9	0.850 6	0.391 9
		$\alpha=0.99$	0.855 3	0.419 7	0.852 9	0.416 9	0.855 5	0.414 2
	GIFP-FCM	$\alpha=0.5$	0.737 9	0.184 7	0.766 7	0.245 0	0.777 5	0.237 1
		$\alpha=0.7$	0.834 9	0.359 8	0.849 8	0.388 9	0.834 8	0.358 5
		$\alpha=0.9$	0.854 4	0.381 5	0.856 7	0.378 0	0.854 0	0.365 1
		$\alpha=0.99$	0.842 2	0.382 1	0.851 1	0.363 4	0.849 8	0.348 2

表 4　噪声实验结果

Table 4　The results on the data sets with or without random noises

数据集 led(SNR=30 dB)			原始数据		加入 10%的随机噪声		加入 20%的随机噪声	
			RI	NMI	RI	NMI	RI	NMI
$m=1.5$	FCM		0.662 4	0.262 9	0.656 2	0.265 6	0.656 0	0.265 3
	HR-FCM	$\alpha=0.5$	0.854 2	0.453 0	0.852 9	0.451 9	0.851 6	0.452 8
		$\alpha=0.7$	0.887 9	0.457 3	0.888 7	0.459 3	0.888 4	0.458 8
		$\alpha=0.9$	0.886 5	0.462 7	0.886 4	0.460 7	0.887 7	0.462 6
		$\alpha=0.99$	0.885 0	0.456 8	0.885 1	0.457 4	0.886 5	0.460 4
	GIFP-FCM	$\alpha=0.5$	0.555 7	0.293 4	0.555 7	0.293 0	0.556 2	0.293 2
		$\alpha=0.7$	0.789 8	0.354 3	0.794 6	0.354 9	0.805 1	0.363 7
		$\alpha=0.9$	0.875 4	0.402 1	0.873 5	0.399 8	0.876 1	0.406 6
		$\alpha=0.99$	0.879 4	0.406 1	0.875 3	0.411 4	0.875 7	0.414 8
$m=2$	FCM		0.718 2	0.242 9	0.703 5	0.255 2	0.697 3	0.250 7
	HR-FCM	$\alpha=0.5$	0.630 8	0.294 7	0.629 4	0.299 7	0.615 4	0.292 9
		$\alpha=0.7$	0.847 2	0.450 2	0.845 4	0.451 1	0.845 4	0.450 8
		$\alpha=0.9$	0.885 4	0.449 9	0.884 3	0.448 4	0.886 7	0.454 4
		$\alpha=0.99$	0.887 6	0.464 3	0.886 1	0.461 9	0.886 9	0.464 1
	GIFP-FCM	$\alpha=0.5$	0.573 3	0.290 9	0.605 1	0.276 5	0.574 5	0.288 0
		$\alpha=0.7$	0.577 2	0.292 1	0.581 7	0.291 0	0.572 9	0.288 1
		$\alpha=0.9$	0.877 3	0.415 9	0.876 4	0.412 3	0.877 0	0.415 7
		$\alpha=0.99$	0.876 8	0.412 7	0.879 1	0.419 6	0.878 3	0.429 6
$m=3$	FCM		0.746 2	0.237 4	0.744 8	0.233 8	0.735 5	0.244 0
	HR-FCM	$\alpha=0.5$	0.645 6	0.283 5	0.645 3	0.284 9	0.649 6	0.278 6
		$\alpha=0.7$	0.674 9	0.341 6	0.674 2	0.339 9	0.676 1	0.339 5
		$\alpha=0.9$	0.889 0	0.460 8	0.889 2	0.461 8	0.888 6	0.459 0
		$\alpha=0.99$	0.885 7	0.452 4	0.884 3	0.450 4	0.885 8	0.453 3
	GIFP-FCM	$\alpha=0.5$	0.637 5	0.271 9	0.676 6	0.249 7	0.598 8	0.286 2
		$\alpha=0.7$	0.559 5	0.292 3	0.564 9	0.288 8	0.557 8	0.292 9
		$\alpha=0.9$	0.856 4	0.473 9	0.854 3	0.435 1	0.857 1	0.474 5
		$\alpha=0.99$	0.875 8	0.406 9	0.874 5	0.399 6	0.874 2	0.402 8
$m=4$	FCM		0.777 1	0.225 7	0.773 7	0.220 2	0.766 1	0.227 8
	HR-FCM	$\alpha=0.5$	0.679 9	0.284 7	0.639 8	0.277 2	0.674 0	0.276 0
		$\alpha=0.7$	0.639 9	0.325 7	0.639 8	0.324 4	0.640 3	0.326 7
		$\alpha=0.9$	0.872 5	0.464 4	0.874 8	0.468 3	0.873 0	0.463 6
		$\alpha=0.99$	0.882 6	0.448 4	0.883 9	0.449 1	0.883 0	0.443 2
	GIFP-FCM	$\alpha=0.5$	0.694 5	0.260 0	0.717 3	0.238 6	0.765 0	0.226 0
		$\alpha=0.7$	0.563 5	0.291 7	0.576 3	0.287 4	0.646 8	0.263 5
		$\alpha=0.9$	0.806 5	0.409 0	0.824 6	0.429 3	0.815 4	0.419 6
		$\alpha=0.99$	0.874 1	0.398 4	0.875 9	0.406 4	0.873 7	0.395 4

4 结束语

在聚类的实际应用中,大多数数据集都含有一定量的辅助信息,这些辅助信息中含有重要的数据特征,但是这些辅助信息在聚类过程中常常被忽略。本文提出了一种利用数据集的辅助信息进行距离学习的方法,进而提出了一种改进的 FCM 算法 HR-FCM。用数据集的辅助信息进行距离学习得到的混合函数,不仅能够反映出数据集本身的特征,而且比欧式距离更加契合数据集,更加适合于实际应用。在含有辅助信息的数据集中,本文提出的 HR-FCM 算法具有较好的聚类性能和鲁棒性。实验结果证明了结论。

参考文献:

[1]王骏, 王士同. 基于混合距离学习的双指数模糊 C 均值算法[J]. 软件学报, 2010, 21(8): 1878-1888.
WANG Jun, WANG Shitong. Double indices FCM algorithm based on hybrid distance metric learning[J]. Journal of software, 2010, 21(8): 1878-1888.

[2]WU L, HOI S C H, JIN R, et al. Learning bregman distance functions for semi-supervised clustering[J]. IEEE transactions on knowledge and data engineering, 2012, 24(3): 478-491.

[3]WU K L, YANG M S. Alternative c-means clustering algorithms[J]. Pattern recognition, 2002, 35(10): 2267-2278.

[4]XING E P, NG A Y, JORDAN M I, et al. Distance metric learning, with application to clustering with side-information[J]. Advances in neural information processing systems, 2003, 15: 505-512.

[5]BAR-Hillel A, HERTZ T, SHENTAL N, et al. Learning a mahalanobis metric from equivalence constraints[J]. Journal of machine learning research, 2005, 6(6): 937-965.

[6]郭瑛洁, 王士同, 许小龙. 基于最大间隔理论的组合距离学习算法[J]. 智能系统学报, 2015, 10(6): 843-850.

[7]YE J, ZHAO Z, LIU H. Adaptive distance metric learning for clustering[C]//IEEE Conference on Computer Vision and Pattern Recognition. Minneapolis, USA, 2007: 1-7.

[8]WANG X, WANG Y, WANG L. Improving fuzzy c-means clustering based on feature-weight learning[J]. Pattern recognition letters, 2004, 25(10): 1123-1132.

[9]HE P, XU X, HU K, et al. Semi-supervised clustering via multi-level random walk[J]. Pattern recognition, 2014, 47(2): 820-832.

[10]HOI S C H, LIU W, LYU M R, et al. Learning distance metrics with contextual constraints for image retrieval[C]// IEEE Conference on Computer Vision and Pattern Recognition. New York, USA, 2006: 2072-2078.

[11]曾令伟,伍振兴,杜文才.基于改进自监督学习群体智能(ISLCI)的高性能聚类算法[J].重庆邮电大学学报: 自然科学版, 2016, 28(1): 131-137.
ZENG Lingwei, WU Zhenxing, DU Wencai. Improved self supervised learning collection intelligence based high performance data clustering approach[J]. Journal of Chongqing university of posts and telecommunications: natural science edition, 2016, 28(1): 131-137.

[12]程旸, 王士同. 基于局部保留投影的多可选聚类发掘算法[J].智能系统学报, 2016, 11(5): 600-607.
CHENG Yang, WANG Shitong. A multiple alternative clusterings mining algorithm using locality preserving projections[J]. CAAI transactions on intelligent systems, 2016, 11(5): 600-607.

[13]DUDA R O, HART P E, STORK D G. Pattern classification[M]// Pattern classification. Wiley, 2001: 119-131.

[14]MEI J P, CHEN L. Fuzzy clustering with weighted medoids for relational data[J]. Pattern recognition, 2010, 43(5): 1964-1974.

[15]HOPPNER F, KLAWONN F. Improved fuzzy partitions for fuzzy regression models[J]. International journal of approximate reasoning, 2003, 32(2/3): 85-102.

[16]ZHU L, CHUNG F L, WANG S. Generalized fuzzy C-means clustering algorithm with improved fuzzy partitions[J]. IEEE transactions on systems man and cybernetics part B, 2009, 39(3): 578-591.

[17]STREHL A, GHOSH J. Cluster ensembles-a knowledge reuse framework for combining multiple partitions[J]. Journal of machine learning research, 2002, 3(3): 583-617.

[18]IWAYAMA M, TOKUNAGA T. Hierarchical Bayesian clustering for automatic text classification[J]. IJCAI, 1996: 1322-1327.

[19]RAND W M. Objective criteria for the evaluation of clustering methods[J]. Journal of the american statistical association, 1971, 66(336): 846-850.

作者简介:

卞则康,男,1993 年生,硕士研究生,主要研究方向为人工智能和模式识别。

王士同,男,1964 年生,教授,博士生导师,主要研究方向为人工智能与模式识别。发表学术论文近百篇,其中被 SCI、EI 检索 50 余篇。