广义洛伦兹内核函数在模糊C均值聚类中的应用研究

王建华[1] 李晓峰[2] 高巍巍[2]
(哈尔滨师范大学 哈尔滨 150025)[1] (黑龙江外国语学院信息科学系 哈尔滨 150025)[2]

摘 要 模糊C均值(FCM)算法是数据聚类分析的主要算法。但在嘈杂环境下,对于抽样大小不一的聚类,数目越多准确性越低,上述弊端可通过替代性FCM(AFCM)的高斯内核映射来解决。鉴于AFCM的不足,提出了针对模糊C均值聚类的广义洛伦兹内核函数。利用该算法对鸢尾数据库进行聚类,将其划分成山鸢尾、变色鸢尾和维吉尼亚鸢尾3类。实验结果表明,广义洛伦兹模糊C均值(GLFCM)可实现对离群聚类和大小不等的聚类数据的分类,其结果优于K均值、FCM、替代性C均值(AFCM)、Gustafson-Kessel(GK)和Gath-Geva(GG)方法,收敛迭代次数比AFCM的更少,其分区索引(SC)效果也好于其他方法。

关键词 广义洛伦兹隶属函数,K均值,替代性模糊C均值,聚类,离群聚类
中图法分类号 TP301.6 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.9.052

Research on Generalized Lorenz Kernel Function in Fuzzy C Means Clustering

WANG Jian-hua[1] LI Xiao-feng[2] GAO Wei-wei[3]
(Harbin Normal University, Harbin 150025, China)[1]
(Information Science Department, Heilongjiang International University, Harbin 150025, China)[2]

Abstract Fuzzy C means(FCM) algorithm is the main algorithm for data clustering analysis. But in a noisy environment, for the clusters of different sampling sizes, accuracy is low when the number of clusters is large. The above disadvantages can be sloved through the Gauss kernel mapping of alternative FCM(AFCM). This paper proposed generalized Lorenz kernel function to the fuzzy C means clustering for the deficiency of AFCM. This algorithm was used to analyze the Iris database cluster, to classify the Iris database into three clusters of Iris setosa, Iris versicolour and Iris virginica. Experimental results show that the generalized lorentzian fuzzy C-means(GLFCM) can classify data of outliers and unequal sized clusters. The GLFCM yields better cluster than K-means(KM), FCM, alternative fuzzy C-means(AFCM), Gustafson-Kessel(GK) and Gath-Geva(GG). It takes less iteration than that of AFCM to converge. Its partition index (SC) is better than the others.

Keywords Generalized lorentzian membership function, K-means, Alternative fuzzy C-means, Clustering, Outlier clustering

1 引言

聚类分析[1,2]是一种利用无监督学习规则将一组固定数据集划分成几组类似个体的科学方法,是机器学习、模式识别和人工智能领域的一项重要任务。K均值(KM)方法是众所周知且被广泛用于分区聚类的算法,易于执行,效率高,线性时间复杂度好。KM将数据集划分为聚类,其中每个对象只能属于一个聚类。离散隶属函数会导致KM算法出错。聚类结果对初始聚类中心敏感,会收敛成局部最佳。

标准的模糊C均值(FCM)[3,4]聚类算法允许每个对象隶属于所有聚类,每个对象均有一个隶属程度来代表该对象与聚类中心的距离[5]。FCM算法利用欧氏距离函数来度量数据点之间的相似度。当情况嘈杂或每个聚类里的抽样大小相差较大时,FCM算法会面临诸多问题。一些学者对欧氏距离函数进行了改进,增加了单调递增功能,称之为替代性函数或高斯内核方法,它可以解决被离群聚类破坏的数据方面的问题[6]。

Gustafson-Kessel(GK)算法[7,8]用具体聚类的马氏距离来取代欧氏距离,以适应不同大小、不同形态的聚类。Gath-Geva(GG)算法[9,10]与GK差不多,只是距离计算法则只用到指数项来发现形状不同的聚类,所以会比内积法则递减得更快。GK和GG算法无法对离群聚类的数据进行分类。

本文提出的新方法利用广义洛伦兹函数来测得数据与中心之间的距离。本文算法实际上解决了KM、FCM、AFCM、GK和GG算法面临的问题。为验证其性能,将其在3组不同的人工生成数据集和鸢尾数据集上进行了实验。

2 相关研究

设 $X=\{x_1,x_2,\cdots,x_N\}$ 是 N 个元素的一维实数。设 $C=$

到稿日期:2014-09-06 返修日期:2014-11-29 本文受黑龙江省智能教育与信息工程重点实验室开放基金项目(1155xnc107),黑龙江省教育厅科学技术研究项目(12543067)资助。

王建华(1976-),女,硕士,教授,CCF高级会员,主要研究方向为智能教育、人工智能;李晓峰(1978-),男,博士生,副教授,CCF高级会员,主要研究方向为数据挖掘、文本挖掘、智能算法,E-mail:mberse@126.com;高巍巍(1976-),女,硕士,副教授,主要研究方向为数据库。

$\{c_1, c_2, \cdots, c_K\}$是由 N 个中心组成的一组数据，其中 C 是某个聚类算法的解。设 $U=[u_{ij}](i=1,2,\cdots,N, j=1,2,\cdots,K)$ 是 X 对 C 的隶属矩阵。

2.1 K 均值算法

K 均值算法将数据划分成 K 个数据集，得到的解就是由 K 个中心组成的一组数据，每个中心是一个局部分区。每个分区由一组被分配给最临近中心的数据构成。对于隶属函数，每个数据点属于离其最近的中心，且与其他中心毫无关联。经 KM 算法优化后的目标函数为

$$KM(X,C)=\sum_{i=1}^{N}\min \| x_i-c_j \|^2 \tag{1}$$

该目标函数提供一种使每个中心与对应数据点之间的平均距离达到最小的算法，即群内方差。KM 的隶属程度和新中心分别是：

$$u_{ij}=\begin{cases}1, & \text{if } j=\arg\min \| x_i-c_j \|^2\\ 0, & \text{otherwise}\end{cases} \tag{2}$$

$$c_j=\frac{\sum_{i=1}^{N}u_{ij}^m x_i}{\sum_{i=1}^{N}u_{ij}^m} \tag{3}$$

其中，m 是一个模糊化参数。

KM 算法有严格的隶属函数，对所有数据点同等重视。该算法因便于理解和执行，成为了聚类分析的惯用算法。

2.2 模糊 C 均值算法(FCM)

模糊 C 均值算法改进后成功地解决了各种聚类问题。FCM 算法将一组有限的 N 个元素划分到模糊聚类的一个集合里。基本的模糊 C 均值目标函数定义如下：

$$J_m(U,C)=\sum_{j=1}^{K}\sum_{i=1}^{N}u_{ij}^m \| x_i-c_j \|^2 \tag{4}$$

其中，$\| \cdot \|^2$ 指欧氏距离度量的平方；模糊化参数 $m\in[1,\infty]$，u_{ij} 是聚类 j 中 x_i 的隶属程度；x_i 是 d 维测量数据；c_j 是聚类的第 j 个 d 维中心。

可根据如下步骤对隶属程度和新中心进行反复更新：

$$u_{ij}=\frac{1}{\sum_{k=1}^{K}\frac{\| x_i-c_j \|^2}{\| x_i-c_k \|^2}} \tag{5}$$

$$c_j=\frac{\sum_{i=1}^{N}u_{ij}^m x_i}{\sum_{i=1}^{N}u_{ij}^m} \tag{6}$$

如果有 $\max_{ij}\{|u_{ij}^{(t+1)}-u_{ij}^{(t)}|\}<\varepsilon$，则停止迭代操作，其中 ε 是较小正值，t 是迭代次数。

2.3 替代性模糊 C 均值(AFCM)

替代性模糊 C 均值(AFCM)，也称为高斯内核模糊 C 均值(GKFCM)，由模糊 C 均值衍生而来，利用替代性高斯内核距离函数来取代欧氏距离。AFCM 的目标函数定义为

$$J_{AFCM}=\sum_{j=1}^{K}\sum_{i=1}^{N}(u_{ij})^m\{1-\exp(-\beta \| x_i-c_j \|^2)\} \tag{7}$$

其中，$m>1$，约束条件 $\sum_{j=1}^{K}u_{ij}=1, i=1,\cdots,N$。根据下列步骤可对应地对新中心和隶属程度进行更新：

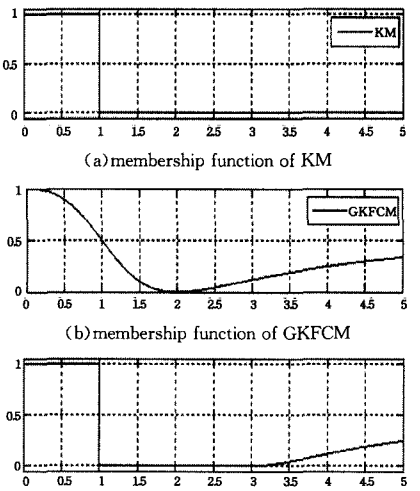
$$u_{ij}=\frac{1/(1-\exp(-\beta \| x_i-c_j \|^2))^{1/(m-1)}}{\sum_{j=1}^{K}1/(1-\exp(-\beta \| x_i-c_j \|^2))^{1/(m-1)}} \tag{8}$$

$$c_j=\frac{\sum_{i=1}^{N}(u_{ij})^m\exp(-\beta \| x_i-c_j \|^2)x_i}{\sum_{i=1}^{N}(u_{ij})^m\exp(-\beta \| x_i-c_j \|^2)} \tag{9}$$

替代性距离函数的主要贡献在于其能解决离群聚类方面的问题，如式(9)里离群聚类的影响将会削弱。

3 广义洛伦兹模糊 C 均值算法(GLFCM)

想要找出聚类的中心，理想的做法就是给最近中心的数据分配一个隶属值，同时给隶属值为 0 或类似值的其他中心的数据分配一个隶属值。图 1 给出的案例中有两个中心：0 和 2。图 1(a)提供了中心 0 的数据的 KM 隶属函数。如果对中心 0 进行更新，只有隶属值为 1 的数据会保留下来。图 1(b)提供了中心 0 的数据的 AFCM 和 GKFCM 隶属函数。2 附近隶属值为 0 或非常接近 0 的数据不多。这些为数不多的数据目前属于中心 2，一旦对中心 0 进行更新后，这些数据将移除。图 1(c)给出了中心 0 的数据的理想隶属函数。2 附近隶属值为 0 或非常接近 0 的数据有很多，这个数据群应保留在中心 2 里。但一旦对中心 0 进行更新，这些数据必须删除。离 0 和 2 较远的数据也有一定的隶属值，但不会大于 0.5。基于该理念，可以得出隶属函数的一个理想应用就是广义洛伦兹函数，如图 1(c)所示。



(a) membership function of KM

(b) membership function of GKFCM

(c) membership function of ideal method

图 1 中心 0 的数据的隶属函数曲线

为达到研究目的，本文提出一种新的距离转换函数 $GLF(d)$，其中 d 是未知欧氏距离的平方。$GLF(\cdot)$ 表述如式(10)。进行函数走向调控时要求有两个可调参数：a 和 b。如果 $a=b$，且 d 是欧氏距离的平方，那么 $GLF(\cdot)$ 模仿洛伦兹内核函数(LF)，得到式(11)。将该函数与高斯函数(GF)进行比较。如图 2 所示，在 0 点附近，GLF 表现为快速递减，并趋于无穷。参数 a 控制着 0 点附近的递减速度和宽度。然而 GLF 的支持范围大于 GF。

$$GLF(d)=\frac{1}{a*d^2+1}-\frac{b*d^2}{(a*d^2+1)^2} \tag{10}$$

$$LF(d)=\frac{1}{(a*d^2+1)^2} \tag{11}$$

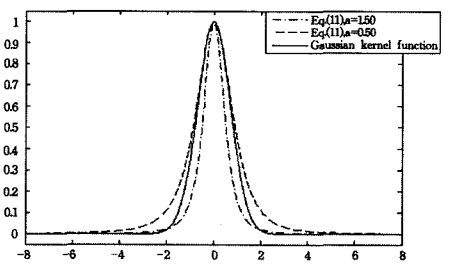


图 2 特征空间与输入空间距离的相关性

图 3 给出了 KM(见式(2))和 GLFCM(见式(13))在中心为 0 和 2 的两个聚类里的隶属曲线。由中心 0 附近的数据的 KM 隶属曲线可见,在[0,1]范围内,隶属度都是 1s;对于 1 以外的数据,其隶属度均是 0s。图 3 没有提供 FCM(见式(5))的隶属曲线,因为该算法包含了除 2 以外的所有数据。AFCM 隶属函数的隶属曲线提供了除 2 以外的所有数据点以及附近其他数据点的非 0 隶属程度。对于 GLFCM 隶属函数的隶属曲线提供了只属于该中心的附近所有数据点的非 0 隶属程度。对于位于[1.5,2.8]之间的所有数据,其隶属值为 0,对旧中心迁移到新中心不存在影响。本文提出的广义洛伦兹模糊 C 均值演变过程如下。

经 GLF 算法调整后,得到 GLFCM 目标函数:

$$J_{GLFCM}=\sum_{j=1}^{K}\sum_{i=1}^{N}(u_{ij})^m\{1-GLF(\beta\|x_i-c_j\|^2)\} \tag{12}$$

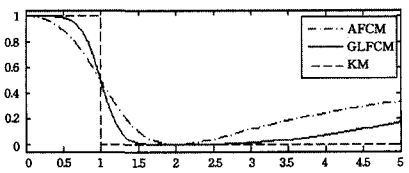
其中,$m>1$,$\sum_{j=1}^{K}u_{ij}=1$,$i=1,\cdots,N$。

利用与 AFCM 的衍生相同的方法对隶属程度和新中心反复进行更新,过程与 AFCM 一样,具体如下:

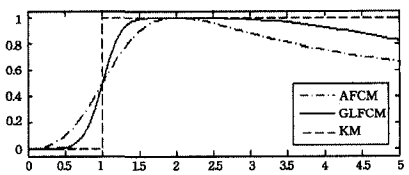
$$u_{ij}=\frac{1/(1-GLF(\beta\|x_i-c_j\|^2))^{1/(m-1)}}{\sum_{j=1}^{K}1/(1-GLF(\beta\|x_i-c_j\|^2))^{1/(m-1)}} \tag{13}$$

$$c_j=\frac{\sum_{i=1}^{N}(u_{ij})^mGLF(\beta\|x_i-c_j\|^2)x_i}{\sum_{i=1}^{N}(u_{ij})^mGLF(\beta\|x_i-c_j\|^2)} \tag{14}$$

由图 3 可知,GLFCM 的隶属曲线比 AFCM 的更接近图 1(c)中的理想曲线。



(a)the reference is centered at 0



(b)the reference is centered at 2

图 3 β 在[0.05,2]范围内,中心 0 和 2 附近数据的 KM、AFCM 和 GLFCM 的隶属曲线

4 实验分析与结果

现利用 4 组数据集来验证所提方法,分别命名为 DataSet1、DataSet2、DataSet3、和 Iris。前 3 组都是人工生成的,最后一组数据来自 UCI Machine Learning Repository。实验得到的 AFCM 和 GLFCM 算法的 β 值如表 1 所列。

表 1 m 等于 2 时 AFCM 和 GLFCM 的 β 值

参数	DataSet1	DataSet2	DataSet3	Iris
AFCM	0.001	0.5969	0.9	0.8
GLFCM	0.0013	0.6	0.72	0.4

4 组数据集及对应的特征如下:

(1)DataSet1($N=84,d=2,K=2$)是一组人工数据集,它是一种双功能的问题,有两个独特的类。样式总数是 84。

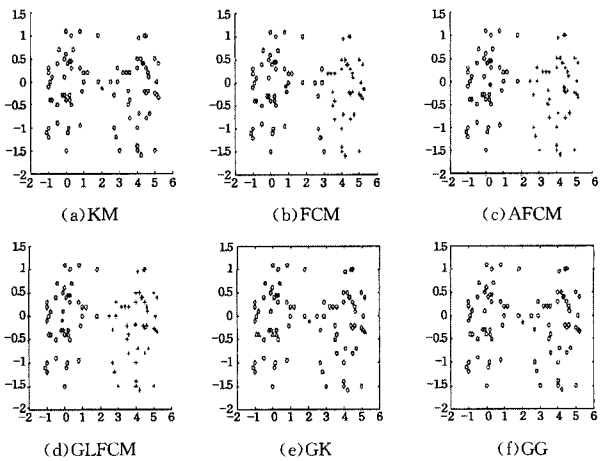
(2)DataSet2($N=39,d=2,K=2$)是一组人工数据集,它是一种双功能的问题,有两个独特的类。样式总数是 39。

(3)DataSet3($N=291,d=2,K=2$)是一组人工数据集,它是一种双功能的问题,有两个独特的类。样式总数是 291。

(4)Iris($N=150,d=4,K=3$)由山鸢尾、变色鸢尾和维吉尼亚鸢尾 3 个鸢尾花品种组成。每个品种按萼片长、萼片宽、花瓣长和花瓣宽这 4 个特征各收集 50 个样本。

本文平均进行了 10 轮模拟实验,每一轮的中心都随机产生并固定不变,只是对算法进行调整。实验用到的电脑配置为英特尔双核酷睿 2.53GHz、4GB RAM、MATLAB 操作环境。

DataSet1 在坐标(100,0)点处存在一个离群聚类。KM 和 FCM 的中心和分区结果分别见图 4(a)和图 4(b)。实心黑点是聚类中心。KM、GK 和 GG 聚类的中心都受到离群聚类的影响。图 4(a)中的中心(2.0339,−0.1387)和(50,50)均不可见。FCM 结果里有许多错误的分类的数据。AFCM 和 GLFCM 里的两个聚类划分正确,未受到离群聚类的影响,如图 4(c)和图 4(d)所示。

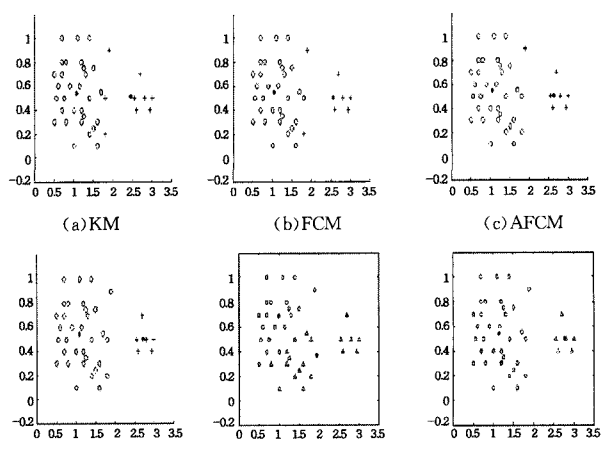


(a)KM (b)FCM (c)AFCM

(d)GLFCM (e)GK (f)GG

图 4 在带离群的坐标点(100,0)的 DataSet1 中,KM、FCM、AFCM、GLFCM、GK 和 GG 的聚类结果

Dataset2 里的抽样聚类数目有两个截然不同的地方。KM 结果里有 3 个错误的分类的数据。GK 里有 12 个错误的分类的数据。FCM 里有两个错误的分类的数据。AFCM 里只有一个错误的分类的数据。只有 GLFCM 和 GG 对这两个聚类的划分正确。有关结果分别见图 5(a)至图 5(f)。



(a)KM (b)FCM (c)AFCM

(d)GLFCM (e)GK (f)GG

图 5 在 DataSet2 中,KM、FCM、AFCM、GLFCM、GK 和 GG 的聚类结果

Dataset3 在右边聚类产生的聚类点数目要比左边的多。KM、GK 和 FCM 里有许多错误分类的数据，而 AFCM 里只有一个。GLFCM 和 GG 对这两个聚类的划分正确，也得到比 AFCM 更多的合理中心点。有关结果分别见图 6(a)至图 6(f)。

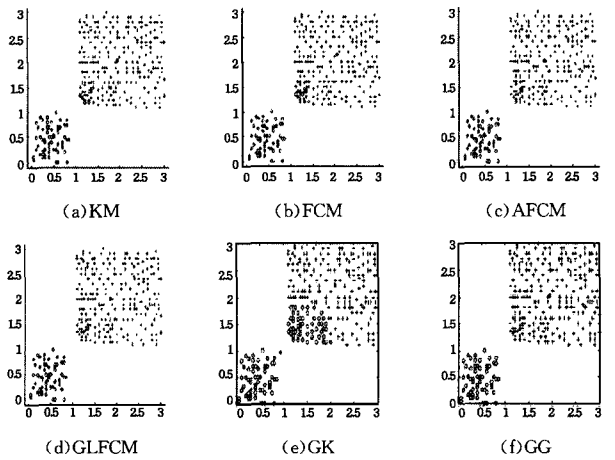


图 6　在 DataSet3 中，KM、FCM、AFCM、GLFCM、GK 和 GG 的聚类结果

对于鸢尾数据集，KM 和 FCM 的总出错数均是 16 秒，AFCM 是 13 秒，GK 是 15 秒，GG 是 40 秒，GLFCM 是 12 秒。由此得出 GLFCM 对鸢尾数据产生的聚类效果最好。以上所有方法的错误计数如表 2 所列；对 4 组数据集进行迭代操作的结果如表 3 所列。

表 2　DataSet1、DataSet2、DataSet3 和 Iris 上的错误计数

数据集	错误计数					
	KM	FCM	AFCM	GLFCM	GK	GG
DataSet1	4.2	7	0	0	42	42
DataSet2	3	2	1	0	12	0
DataSet3	5	19	1	0	58	0
Iris	16	16	13	12	15	40

表 3　各种方法在 4 组数据集上进行迭代操作的结果

数据集	迭代数目					
	KM	FCM	AFCM	GLFCM	GK	GG
DataSet1	5.2	27.3	9.1	7.6	23.5	11.9
DataSet2	4	43.4	46.2	21	30.7	22.1
DataSet3	6.2	48.9	31	29.5	9.8	20.7
Iris	3.3	21.1	33	4	18.5	18.7

本文再通过分区索引(SC)对上述方法做进一步验证。SC 是聚类紧凑性与分离程度之和所占比，是经每个聚类的模糊基数划分后再经规范处理得到单个聚类有效性度量值的总和。只有当对聚类数相同的不同分区进行比较时，SC 才有一定意义。SC 值越低表明划分效果越好。有关结果(见表 4)说明 GLFCM 生成的分区索引优于其他方法。

表 4　4 组数据集里的聚类有效值(即分区索引)对比情况

数据集	聚类算法					
	KM	FCM	AFCM	GLFCM	GK	GG
DataSet1	1.03	0.268	0.105	0.0018	0.0098	1.13
DataSet2	0.226	0.446	0.168	0.0399	0.8638	3.642
DataSet3	0.191	0.405	0.2335	0.1235	0.5109	0.443
Iris	0.084	0.116	0.0917	0.0509	0.0982	0.106

由表 4 可知，GLFCM 的迭代次数要少于 AFCM，且运行至终止条件所用的时间也更短。

结束语　针对聚类大小不一的问题，文章提出用广义洛伦兹内核函数来取代高斯内核 AFCM 聚类函数。为验证其有效性，本文抽取 4 组数据进行实验，结果表明 KM、FCM、GK 和 GG 算法均无法解决问题，且结果还受到离群聚类的影响，而 GLFCM 能对带离群聚类的数据以及大小不一的聚类进行分类。GLFCM 的出错最少，迭代次数也较 AFCM 少。GLFCM 的分区索引低于 KM、FCM、AFCM、GK 和 GG 这 5 种方法。

参 考 文 献

[1] Kaufman L, Rousseeuw P. Finding Groups in Data[M]. Wiley Series in Probability and Statistic, 2005: 56-67

[2] Mirkin B. Clustering for Data Mining: A Data Recovery Approach[M]. Chapman and Hall, 2005: 12-24

[3] Wang Xiang, Guo Rui, et al. A Novel Alternative Weighted Fuzzy C-means Algorithm and Cluster Validity Analysis [C]// IEEE Pacific-Asia Workshop on Computational Intelligence and Industrial Application. 2008: 130-134

[4] Hammerly G, Elkan C. Alternatives to the k-mean algorithm that find better clusterings[C]// Proceedings of the 11th International Conference on Information and Knowledge Management, 2002: 600-607

[5] 郭小芳，李锋，宋晓宁，等. 基于连续域混合蚁群优化的核模糊 C-均值聚类算法研究[J]. 模式识别与人工智能，2014，27(9)：841-846
Guo Xiao-fang, Li Feng, Song Xiao-ning, et al. Kernelized Fuzzy C-means Clusterling Algorithm Based on Hybrid Ant Colony Optimization for Continuons Domains[J]. Pattern Recognition and Artificial Intelligence, 2014, 27(9): 841-846

[6] 李广原，杨炳儒，刘英华，等. 基于模糊论的数据挖掘研究综述[J]. 计算机工程与设计，2011，32(12)：4064-4067
Li Guang-yuan, Yang Bing-ru, Liu Ying-hua, et al. Survey of data mining based on fuzzy set theory[J]. Computer Engineering and Design, 2011, 32(12): 4064-4067

[7] 李丽丽，李明，刘希玉. 基于粒子群模糊 C-均值聚类的图像分割算法[J]. 计算机工程与应用，2009，45(31)：158-160
Li Li-li, Li Ming, Liu Xi-yu. Image segmentation algorithm based on particle swarm optimization fuzzy C-means clustering[J]. Computer Engineering and Applications, 2009, 45(31): 158-160

[8] Liu X, Yang C. Performance research of Gaussian function weighted fuzzy C-means algorithm[C]// Proceedings of SPIE. 2007

[9] Yang M S, Tsai H S. A Gaussian kernel-based fuzzy c-means algotihm with a spatial bias correction[J]. Pattern Recognition Letters, 2008, 29(12): 1713-1725

[10] Ramathilagam S, Huang Yueh-min. Extended Gaussian kernel version of fuzzy c-means in the problem of data analyzing[J]. Expert Systems with Applications: An International Journal, 2011, 38(4): 3793-3805