基于自适应权重的模糊 C-均值聚类算法*

任丽娜，秦永彬，许道云
（贵州大学 计算机科学与信息学院，贵阳 550025）

摘　要：针对模糊 C-均值聚类算法过度依赖初始聚类中心的选取，从而易受孤立点和样本分布不均衡的影响而陷入局部最优状态的不足，提出一种基于自适应权重的模糊 C-均值聚类算法。该算法采用高斯距离比例表示权重，在每一次迭代过程中，根据当前数据的聚类划分情况，动态计算每个样本对于类的权重，降低了算法对初始聚类中心的依赖，减弱了孤立点和样本分布不均衡的影响。实验结果表明，该算法是一种较优的聚类算法，具有更好的健壮性和聚类效果。

关键词：模糊 C-均值聚类算法；自适应权重；高斯距离；隶属矩阵
中图分类号：TP301.6　　**文献标志码**：A　　**文章编号**：1001-3695(2012)08-2849-03
doi:10.3969/j.issn.1001-3695.2012.08.012

Fuzzy C-means clustering based on self-adaptive weight

REN Li-na, QIN Yong-bin, XU Dao-yun
(*College of Computer Science & Information, Guizhou University, Guiyang* 550025, *China*)

Abstract: Due to fuzzy C-means clustering algorithm rely heavily on randomly select C clustering centers, so outlier and uneven distribution of the samples easily influenced and made it easy to fall into the local optimum states. Therefore, this paper proposed an improved fuzzy C-means clustering algorithm based on self-adaptive weights. The new method expressed weight by using the Gaussian distance ratio, it computed the weights for every data according to the current clustering state and no more did rely on the initial clustering center, weakened the influence of outlier and uneven distribution of the samples. The experiments indicate that the fuzzy C-means clustering algorithm based on self-adaptive weights is an effective fuzzy clustering algorithm, has more robust and higher clustering accuracy.

Key words: fuzzy C-means clustering algorithm; self-adaptive weights; Gaussian distance; membership matrix

0　引言

传统的聚类算法中每一个样本必须只属于一个类，对数据形成精确的划分，即硬聚类[1]。但是现实中要处理的很多数据具有不确定性和不精确性，因此传统算法在一些应用上无法达到理想的效果。模糊聚类作为一种软聚类算法，利用隶属度使得类与类之间没有明显的界限，用来处理不分明的对象是行之有效的，使聚类效果更合乎自然，更符合客观实际[2]。

目前，在众多模糊聚类算法中，模糊 C-均值聚类（fuzzy C-means clustering，FCM）算法[3]应用最广泛且较成功，它通过优化目标函数得到每个样本点对所有类中心的隶属度，从而决定样本点的类属以达到自动对数据样本进行聚类的目的。然而该算法仍存在着一些不足：a）算法的性能依赖于初始聚类中心的选取；b）聚类的类数不能自动确定；c）对于样本中的孤立点、噪声数据比较敏感；d）目标函数没有充分考虑到样本分布不均衡的问题[4]。

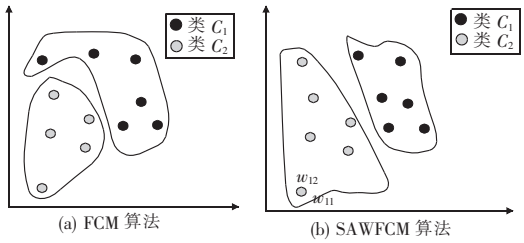
国内外相关文献针对上述的不足进行了深入的研究。齐淼等人[5]通过改进隶属度函数，以消除孤立点对聚类结果的影响，为每个样本点赋予一定量的权值，以改善噪声和分布不均衡样本集的聚类结果；朱林等人[6]提出了一般化的改进模糊划分的 GIFP-FCM 算法，通过引入新的隶属度约束和模糊程度系数，使得算法具有更好的鲁棒性和参数适应性；考虑到使算法不受聚类中心的影响并自动确定聚类个数，魏娜等人[7]提出了一种无监督多尺度聚类算法；Graves 等人[8]采用基于内核的 FCM 算法明显改善了分析数据集的边际问题。

本文提出一种基于自适应权重的模糊 C-均值聚类（fuzzy C-means clustering based on self-adaptive weights，SAWFCM）算法，根据数据样本的具体分布情况设计自适应权重计算策略以改进聚类效果。SAWFCM 算法认为在聚类过程中，同一个类中的样本对于聚类中心的影响并不相同，对每一个样本来说不包含样本的类也有一个产生抑制效果的权重，因此将针对每个样本设计权重并用其改进 FCM 的隶属矩阵，自适应权重的模糊 C-均值聚类在每一次的迭代过程中根据当前的数据划分状态，动态地计算样本权重，降低了 FCM 算法对初始聚类中心和隶属矩阵的依赖，减弱了孤立点和样本分布不均衡的影响。此外，算法采用高斯距离比例表示权重，从而可以有效地减弱数据密度变化对聚类效果的影响[9,10]。如图 1 所示，基本的 FCM 算法在聚类过程中随机选取初始聚类中心，图 1（a）显示了基本 FCM 算法在聚类过程中由于依赖初始聚类中心，在存在孤

收稿日期：2011-12-20；**修回日期**：2012-02-23　　**基金项目**：国家自然科学基金资助项目（60863005）；贵州省科学技术基金资助项目（黔科台 J 字[2012]2125 号）；贵州大学引进人才科研资助项目（贵大人基合字[2011]14 号）

作者简介：任丽娜（1987-），女，辽宁阜新人，硕士研究生，主要研究方向为数据库技术与应用系统（renlina111@163.com）；秦永彬（1980-），男，山东招远人，副教授，博士，主要研究方向为可计算性与计算复杂性；许道云（1959-），男，贵州安顺人，教授，博导，主要研究向为可计算性与计算复杂性、算法设计与分析.

立点时产生了局部最优的聚类结果；而 SAWFCM 算法中同一个类中的每个样本点对类的影响都是不同的，对不包含该样本点的类由于距离远而产生一个抑制权重，图 1(b)则展示了 SAWFCM 算法每个样本点，例如样本点 1 对类 C_1 与 C_2 具有不同的权重 w_{11}、w_{12}，算法在每一次迭代中，依据当前数据划分状态动态计算每个样本所属类别的权重，得到较好的聚类结果。



(a) FCM 算法 (b) SAWFCM 算法

图 1 FCM 与 SAWFCM 算法的聚类

1 基本 FCM 算法

FCM 算法将数据集 $X=\{x_i, i=1,\cdots,n\}$ 分为 c 个模糊类，并求每个类的聚类中心，使得目标函数即非相似性指标的价值函数达到最小。该算法采用模糊度划分，使得每个给定数据样本点用值在[0,1]间的隶属度来确定其属于各个组的隶属度。根据归一化规定，一个数据集的隶属度的和总等于 1：

$$\sum_{k=1}^{c} u_{ik}=1, \forall i=1,\cdots,n \tag{1}$$

那么，FCM 的目标函数为

$$J(U,c_1,\cdots,c_c)=\sum_{k=1}^{c}\sum_{i=1}^{n} u_{ik}^m d_{ik}^2 \tag{2}$$

其中：u_{ik}表示介于[0,1]间的第 i 个样本对于第 k 个类的隶属度；c_k 为类 C_k 的聚类中心；$d_{ik}=\|x_i-c_k\|$ 为第 k 个聚类中心与第 i 个数据点间的欧几里德距离；且 $m\in[1,\infty)$是一个加权指数作为模糊因子，随着 m 的增大模糊度增大，通常在不作特殊要求的情况下取 $m=2$[11]。

聚类准则是求得适当的模糊隶属矩阵 $U=\{u_{ik}\}$ 与聚类中心 c_k 使得目标函数 $J(U,c_1,\cdots,c_c)$ 达到极小值，根据拉格朗日乘数法求得 u_{ik}、c_k 分别为

$$c_k=\sum_{i=1}^{n} u_{ik}^m x_i / \sum_{i=1}^{n} u_{ik}^m \tag{3}$$

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

FCM 算法是从一个随机的聚类中心开始，通过搜索目标函数的最小值，不断调整聚类中心和每一个样本的模糊隶属度，达到确定样本类别的过程。此算法也可以先用满足式(1)的[0,1]区间的随机数初始化隶属矩阵，再执行迭代过程。

2 基于自适应权重的模糊 C-均值聚类

在 FCM 算法里，由于初始聚类中心的随机选取，导致聚类结果对初始聚类中心的过度依赖，在存在孤立点和样本分布不均衡等情况时，聚类会陷入局部最优状态。SAWFCM 算法针对 FCM 算法的上述不足进行了改进，算法的关键在于增加了自适应权重的计算策略。下面将介绍自适应权重的计算策略以及 SAWFCM 算法。

2.1 自适应权重计算

给定具有 n 个样本的数据集 $X=\{x_i, i=1,\cdots,n\}$，利用高斯距离计算样本对于聚类的影响，样本的作用随着与类心距离的不断增大而平滑降低，从而降低孤立点对聚类效果的影响。样本 x_i 在聚类过程中对于第 k 类的影响 $f(i,k)$ 计算如下：

$$f(i,k)=e^{-d_{ik}^2/\vartheta_k^2} \tag{5}$$

其中：d_{ik}为第 k 个聚类中心与第 i 个数据点间的欧氏距离；ϑ_k 表示第 k 类的有效半径，其取值为第 k 类中所有样本与聚类中心的欧氏距离的平均值，即

$$\vartheta_k=\frac{1}{|C_K|}\sum_{x_i\in c_k} d_{ik} \tag{6}$$

其中：C_k 表示属于第 k 类的样本集，$|\cdot|$为集合的势。

根据样本的划分状态，归一化单个样本的影响可以计算出样本 x_i 在聚类过程中对于第 k 类的权重 $w_{i,k}$，即单个样本对类的影响与样本集对类的影响的比值：

$$w_{i,k}=f(i,k)/\sum_{i=1}^{n} f(i,k) \tag{7}$$

归一化处理不仅确保属于某一类的样本权重总和为 1，而且由于采用高斯距离比例表示权重，从而可以有效地减弱数据密度变化对聚类效果的影响，可以在一定程度上减弱样本分布不均衡对于聚类的干扰。

2.2 加权的隶属矩阵

将上述自适应权重的计算策略与隶属矩阵相结合，采用加权的隶属矩阵来计算下一次迭代的聚类中心，使得到的聚类中心 c_k 更合理，从而提高 FCM 算法的聚类准确率和收敛速度。

加权后的隶属矩阵使得每一个样本离聚类中心距离越远其属于该聚类中心的隶属度越小。由式(8)获得的样本权重，可以按如下公式计算并更新隶属矩阵 U'：

$$u_{ik}'=w_{i,k}\otimes u_{ik} \tag{8}$$

其中：符号⊗的运算法则定义为两矩阵中对应的元素相乘所得矩阵为最终结果。例如：

$$(a\quad b)\otimes(c\quad d)=(ac\quad bd)$$

2.3 改进后的 FCM 算法

本文从自适应权重方面着手对 FCM 算法进行改进，改进算法的主要思想与 FCM 算法类似，需要根据更新后的隶属矩阵来更新聚类中心，通过不断地调整聚类中心和隶属矩阵使得目标函数达到最小值，得到趋于稳定的聚类结果。

根据加权后的隶属矩阵，重新定义 FCM 的计算公式。SAWFCM 算法的目标函数和聚类中心 c_k 为

$$J(U,c_1,\cdots,c_c)=\sum_{k=1}^{c}\sum_{i=1}^{n} u'^m_{ik} d_{ik}^2 \tag{9}$$

$$c_k=\sum_{i=1}^{n} u'^m_{ik} x_i / \sum_{i=1}^{n} u'^m_{ik} \tag{10}$$

由上述的自适应权重及距离的计算方法，可以给出基于自适应权重的模糊 C-均值聚类 SAWFCM 算法的具体步骤。

输入：数据集 X，聚类数目 c。

输出：c 个聚类。

a)初始化。给定聚类类别数 c，$2\leqslant c\leqslant n$，n 为数据集大小；设置迭代终止阈值 d；模糊因子 $m=2$。用值在[0,1]间的随机数初始化隶属矩阵 U，使其满足式(1)中的约束条件。

b)计算聚类中心 c_k。用式(3)计算 c 个聚类中心 $c_k(k=1,\cdots,c)$。

c)计算半径参数 ϑ。根据当前数据划分状态由式(6)计算半径参数 ϑ。

d)计算 $f(i,k)$。由式(5)计算每一个样本 x_i 对应其各所属类的影响 $f(i,k)$。

e)更新隶属矩阵 U。将 $f(i,k)$ 作为样本权重，依据式(8)计算修正后加权的隶属矩阵 U'。

f)更新聚类中心 c_k。依据修正后加权的隶属矩阵 U'，由式(10)更

新聚类中心 c_k。

g)若聚类趋于稳定或迭代次数超过阈值,则算法结束输出结果,否则转至步骤 c)。

2.4　迭代控制

SAWFCM 算法主要通过判定样本集与各类的均方差是否趋于稳定来停止迭代,可以根据文献[10]的迭代控制思想来度量相邻迭代步中的均方差变化

$$d = (\alpha_{\text{sum}}^{t} - \alpha_{\text{sum}}^{t-1})/\alpha_{\text{sum}}^{t-1} \tag{11}$$

其中:α_{sum}^{t}表示迭代步 t 得到的聚类结果中样本集与各类的均方差,设 c_k^t 和 C_k^t 为迭代步 t 得到的第 k 类的聚类中心和样本集,则有:

$$\alpha_{\text{sum}}^{t} = \sum_{k=1}^{c} \alpha_{C_k}^{t} \tag{12}$$

$$\alpha_{C_k}^{t} = \frac{1}{|C_k| - 1} \sum_{x_i \in C_k} \sqrt{(x_i - c_k)^2} \tag{13}$$

在具体实验中,设当 $d < 10^{-4}$时聚类结果稳定,迭代停止。

3　实验结果与分析

为了验证 SAWFCM 算法的性能,实验采用了 UCI 标准数据集[12]对改进算法的聚类效果进行测试。表 1 展示了本文所采用的数据集的基本信息。

表 1　UCI 数据集基本信息

数据集	样本点个数	维数	类数
Glass Identification	214	10	7
Iris	150	4	3
Breast cancer Wisconsin	699	9	2
Ionosphere	351	34	2
Labor	57	16	2
Hepatitis	155	19	2
Diabetes	768	8	2

3.1　标准数据集实验结果

实验将应用基本的 FCM 算法和 SAWFCM 算法对每个 UCI 数据集进行 20 次聚类,并将 FCM 算法和 SAWFCM 算法相对于齐淼等人提出的 SWFCM 算法[5],在聚类的准确率上进行比较,结果如表 2 所示。

表 2　FCM、SWFCM 与 SAWFCM 算法的性能比较

数据集	FCM 算法准确率/%	SWFCM 算法准确率/%	SAWFCM 算法准确率/%
Glass Identification	55.607	56.542	57.266
Iris	89.333	91.333	91.911
Breast cancer-W	95.279	96.424	97.644
Ionosphere	70.940	72.085	80.171
Labor	77.193	78.912	79.667
Hepatitis	79.355	*79.355*	*85.484*
Diabetes	66.667	*66.146*	*77.700*

从表 2 的实验结果中可以看出,本文中所提出的 SAWFCM 算法比传统的 FCM 算法在整体聚类精度上有很大提高,该算法相较于 SWFCM 算法也具有较高的性能,特别在 Hepatitis 和 Diabetes 两个数据集上可以看出 SWFCM 算法在准确率上等于或低于 FCM 算法,而本文的 SAWFCM 算法则高于两种算法。对 UCI 上的多个公共数据集进行聚类分析应用发现,SAWFCM 算法的稳定性高于 FCM 算法。总体上讲,SAWFCM 算法具有更好的聚类效果与稳定性,算法的健壮性也更强。

3.2　算法运行效果比较分析

为了验证改进后算法的效率,限于篇幅有限,本文取聚类的经典数据集 Iris 和 Breast cancer Wisconsin 为例,比较 FCM 算法与 SAWFCM 算法在相同迭代次数的条件下 10 次平均聚类结果的收敛情况,实验结果如图 2、3 所示。

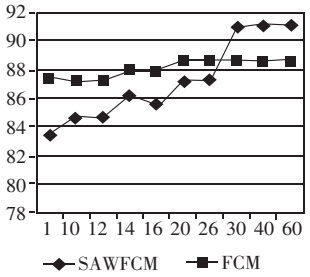


图 2　Iris 运行效果测试结果

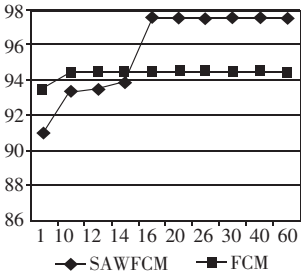


图 3　Breast cancer Wisconsin 运行效果测试结果

从图 2、3 中可以看出,虽然 FCM 算法较快地收敛到一个稳定的聚类状态,但是却由于受到初始聚类中心、孤立点或样本的不均匀分布的影响而陷入了局部最优的状态,而 SAWFCM 算法可以降低上述 FCM 算法的不足,以较快的速度得到较优的聚类效果。

4　结束语

通过对传统 FCM 算法的分析易知,该算法易受初始聚类中心和初始隶属矩阵的影响而陷入局部最优,又易受孤立点和样本分布不均衡的影响。而本文根据高斯距离所具有的密度性质,提出一种基于自适应权重的模糊 C-均值聚类算法。新聚类算法根据每次迭代的聚类结果动态地确定下一次迭代聚类中样本关于类的权重,并以此改进原始 FCM 算法对于初始聚类中心和隶属矩阵的依赖,从而使权重的选择更加科学,聚类的结果更加精确。SAWFCM 算法是一种基于目标函数的聚类算法,同传统的聚类算法一样,同样存在计算大数据量费时的问题,这个问题可以通过在此算法上采取分割数据集实现动态增量式聚类解决。

参考文献:

[1] MacQUEEN J B. Some methods for classification and analysis of multivariate observation[C]//Proc of the 5th Berkeley Symposium on Mathematical Statistics and Probability. 1967:281-297.

[2] 高新波. 模糊聚类分析及其应用[M]. 西安:西安电子科技大学出版社,2003.

[3] BEZDEK J C. Pattern recognition with fuzzy objective function algorithms[M]. New York:Plenum Press,1981:10-18.

[4] 严骏. 模糊聚类算法应用研究[D]. 杭州:浙江大学,2006.

[5] 齐淼,张化祥. 改进的模糊 C-均值聚类算法研究[J]. 计算机工程与应用,2009,45(20):133-135.

[6] 朱林,王士同,邓赵红,等. 改进模糊划分的 FCM 聚类算法的一般化研究[J]. 计算机研究与发展,2009,46(5):814-822.

[7] 魏娜,王建勋,兰文祥,等. 无监督多尺度模糊聚类算法研究[J]. 空军工程大学学报:自然科学版,2011,12(1):78-82.

[8] GRAVES D, PEDRYCZ W. Kernel-based fuzzy clustering and fuzzy clustering:a comparative experimental study[J]. Fuzzy Sets and Systems,2010,161(4):522-532.

[9] 郑超,苗夺谦,王睿智,等. 基于密度加权的粗糙 K-均值聚类改进算法[J]. 计算机科学,2009,36(3):220-222.

[10] 周杨,苗夺谦,岳晓冬. 基于自适应权重的粗糙 K-均值聚类算法[J]. 计算机科学,2011,38(6):237-242.

[11] PAL N R, BEZDEK J C. On cluster validity for the fuzzy C-means model[J]. IEEE Trans on Fuzzy Systems,1995,3(3):370 -379.

[12] UCI machine learning database[DB/OL]. http://archive.ics.uci.edu/ml/.