Doi:10.3969/j.issn.1672-0105.2015.03.013

基于Bootstrap的加权FCM改进算法*

洪年松，范 渊，郭华峰
（浙江工贸职业技术学院 信息传媒学院，浙江 温州 325003）

摘 要：为了提高模糊C-均值聚类算法的收敛速度，Yang等引入截断因子，提出了截断阈值模糊C-均值（FCMα）聚类算法，加快了算法的收敛。然而该算法也存在着聚类效果不佳的问题。针对该问题，使用Bootstrap统计方法进行特征加权，提出了基于Bootstrap的加权模糊C-均值聚类改进算法。实验表明，该算法具有更高的聚类准确率，也更有效。

关键字：FCM；截断阈值；Bootstrap；特征加权
中图分类号：TP301.6 **文献标识码：**A **文章编号：**1672-0105(2015)03-0057-03

Weighted Improved Fuzzy C-Means Clustering Algorithm based on Bootstrap

HONG Nian-song, FAN Yuan, GUO Hua-feng
(College of Information and Communications, Zhejiang Industry & Trade Vocational College, Wenzhou, 325003, China)

Abstract: To improve the convergence speed of the fuzzy C- means clustering algorithm, alpha-cut factor is introduced by Yang et al. and the alpha-cut threshold fuzzy C- means(FCMα) clustering algorithm is proposed, which accelerates the convergence. However, the FCMα algorithm also has the problem of poor clustering effect. To solve the problem, the Bootstrap statistical method is used to carry on the feature weighting, and the weighted improved fuzzy C- mean clustering algorithm based on Bootstrap is proposed. The experiments show that the proposed algorithm has higher clustering accuracy and been more effective.

Key words: FCM; alpha-cut threshold; Bootstrap; feature weighting

0 引言

为了改善模糊C–均值（FCM）聚类算法的聚类效果，学者做了多方面的研究，包括特征加权和样本加权等几个方面。特征加权方面，袁正午等通过ReliefF算法实现特征权值的自动确定，提出了一种特征加权自适应FCM算法[1]，提高了算法的聚类精度。蔡静颖等则引入自适应马氏距离对特征加权进行处理，提出了一种基于马氏距离特征加权的模糊聚类算法[2]，提升了分类的有效性。样本加权方面，刘兵等提出了一种基于样本加权的可能性模糊聚类算法[3]，该算法具有更快的收敛速度和更强的鲁棒性。段林珊等则受热力学中熵定义的启发，提出了一种基于模拟退火的样本加权FCM算法[4]，提高了分类的准确数和准确率。与FCM算法类似，FCM的改进算法——FCMα算法[5]也存在着聚类效果不佳的问题。为了解决这个问题，本文将引入Bootstrap统计方法，对数据进行特征加权，改善FCMα算法的收敛效果。

1 截断阈值模糊C-均值聚类算法

为了改善FCM算法收敛速度慢的问题，YANG等引入截断因子，提出了截断阈值模糊C–均值（FCMα）聚类算法[5]，加快了算法的收敛。FCMα算法步骤如下：

① 设定模糊指数 $1<m<\infty$，分类数 $2\leqslant c\leqslant n$，迭代次数 $l=0$，迭代终止阈值 $\varepsilon>0$，截断阈值 $0.5\leqslant\alpha\leqslant1$，初始化中心点起始值 $Z^{(0)}$。

收稿日期：2015–07–25
基金项目：温州市公益性科技计划项目（G20140049），浙江工贸职业技术学院教师科技创新活动计划项目（X140203）
作者简介：洪年松，硕士，浙江工贸职业技术学院讲师，主要研究方向：模式识别、图像处理；范渊，硕士，浙江工贸职业技术学院讲师，主要研究方向：图像处理；郭华峰，硕士，浙江工贸职业技术学院讲师，主要研究方向：图像处理、模式识别。

② 根据公式(1)计算隶属值 μ_{ij}。

$$\mu_{ij}=\left(\sum_{k=1}^{c}\frac{\left\|x_j-Z_i\right\|^{2/(m-1)}}{\left\|x_j-Z_k\right\|^{2/(m-1)}}\right)^{-1},\ j=1,2,\cdots,c,\ i=1,2,\cdots,n \quad (1)$$

③ 使用以下方法更新 μ_{ij}，i=1,...,n，j=1,...,c：如果 $\mu_{is}=\max_{1\leqslant j\leqslant c}\mu_{ij}>\alpha$，那么 $\mu_{is}=1$，$\mu_{is'}=0$，$s'\neq s$。

④ 根据公式(2)计算中心点 $Z^{(l+1)}$。

$$Z_j=\frac{\sum_i \mu_{ij}^m x_i}{\sum_i \mu_{ij}^m},\ j=1,2,\cdots,c \quad (2)$$

⑤ 若 $\left\|Z^{(l+1)}-Z^{(l)}\right\|<\varepsilon$，则停止；

否则 $l=l+1$，并回到步骤②。

FCMα算法虽然收敛速度得到了提升，但其聚类效果仍有改善的空间，下面将在该算法中引入Bootstrap特征加权方法。

2 截断阈值FCM算法的特征加权

2.1 基于Bootstrap的特征加权方法

Bootstrap统计方法是现代统计学较为流行的一种统计方法，在小样本统计时效果很好。Bootstrap方法由B. Efron提出[6]，其主要思想是从原始数据中重复抽样以帮助我们认识样本特征，从而进一步认识母体特征。其基本步骤为：① 采用重复抽样技术从原始样本中抽取给定数量的样本。② 根据抽出的样本计算给定的统计量T。③ 重复上述B次，得到B个统计量T。④ 计算上述B个统计量T的样本方差。

根据Bootstrap方法的以上特点，FCMα算法所处理的数据样本可以使用该方法来计算特征权重，其计算步骤设计如下：

① 给定重复抽样次数B，使用重复抽样技术从样本 $X=(x_1,x_2,\cdots,x_n)\subset R^P$ 中进行B次有放回的重复抽样，得到样本 $X^b=(x_1^b,x_2^b,\cdots,x_n^b)\subset R^P$，$x_i^b=(x_{i1}^b,x_{i2}^b,\cdots,x_{iP}^b)^T$，$i=1,2,\cdots,n$，$b=1,2,\cdots,B$，其中P为样本维数。

② 使用公式(3)计算各次重复抽样的特征权重：

$$w_j(b)=\frac{CV_j(b)}{\sum_{i=1}^{P}CV_i(b)},\ j=1,2,\cdots,P \quad (3)$$

其中 $CV_j(b)=\dfrac{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_{ij}^b-\bar{x}_j^b\right)^2}}{\bar{x}_j^b}$，$\bar{x}_j^b=\frac{1}{n}\sum_{i=1}^{n}x_{ij}^b$,。

③ 使用公式（4）计算样本最终的特征权重：

$$w_j=\frac{1}{B}\sum_{b=1}^{B}w_j(b),\ j=1,2,\cdots,P \quad (4)$$

2.2 基于Bootstrap的特征加权FCM改进算法

特征权重 $w=(w_1,w_2,\cdots,w_p)^T$ 获取之后，就可以代入FCMα算法，使用加权欧氏距离来代替一般欧氏距离。其加权欧氏距离定义为：

$$d_{ij}^{(w)}=\sqrt{\sum_{k=1}^{p}w_k\left(x_{ik}-z_{jk}\right)^2} \quad (5)$$

其中 w_k 表示样本在第k个特征的加权，$k=1,2,\cdots,P$。使用新距离对FCMα算法进行推算，就可以得到新的特征加权FCMα（简称WFCMα）算法，其迭代方程更新如下：

$$Z_j=\frac{\sum_i \mu_{ij}^m x_i}{\sum_i \mu_{ij}^m},\ j=1,2,\cdots,c \quad (6)$$

$$\mu_{ij}=\left(\sum_{k=1}^{c}\frac{\left(d_{ij}^{(w)}\right)^{2/(m-1)}}{\left(d_{ik}^{(w)}\right)^{2/(m-1)}}\right)^{-1},\ j=1,2,\cdots,c,\ i=1,2,\cdots,n \quad (7)$$

则新提出的WFCMα算法步骤如下：

①使用基于Bootstrap的特征加权方法计算样本特征权重 $w=(w_1,w_2,\cdots,w_p)^T$。

②设定模糊指数 $1<m<\infty$，分类数 $2\leqslant c\leqslant n$，迭代次数 $l=0$，迭代终止阈值 $\varepsilon>0$，截断阈值 $0.5\leqslant\alpha\leqslant1$，初始化中心点起始值 $Z^{(0)}$。

③根据公式(7)计算隶属值 μ_{ij}。

④使用以下方法更新 μ_{ij}，i=1,...,n，j=1,...,c：如果 $\mu_{is}=\max_{1\leqslant j\leqslant c}\mu_{ij}>\alpha$，那么 $\mu_{is}=1$，$\mu_{is'}=0$，$s'\neq s$。

⑤根据公式(6)计算中心点 $Z^{(l+1)}$。

⑥若 $\left\|Z^{(l+1)}-Z^{(l)}\right\|<\varepsilon$，则停止；

否则 $l=l+1$，并回到步骤③。

观察算法可以发现，与FCMα算法相比，WFCMα算法多了第一步，即特征权重的计算，以及特征权重到算法的代入。为了验证WFCMα算法的效果，进行如下实验。

3 仿真实验

3.1 数据集

采用正态分布的二维人工数据集和真实的高维数据集来验证算法的有效性。其中二维人工数据集有两个。第一个数据集分别以（3,4）、（7,4）为期望，单位矩阵I2为协方差矩阵。随机生成2堆正态分布的数据点，每堆100个数据，共200个数据点，如图1所示。第二个数据集分别以（3,4）为期望，单位矩阵I2为协方差矩阵，（7,4）为期望，矩阵$\begin{pmatrix}2 & 0\\0 & 4\end{pmatrix}$为协方差矩阵。随机生成2堆正态分布的数据点，每堆100个数据，共200个数据点，如图2所示。真实的高维数据集则采用经典的Iris数据集[7]。

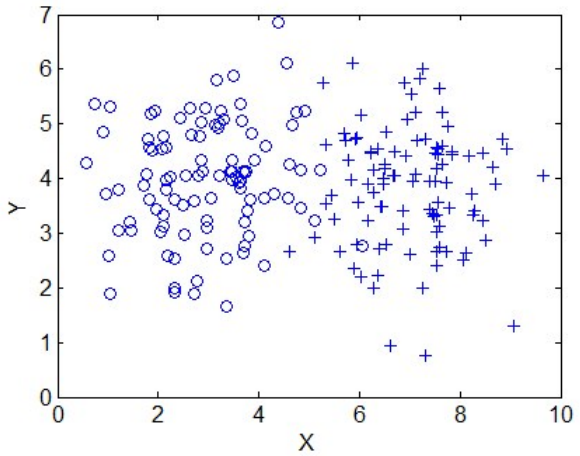

图1 随机生成的正态分布数据集1

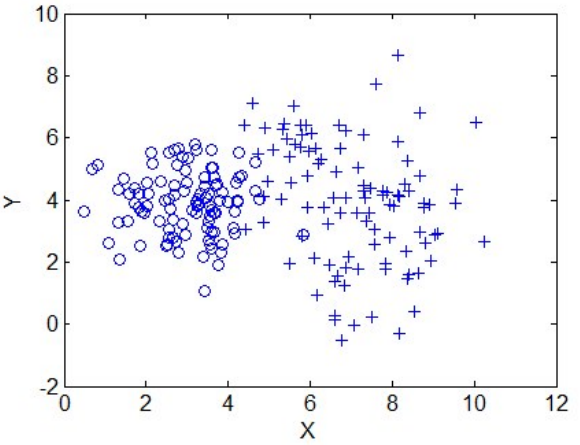

图2 随机生成的正态分布数据集2

3.2 实验

在图1和图2所示的数据集中使用新提出的WFCMα算法和FCMα算法，取模糊度m=2，分类数c=2，截断阈值α=0.6，其他的初始条件也设置相同，执行上述两种算法，得到如表1所示的结果：

表1 WFCMα算法和FCMα算法对图1和图2数据集的聚类结果

数据集	算法	权重因子	准确率
图1	FCMα	（0.5，0.5）	0.9750
	WFCMα	（0.6332，0.3668）	0.9800
图2	FCMα	（0.5，0.5）	0.9200
	WFCMα	（0.5385，0.4615）	0.9250

从表1的结果可知，相对于FCMα算法的平均权重，WFCMα算法给出了更自然的权重因子，其聚类准确率也得到了一定的提升。以上是算法在人工数据集中的表现，下面进行真实高维数据集的实验。

实验二，在Iris数据集中分别使用WFCMα算法和FCMα算法，设定分类数c=2，截断阈值α=0.6，其他的初始条件也设置相同，在模糊指数m等于1.2，2.0，5.0，10.0的情况下，执行上述两种算法，得到如表2所示的结果：

表2 WFCMα算法和FCMα算法对Iris数据集的聚类结果

m	算法	权重因子	准确率
1.2	FCMα	（0.25，0.25，0.25，0.25）	0.8933
	WFCMα	（0.1021，0.1013，0.3379，0.4588）	0.9600
2.0	FCMα	（0.25，0.25，0.25，0.25）	0.8933
	WFCMα	（0.1018，0.1027，0.3372，0.4583）	0.9400
5.0	FCMα	（0.25，0.25，0.25，0.25）	0.9133
	WFCMα	（0.1016，0.1021，0.3380，0.4582）	0.9600
10.0	FCMα	（0.25，0.25，0.25，0.25）	0.9200
	WFCMα	（0.1013，0.1024，0.3377，0.4586）	0.9533

表2的数据表明，在真实的数据集中，WFCMα算法的表现非常优异，在不同的模糊指数下都取得了更好的准确率，聚类精度更高。这说明，相比较FCMα算法的平均特征权重取值，使用Bootstrap方法得到的特征权重更符合实际，基于Bootstrap的WFCMα算法也更有效。

（下转 第77页）

“创新–发展–再创新–再发展”的循环往复过程，只有不断适应时代的需求，不断根据学校的办学特色、学科发展、生源特点适时调整，才能具备经久不衰的影响力，才能让高校走上欣欣向荣之路。

2.加强管理，保障校园文化品牌的建设投入

学校应加强对已有校园文化品牌的管理，注重管理模式的更新，建立规范有序的管理机制、评价体系和奖惩制度，克服建设的盲目性和重复性。其次高校按照建设高规格、高品位校园文化的要求保证投入力度，广开渠道筹集资金，尽可能地为那些有潜力发展为名牌的文化品牌提供物力、人力和财力支持，使之能健康、持续、稳步地发展。

参考文献:

[1] 余明阳.品牌学教程[M].上海:复旦大学出版社,2005:3–10
[2] 沈威,姜国玉.当前高校校园文化品牌培育的现状分析与对策[J].思想政治教育研究,2011(1):102–104
[3] 郑予捷.高校特色校园文化建设的实施路径探析[J].西南名族大学学报:社会科学版,2013(10):234–236
[4] 高海榕.从品牌战略管理的黄金法则看高校品牌的建设[J].山东科技大学学报:社会科学版,2010(2):98
[5] 辛承其.高校品牌经营策略研究[D].福建师范大学,2011.
[6] 蒋赟.高职校园文化品牌建设的现状与对策—基于浙江省的个案研究[J].天津职业大学学报,2013(2):13–17.
[7] 贾滕.大学特色校园文化品牌建设微探[J].学校党建与思想教育,2015(3):84–85.
[8] 周远.从市场营销学角度谈高校校园文化品牌建设[J].考试周刊,2009(22):51–52.

（责任编辑:王　艳）

（上接 第59页）

4 结论

为了提高FCMα算法的聚类效果，在算法中引入Bootstrap统计方法，提出了基于Bootstrap的特征加权FCM改进（WFCMα）算法。仿真实验表明，相对于FCMα算法的平均权重，WFCMα算法得到的样本特征权重更自然，更适用于实际情况，改进后的算法也具有更好的聚类效果，更有效。

参考文献:

[1] 袁正午,魏荣,叶明星.一种适用于基因表达数据的特征加权FCM算法[J].计算机应用研究,2010,27(7):2483–2485.
[2] 蔡静颖,谢福鼎,张永.基于马氏距离特征加权的模糊聚类新算法[J].计算机工程与应用,2012,48(5):198–200.
[3] 刘兵,夏士雄,周勇,韩旭东.基于样本加权的可能性模糊聚类算法[J].电子学报,2012,40(2):371–375.
[4] 段林珊,刘培玉,谢方方.基于模拟退火的样本加权FCM算法[J].计算机工程与设计,2013,34(6):2004–2008.
[5] M.S. Yang, K.L. Wu, J.N. Hsieh, et al.Alpha–cut implemented fuzzy clustering algorithms and switching regressions[J].IEEE Transactions on Systems, Man,and Cybernetics, 2008, 38(3): 588–603.
[6] B. Efron. Bootstrap Methods–– Another Look at the Jackknife [J]. The Annals of Statistics, 1979 ,7(1): 1–26.
[7] C.L. Blake, C.J. Merz. UCI repository of machine learning databases.[http://archive.ics.uci.edu/ml/]. Irvine, CA: University of California, Department of Information and Computer Science, 1998.

（责任编辑:潘修强）