Journal of Computer Applications
计算机应用,2012,32(3):646-648
ISSN 1001-9081
CODEN JYIIDU
2012-03-01
http://www.joca.cn

文章编号:1001-9081(2012)03-0646-03
doi:10.3724/SP.J.1087.2012.00646

距离修正的模糊 C 均值聚类算法

楼晓俊[1*],李隽颖[1],刘海涛[1,2]
(1.中国科学院 上海微系统与信息技术研究所,上海 200050; 2.无锡物联网产业研究院,江苏 无锡 214135)
(*通信作者电子邮箱 louxjanan@gmail.com)

摘　要:经典的模糊 C 均值算法基于欧氏距离,存在等划分趋势的缺陷,分错率较高,只适用于球形结构的聚类。针对这一问题,利用数据的点密度信息,在数据点与聚类中心的距离度量中引入了调节因子,提出了一种基于密度的距离修正矩阵,并用其代替经典模糊 C 均值算法中的距离度量矩阵。通过人造数据集和 UCI 数据集的两组聚类实验,证实了改进算法对非球形结构的数据同样适用,且相比经典的模糊 C 均值算法具有更高的聚类准确率。

关键词:聚类;模糊 C 均值;距离度量;点密度;调节因子
中图分类号:TP18;TP391.4;TP301.6　　**文献标志码**:A

Improved fuzzy C-means clustering algorithm based on distance correction

LOU Xiao-jun[1*], LI Jun-ying[1], LIU Hai-tao[1,2]
(1. *Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, Shanghai* 200050, *China*;
2. *Wuxi SensingNet Industrialization Research Institute, Wuxi Jiangsu* 214135, *China*)

Abstract: Based on Euclidean distance, the classic Fuzzy C-Means (FCM) clustering algorithm has the limitation of equal partition trend for data sets. And the clustering accuracy is lower when the distribution of data points is not spherical. To solve these problems, a distance correction factor based on dot density was introduced. Then a distance matrix with this factor was built for measuring the differences between data points. Finally, the new matrix was applied to modify the classic FCM algorithm. Two sets of experiments using artificial data and UCI data were operated, and the results show that the proposed algorithm is suitable for non-spherical data sets and outperforms the classic FCM algorithm in clustering accuracy.

Key words: clustering; Fuzzy C-Means (FCM); distance measurement; dot density; regulatory factor

模糊聚类技术基于模糊集合论,被广泛应用于数据挖掘、模式识别、控制决策等领域,具有重要的理论和实际应用价值。模糊 C 均值(Fuzzy C-Means, FCM)算法是模糊聚类中最基本也是应用最广泛的方法之一,它是一种基于划分的聚类算法,依据最小二乘原理,采用迭代方法优化目标函数,最终得到每个样本点的归属[1-4]。然而经典的 FCM 算法基于欧氏距离,只能适用于球形结构的聚类,最小化目标函数的方法具有对数据集进行等划分的趋势,对于其他结构的聚类分错率较高[5-6]。因此出现了许多 FCM 的改进算法,有学者考虑了样本点不同维度对聚类效果的不同贡献,通过特征加权的方式来优化算法[7-9];有学者引入了不同的距离度量方法,如马氏距离[10-11]、组合距离[12]、权重距离[13];还有学者通过修正隶属度函数和目标函数来优化算法[14-15]。本文基于数据的点密度信息,提出了一种基于距离修正的 FCM(FCM based on Distance Correction, FCM-DC)改进算法,引入了距离度量的调节因子,弥补了欧氏距离等划分趋势的影响,通过人造数据和 UCI 数据的两组聚类实验,证实了该算法相比经典的 FCM 具有更广的适用范围和更高的聚类准确率。

1　FCM 算法

设 $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}$ 为 n 元数据集合,$\boldsymbol{x}_i\in\mathbf{R}^s$。FCM 聚类方法就是按照特定规则把 $\boldsymbol{X}$ 划分为 c 个子集 $\boldsymbol{S}_1,\boldsymbol{S}_2,\cdots,\boldsymbol{S}_c$,若用 $\boldsymbol{V}=\{\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_c\}$ 表示这 c 个子集的聚类中心,u_{ij} 表示元素 $\boldsymbol{x}_j$ 对 $\boldsymbol{S}_i$ 的隶属度,则 FCM 算法的优化目标函数为:

$$J^m(\boldsymbol{U},\boldsymbol{V},\boldsymbol{X})=\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^m d_{ij}^2 \tag{1}$$

其中:

$$d_{ij}^2=\left\|\boldsymbol{x}_j-\boldsymbol{v}_i\right\|_{\boldsymbol{A}}^2=(\boldsymbol{x}_j-\boldsymbol{v}_i)\boldsymbol{A}(\boldsymbol{x}_j-\boldsymbol{v}_i)^{\mathrm{T}} \tag{2}$$

u_{ij} 满足约束条件:

$$\sum_{i=1}^{c}u_{ij}=1;1\leqslant j\leqslant n$$

$$u_{ij}\geqslant 0;1\leqslant i\leqslant c,1\leqslant j\leqslant n$$

式(1)~(2)中:$\boldsymbol{U}=\{u_{ij}\}$ 为 $c\times n$ 阶矩阵;$\boldsymbol{V}=\{\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_c\}$ 为 $s\times c$ 阶矩阵;$\boldsymbol{A}$ 为 $s\times s$ 阶对称正定矩阵;d_{ij} 为数据元 $\boldsymbol{x}_j$ 与聚类中心 $\boldsymbol{v}_i$ 的距离,经典的 FCM 算法中使用欧氏距离,即 $\boldsymbol{A}$ 取单位矩阵 $\boldsymbol{I}$;m 为大于 1 的模糊指数,控制分类矩阵 $\boldsymbol{U}$ 的模糊程度,m 越大,聚类的模糊程度越大,在实际应用中 m 最佳取值范围为(1.5,2.5),一般使用 $m=2$[16-17]。FCM 算法是使目标函数最小化的迭代收敛过程,通过 Lagrange 乘数法求解式(1)可得到:

$$\boldsymbol{v}_i=\frac{\sum_{j=1}^{n}u_{ij}^m\boldsymbol{x}_j}{\sum_{j=1}^{n}u_{ij}^m} \tag{3}$$

收稿日期:2011-08-22;**修回日期**:2011-12-08。
基金项目:国家科技重大专项(2010ZX03006-004);国家 973 计划项目(2011CB302906)。
作者简介:楼晓俊(1984-),男,浙江杭州人,博士研究生,CCF 会员,主要研究方向:传感器网络信号处理、模式识别; 李隽颖(1982-),男,湖北云梦人,博士研究生,CCF 会员,主要研究方向:传感器网络信号处理、模式识别; 刘海涛(1968-),男,新疆昌吉人,研究员,博士生导师,主要研究方向:传感器网络、物联网体系架构。

$$u_{ij} = 1/\sum_{k=1}^{c}\left(\frac{d_{ij}}{d_{kj}}\right)^{2/(m-1)} \tag{4}$$

2　FCM-DC 算法

模式识别中通常利用样本点之间的距离来度量其差异性,然后作为判别其归属的依据。经典的模糊 C 均值算法采用欧氏距离来度量样本间的差异性,虽然运算简单,但是对于非球形结构或者非对称结构的聚类存在缺陷。样本点密度能够反映其合群程度,同时也能一定程度上说明其对聚类的影响能力,因此利用样本点密度信息,构造了距离的调节因子,形成样本与聚类中心的距离矩阵,用于修正经典 FCM。

2.1　基于点密度的距离调节因子

在样本集 $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n\}$ 中,对于每个样本点 $\boldsymbol{x}_i$,通常点密度函数的表达式定义为:

$$z_i = \sum_{\substack{j=1 \\ j\neq i}}^{n}\frac{1}{d_{ij}}; d_{ij} \leqslant \sigma, 1 \leqslant i \leqslant n \tag{5}$$

其中:d_{ij} 表示样本 $\boldsymbol{x}_i$ 与 $\boldsymbol{x}_j$ 之间的欧氏距离;σ 是点密度的有效半径,可根据实际情况设置,σ 越大得到的点密度相对值也越大。为了简化算法,提出了一种新的点密度定义方式:

$$z_i = 1/\min(\{d_{ij}\}); 1 \leqslant i \leqslant n \tag{6}$$

即将样本 $\boldsymbol{x}_i$ 到其最近邻样本之间的距离的倒数作为其点密度。

利用式(6)定义的点密度信息,提出了用于 FCM 距离修正的调节因子,定义为:

$$w_i = \frac{\sum_{j=1}^{n}\alpha_j z_j}{\sum_{j=1}^{n}\alpha_j}; 1 \leqslant i \leqslant c \tag{7}$$

其中 α_j 是样本 $\boldsymbol{x}_j$ 的类别归属标识。当样本 $\boldsymbol{x}_j \in \boldsymbol{S}_i$,即 $\boldsymbol{x}_j$ 归属于子集 $\boldsymbol{S}_i$,则 $\alpha_j = 1$;否则 $\alpha_j = 0$。式(7)调节因子 w_i 反映了样本所在类别中的整体点密度信息。

2.2　FCM-DC 聚类算法

得到调节因子 w_i 之后,样本点与聚类中心的距离度量修正为:

$$d_{ij}^2 = \frac{\left\|\boldsymbol{x}_j - \boldsymbol{v}_i\right\|^2}{w_i}; 1 \leqslant i \leqslant c, 1 \leqslant j \leqslant n \tag{8}$$

因此,将式(8)代入式(1)可以得到FCM-DC算法的目标函数为:

$$J_w^m(\boldsymbol{U},\boldsymbol{V},\boldsymbol{X}) = \sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^m\frac{\left\|\boldsymbol{x}_j - \boldsymbol{v}_i\right\|^2}{w_i} \tag{9}$$

构造 Lagrange 函数,对目标函数求极小值,可得:

$$\boldsymbol{v}_i = \frac{\sum_{j=1}^{n}u_{ij}^m\boldsymbol{x}_j}{\sum_{j=1}^{n}u_{ij}^m}; 1 \leqslant i \leqslant c \tag{10}$$

$$u_{ij} = 1\Big/\sum_{k=1}^{c}\left(\frac{w_k\left\|\boldsymbol{x}_j - \boldsymbol{v}_i\right\|^2}{w_i\left\|\boldsymbol{x}_j - \boldsymbol{v}_k\right\|^2}\right)^{1/(m-1)} \tag{11}$$

由上述的推导过程可看出:FCM-DC 算法相对于 FCM 算法的改进主要在于样本差异性的度量中在欧氏距离的基础上引入了调节因子 w_i,而 w_i 是基于样本点密度信息的。

FCM-DC 算法步骤如下:

1) 根据式(6),计算每个样本的点密度 z_i。

2) 设定聚类类别数目 c,模糊指数 m,迭代截止误差 ε 和最大迭代次数 T,初始化隶属度矩阵 $(u_{ij})_0$。

3) 由隶属度矩阵 $(u_{ij})_t$,根据式(10)计算聚类中心 $(\boldsymbol{v}_i)_t$。

4) 根据式(9)计算目标函数 $(J)_t$。

5) 根据式(11),更新隶属度矩阵 $(u_{ij})_{t+1}$。

6) 判决迭代截止条件,如果 $|(u_{ij})_t - (u_{ij})_{t+1}| \leqslant \varepsilon$ 或者 $t = T$,则迭代结束;否则令 $t = t + 1$,跳回 3) 继续。

3　实验与分析

为了验证算法的有效性,对 FCM-DC 进行人造数据集和 UCI 数据集两组实验,并与经典 FCM 算法进行了比较与分析。实验采用 Matlab 程序仿真,参数都选择默认的常规配置,即 $m = 2, \varepsilon = 10^{-5}, T = 100$。

3.1　人造数据集实验

为了能够直观地分析与对比聚类算法的性能,在二维坐标轴上随机生成两组样本点,分别代表两个类别子集。随机样本点满足以下规则:第一组样本点均匀分布在圆心(0,0),半径为 1 的圆形区域内,第二组样本点均匀分布在圆心(5.5,0),半径为 5 的圆形区域内;两组样本各 100 个点,因此第二组样本的密度小于第一组样本。对产生的样本点分别进行经典 FCM 聚类与 FCM-DC 聚类,实验重复进行 20 次,图 1 是某次实验的聚类结果对比图。

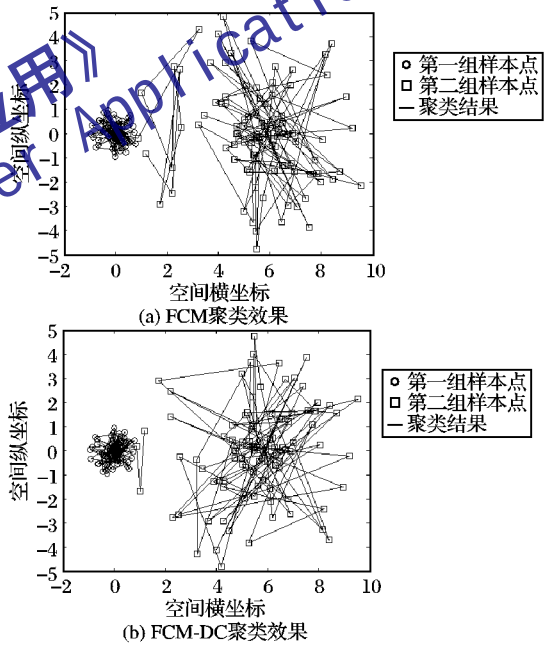

图 1　人造数据集上的单次聚类性能对比

表 1 给出了重复 20 次实验后,两种聚类算法的平均分错率及聚类中心的平均偏差。聚类中心的偏差定义为实际运算得到的聚类中心与真实聚类中心的绝对空间距离。从表 1 可看出:FCM-DC 算法的聚类效果与真实情况更为接近,具有更高的聚类准确率。这是因为经典 FCM 算法具有等划分趋势的限制,样本点被自然划分到离聚类中心近的类别中,因此当样本点的分布密度不同时,FCM 就会出现较高的分错率。而 FCM-DC 算法正是基于点密度信息,引入了距离修正的调节因子,因此在此类情况下,能够大大提升聚类的准确率。

表 1　人造数据集 20 次实验平均聚类性能对比

聚类算法	聚类平均分错率/%	聚类中心平均偏差
FCM 聚类	10.5	0.97
FCM-DC 聚类	1.8	0.26

3.2 UCI 数据集实验

为了进一步验证 FCM-DC 算法的有效性,采用 UCI 数据集中的 IRIS 和 WINE 数据进行实验,因为这两组数据是国际公认的比较无监督聚类方法效果好坏的典型数据。其中 IRIS 数据包含 150 个 4 维的样本点,类别数为 3,每类 50 个样本点。第一类数据与其他两类数据离得较远,第二类数据与第三类数据离得较近,且部分重叠;WINE 数据包含 178 个 13 维数据,类别数也为 3,三类样本数目各为 59,71 和 48。

表 2 ~ 3 是 FCM 算法与 FCM-DC 算法对 UCI 数据集的聚类效果对比。在 IRIS 数据实验中,对于相隔较远的第一类数据,两种算法都能够完全做到没有错误,但对于有交叉重合的第二类和第三类数据,FCM-DC 算法表现出了更高的准确率,总体准确率比 FCM 提升了 12.6%;在 WINE 数据实验中,由于数据本身的高维稀疏特性,两种算法的聚类分错率都比较高,但 FCM-DC 算法仍然比 FCM 提升了 9.1% 的聚类准确率。实验说明,调节因子 w_i 起到了距离度量的修正优化效果,提升了聚类的准确率。同时,由于这两组数据集是典型的非球形结构数据,因此实验也说明 FCM-DC 算法适用面更广。

表 2 IRIS 数据的聚类效果对比

聚类算法	数据类别	分错个数	聚类分错率/%
FCM 聚类	第一类	0	10.67
	第二类	13	
	第三类	3	
FCM-DC 聚类	第一类	0	9.33
	第二类	9	
	第三类	5	

表 3 WINE 数据的聚类效果对比

聚类算法	数据类别	分错个数	聚类分错率/%
FCM 聚类	第一类	40	49.44
	第二类	25	
	第三类	23	
FCM-DC 聚类	第一类	33	44.94
	第二类	22	
	第三类	25	

4 结语

针对经典模糊 C 均值算法存在的等划分趋势的缺陷,本文提出了一种距离修正的改进算法——FCM-DC。基于样本点密度信息,引入了距离修正的调节因子,对 FCM 算法样本差异性度量进行了修正。通过人造数据集与 UCI 数据集两组实验,对比分析 FCM-DC 与 FCM 算法的性能,结果表明,FCM-DC 算法对于非球形结构的数据同样适用,且具有更高的聚类准确率。

参考文献:

[1] PEDRYCZ W. Conditional fuzzy C-means [J]. Pattern Recognition Letters, 1996, 17(6): 625 - 631.

[2] GRAVES D, PEDRYCZ W. Kernel-based fuzzy clustering and fuzzy clustering: A comparative experimental study [J]. Fuzzy Sets and Systems, 2010, 161(4): 522 - 543.

[3] SONG Q, YANG X L, SOH Y C, *et al.* An information-theoretic fuzzy C-spherical shells clustering algorithm[J]. Fuzzy Sets and Systems, 2010, 161(13): 1755 - 1773.

[4] LEE M, PEDRYCZ W. The fuzzy C-means algorithm with fuzzy P-mode prototypes for clustering objects having mixed features[J]. Fuzzy Sets and Systems, 2009, 160(24): 3590 - 3600.

[5] 刘小芳,曾黄麟,吕炳朝. 点密度函数加权模糊 C-均值算法的聚类分析[J]. 计算机工程与应用,2004,40(24):64 - 65.

[6] TANG C L, WANG S G, XU W. New fuzzy C-means clustering model based on the data weighted approach [J]. Data & Knowledge Engineering, 2010, 69(9): 881 - 900.

[7] 王惠,申石磊. 一种改进的特征加权 *K*-means 聚类算法[J]. 微电子学与计算机,2010,27(7):161 - 163.

[8] 李丹,顾宏,张立勇. 基于属性权重区间监督的模糊 C 均值聚类算法[J]. 控制与决策,2010,25(3):457 - 460.

[9] BAI L, LIANG J Y, DANG C Y, *et al.* A novel attribute weighting algorithm for clustering high-dimensional categorical data[J]. Pattern Recognition, 2011, 44(12): 2843 - 2861.

[10] 蔡静颖,周福娟,张永. 基于自适应马氏距离的模糊 C 均值算法[J]. 计算机工程与应用,2010,46(34):174 - 176.

[11] XIANG S M, NIE F P, ZHANG C S. Learning a Mahalanobis distance metric for data clustering and classification[J]. Pattern Recognition, 2008, 41(12): 3600 - 3612.

[12] 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.

[13] TSAI D M, LIN C C. Fuzzy C-means based clustering for linearly and nonlinearly separable data[J]. Pattern Recognition, 2011, 44(8): 1750 - 1760.

[14] 于迪,李义杰. 基于减法聚类改进的模糊 C 均值算法的模糊聚类研究[J]. 微型机与应用,2010,29(16):14 - 16.

[15] 李雷,罗红旗,丁亚丽. 自适应约束模糊 C 均值聚类算法[J]. 模糊系统与数学,2010,24(5):126 - 130.

[16] YU J, CHENG Q S, HUANG H K. Analysis of the weighting exponent in the FCM[J]. IEEE Transactions on System, Man and Cybernetics: Part B: Cybernetics, 2004, 34(1): 634 - 639.

[17] 肖满生,阳娣兰,张居武,等. 基于模糊相关度的模糊 C 均值聚类加权指数研究[J]. 计算机应用,2010,30(12):3388 - 3390.

(上接第 645 页)

[8] BELKIN M, NIYOGI P, SINDHWANI V. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples [J]. Journal of Machine Learning Research, 2006, 7(11): 2399 - 2434.

[9] ZHOU D Y, BOUSQUET O, LAL T N, *et al.* Learning with local and global consistency [C]// THRUM S, SAUL L, SCHÖLKOPF B. Proceedings of the 18th Annual Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2003: 321 - 328.

[10] WU M R, SCHOLKOPF B. Transductive classification via local learning regularization [C]// MEILA M, SHEN X. Proceedings of the 11th International Conference on Artificial Intelligence and Statistics. Cambridge: MIT Press, 2007: 624 - 631.

[11] XIANG S M, NIE F P, ZHANG C S. Semi-supervised classification via local spline regression [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 32(11): 2039 - 2053.

[12] WANG F. A general learning framework using local and global regularization [J]. Pattern Recognition, 2010, 43(9): 3120 - 3129.

[13] VAPNIK V. The nature of statistical learning theory [M]. Berlin: Springer-Verlag, 1995.

[14] BOTTOU L, VAPNIK V. Local learning algorithms [J]. Neural Computation, 1992, 4(6): 888 - 900.

[15] VAPNIK V, BOTTOU L. Local algorithms for pattern recognition and dependencies estimation [J]. Neural Computation, 1993, 5(6): 893 - 909.