关于 FCM 算法中的权重指数 m 的一点注记

于 剑[1],程乾生[2]
(1. 北方交通大学计算机学院,北京 100044;2. 北京大学数学科学院信息科学系,北京 100871)

摘 要: 模糊 c 均值算法(FCM)是经常使用的聚类算法之一. 使用模糊 c 均值算法时,如何选取模糊指标 m 一直是一个悬而未决的问题. 部分文献根据实验结果建议最佳的权重指数可能位于区间[1.5,2.5],但大多数研究者使用 $m = 2$. 本文阐述了 FCM 算法有效性与聚类有效性之间的理论联系,指出如果某个权重指数使得 FCM 算法作为聚类算法不能有效工作,则其不能作为最佳的权重指数. 据此,我们进行了数据实验,数据实验结果说明了权重指数的最佳取值未必位于区间[1.5,2.5].

关键词: 权重指数; 聚类有效性; FCM 算法; 划分熵
中图分类号: TP391.41 **文献标识码**: A **文章编号**: 0372-2112 (2003) 03-0478-03

A Note on the Weighting Exponent m in FCM Algorithm

YU Jian[1], CEHNG Qian-sheng[2]
(1. *School of Computer & Information Technology, Northern Jiaotong University, Beijing* 100044, *China*;
2. *School of Mathematical Sciences, Peking University, Beijing* 100871, *China*)

Abstract: The fuzzy c-means algorithm (FCM) is one of widely used clustering algorithms. It is an open problem how to select an appropriate fuzziness index m when implementing the FCM. Some researchers have suggested that the best choice for m is probably in the interval[1.5,2.5] based on their experimental results. In this paper, we discovered the theoretical connection between the validity of FCM algorithm as clustering algorithm and clustering validity, and pointed out that the weighting exponent m is not the optimal if it makes the FCM not to work properly as a clustering algorithm. According to this analysis, we carried on one experiment. The experimental result shows that the optimal weighting exponent in FCM algorithm could not always belong to the range [1.5,2.5].

Key words: weighting exponent; clustering validity; FCM algorithm; partition entropy

1 引言

FCM 算法是模糊聚类算法中的一种比较重要的算法,其目标函数 J_m 含有的权重指数,对 FCM 算法的聚类效果有重要影响. 但是至少在理论上,如何选取合适的权重指数是一个悬而未决的问题. 通过对聚类有效性函数的评估实验,文献[2]宣称是最佳的权重指数可能位于区间[1.5,2.5],大多数研究者使用 $m = 2$. (其原文为:"the best choice for m is probably in the interval[1.5,2.5], whose mean and midpoint $m = 2$, have often been the preferred choice for many users of FCM"). 2000 年,文献[1]根据他们的理论设计的实验得出的实验结果,作出了如下断言:"在实际应用中 m 的最佳取值范围为[1.5,2.5],这与 Pal 等的实验结论一致". 在下文中,我们将从算法的有效性出发,来说明对于某些数据来说,m 的最佳取值也可能不在区间[1.5,2.5]之内.

2 算法的有效性和聚类有效性

所谓 FCM 算法的最佳权重指数 m_{opt},应该是指当权重指数为 m_{opt} 时,FCM 算法此时的聚类效果要比权重指数取它值为最佳. 如何评价 FCM 算法在不同的权重指数时的聚类效果,属于聚类有效性问题. 一般地,文献中使用聚类有效性函数解决这一问题,FCM 算法聚类效果最好时对应的权重指数为最佳权重指数 m_{opt}. 本文选取划分熵作为聚类有效性函数. 注意到几乎所有讨论聚类有效性的文章,都是在假设算法是有效的基础之上的. 因此可以推断最佳权重指数 m_{opt} 至少满足如下条件:当权重指数为 m_{opt},FCM 算法作为聚类算法应该是有效的. 这实际上是一个非常自然的结论.

但是,"FCM 算法作为聚类算法是有效的"这一事实意味着什么呢?为此,我们先引入文献中的几个著名结果与划分熵的定义,所有的证明均可参考相关文献.

令 $X = \{x_1, x_2, \ldots, x_n\}$ 是数据集,$u = \{u_{ik}\}$ 是划分矩阵,$V = \{v_1, v_2, \ldots, v_c\}$ 为其对应的类中心,目标函数为 $J_m(u, v, X) = \sum_{k=1}^{n}\sum_{i=1}^{c}(u_{ik})^m \| x_k - v_i \|^2$, $1 < m < +\infty$. FCM 算法即是求出使目标函数 J_m 达到最小值的划分矩阵 $u = \{u_{ik}\}_{c\times n}$ 与类

收稿日期:2000-07-05;修回日期:2002-10-20
基金项目:教育部科学技术研究重点项目(No. 02031);北方交通大学校基金

中心 $V=\{v_1,v_2,\ldots,v_c\}$,其中 $\forall k\in\{k\mid 1\le k\le n\}$, $\forall i\in\{i\mid 1\le i\le c\}$, x_k 属于模糊集合 X_i 的隶属度为 u_{ik},并满足如下条件

$$\sum_{i=1}^{c}u_{ik}=1,\ u_{ik}\ge 0,0<\sum_{i=1}^{c}u_{ik}<n.$$

利用拉格朗日乘子法,可求得使 $J_m(u,v,X)$ 达到极小值的必要条件为[2]:

$$v_i=\frac{\sum_{k=1}^{n}(u_{ik})^m x_k}{\sum_{k=1}^{n}(u_{ik})^m} \tag{1}$$

$$u_{ik}=\left(\sum_{j=1}^{c}\left(\frac{\|x_k-v_i\|^2}{\|x_k-v_j\|^2}\right)^{\frac{1}{m-1}}\right)^{-1} \tag{2}$$

因此,模糊 c 均值聚类算法的基本步骤为如下:

Step 1. 固定聚类数 c,权重指数 m,最大迭代步数 T 和一个阈值 ε,设目标函数 $J_m(u,v,X)$ 的初始值为正无穷;

Step 2. 初始化模糊 c 划分矩阵 u,使其满足约束条件;

Step 3. 使用公式(1)和计算 c 个聚类中心 v_i,$(1\le i\le c)$;

Step 4. 计算新的目标函数 $J_m(u,v,X)$ 值. 如果它相对上次目标函数数值的减少量小于阈值或者迭代步数大于 T,则算法停止;否则,用式(2)重新计算划矩阵,返回第 3 步.

定理 1[2]　$u=\left[\frac{1}{c}\right]\Leftrightarrow v_i=\frac{\sum_{k=1}^{n}x_k}{n}$, $\forall i\in\{1,2,\cdots,c\}$,

Ruspini(1969)[3]定义了下面的划分熵:

$$V_{pe}(u,c)=-\left(\frac{1}{n\log_a c}\right)\sum_{k=1}^{n}\sum_{i=1}^{c}u_{ik}\log_a(u_{ik}),\ 1<a<+\infty$$

定理 2[2]　当 $1<c<n$,

(1) $0\le V_{pe}(u,c)\le 1$

(2) $V_{pe}(u,c)=0\Leftrightarrow u$ 硬划分;

(3) $V_{pe}(u,c)=1\Leftrightarrow u=\left[\frac{1}{c}\right]$

定理 3[4]

(1) $\lim_{m\to+\infty}u_{ik}=\left[\lim_{m\to+\infty}\left(\frac{\|x_k-v_i\|}{\|x_k-v_j\|}\right)^{2/(m-1)}\right]^{-1}=\frac{1}{c}$

其中　$1\le i\le c;1\le k\le n$

(2) $\lim_{m\to+\infty}v_i=\lim_{m\to+\infty}\frac{\sum_{k=1}^{n}u_{ik}^m x_k}{\sum_{k=1}^{n}u_{ik}^m}=\frac{\sum_{k=1}^{n}x_k}{n}$

其中　$1\le i\le c$

定理 4[2]　$\lim_{m\to+\infty}V_{pe}(u,c)=1$

由以上定理可以知道,当权重指数 m 变得很大的时候,由于 FCM 算法选出的聚类结果几乎总是数据集的中心,而不管数据本身具有多么清晰的适合 FCM 算法聚类的子类结构,显然此时的 FCM 算法已经不可用了,即此时的 FCM 算法作为聚类算法已经无效了.

反过来,如果存在一个适合于 FCM 算法聚类的数据集 X,有一个有限的权重指数 m,使得对于 FCM 算法选出的聚类结果几乎总是数据集的中心,则此权重指数 m 显然不会是最佳权重指数或最佳权重指数的侯选值. 因此,"FCM 算法作为聚类算法是有效的"这一事实至少意味着 FCM 算法选出的聚类结果不总是数据集的中心或 $u=\left[\frac{1}{c}\right]$.

下文中,将仿照[2]中的构造 Normal-4 的方法构造一个数据 Normal-12,并用它测试对 FCM 算法有效或无效的权重指数. 方法为:选定一个较大的权重指数 m 及 $c_{\max}$,如果对满足 $2\le c\le c_{\max}$ 条件的任意聚类数 c,用 FCM 算法对数据聚类得到划分矩阵 u 及类中心 V,都使得 $V_{pe}(u,c)=1$ 成立,则有 $u=\left[\frac{1}{c}\right]$,可以判断此权重指数 m 不会是最佳权重指数或最佳权重指数的候选值,然后逐步减小权重指数 m 的取值,直到 FCM 算法对数据的聚类结果不再是数据集的中心. 在本文中的实验中,我们采用 Matlab5.3 工具箱中提供的 FCM 算法,其中迭代步数为 200,阈值 $\varepsilon=1e-8$.

3　实验数据、实验结果与实验分析

实验 1　数据为 Normal-12,其构成如下:分为 12 类,每类含有 150 个样本,每个样本是一 12 维数据,类 I 的中心是矩阵 $3\times I_{12}$ 的第 I 列,每类样本皆服从类中心为其均值,方差矩阵为 $0.5\times I_{12}$ 的多元正态分布. 权重指数的变化范围为 $[1.4,2.6]$.

实验 1 的结果如下:当权重指数 m 从 1.4 以 0.1 的间隔增到 2.6,同时聚类数 c 从 2 增到 17 时,根据 FCM 算法出的聚类结果得到的 $V_{pe}(u,c)=1$ 恒为 1,由前面的定理可以知道 FCM 算法得出的聚类结果总是数据 Normal-12 的中心,因此,对数据 Normal-12 来说,权重指数属于区间 $[1.4,2.6]$ 时,FCM 算法已经失效了. 但是可以看出,在几何上,数据 Normal-12 明显有 12 类构成,每类具有球型结构,应该是适合用 FCM 来聚类的. 实际上,当 $m=1.2$ 时,FCM 算法的聚类效果非常好,得出了与期望值几乎一样的结果. 现在,我们可以证明,当权重指数属于区间[1.4,2.6]时,数据 Normal-12 的中心是 FCM 算法的一个稳定解. 具体证明参考文献[7,8]. 为了更清楚说明问题,我们构造了一个人造数据 CUBE-S.

实验 2　$CUBE\text{-}S=\{x\in R^s\mid x=y+z,\ \forall y\in A,\ \forall z\in B,\}$,这里

$$A=\left\{y=[y_1,y_2,\cdots,y_j,\cdots,y_s]\in R^s\ \middle|\ \begin{array}{l}y_j=0\text{ or }1,\\ \forall 1\le j\le s\end{array}\right\}$$

$$B=\left\{z=[z_1,z_2,\cdots,z_s]\in R^s\ \middle|\ \begin{array}{c}\exists!\,1\le j\le s\\ \text{such that } z_j=\pm s,\\ \text{else } z_j=0\end{array}\right\}$$

显然,$CUBE\text{-}S$ 空间上由 $2s$ 个类组成,每个类包含 2^s 个点.

重复对数据 Normal-12 的数值实验,得到实验结果如表 1.

由表 1 容易知道,实验 2 的更加清楚说明了本文理论分析的正确性,即对于有些数据来说,最佳权重指数并不位于区间 $[1.5,2.5]$.

表 1　$CUBE.S$ 的模糊指标实验有效值

数据维数	m 实验有效值
3	$m<3.1$
4	$m<2.1$
5	$m<1.7$
6	$m<1.5$
7	$m<1.4$
8	$m<1.34$
9	$m<1.3$

4 结论

综上所述,我们可以得出如下结论,对数据 Normal-12 来说,最佳权重指数并不位于区间[1.4,2.6],这显然有助于人们对于 FCM 算法的理解与正确使用.实际上,上述结论并不仅限于对数据 Normal-12 和 CUBE-S 成立,更深入的结果与讨论可见文[5～8].

另外,我们注意到文[1]中两种计算最佳权重指数的方法皆需依赖于数据与聚类数,以致于 FCM 算法的初值.更重要的是,文[1]中两种计算最佳权重指数的方法皆缺乏明确的理论依据,基于模糊决策的方法甚至引入了新的参数,对于如何确定新的参数文[1]并没有给出明确的理论说明,基于目标函数拐点的方法,更仅仅是由于实验结果的暗示.而本文使用聚类有效性函数来计算最佳权重指数可能存在的范围,显然理论上消除了聚类数与 FCM 算法的初值的影响,而且由于定理 1-4 的成立,本文的方法理论上得以成立的理由是不言自明的.但是,本文的计算量较大,因此,直接根据数据,理论上给出一个合理的最佳权重指数可能存在的范围是一件非常有意义的工作,我们目前正在做这方面的工作,目前的结果已经体现在文[7,8]中.

参考文献:

[1] 高新波,裴继红,谢维信.模糊 c-均值聚类算法中加权指数 m 的研究[J].电子学报,2000,28(4):80-83.

[2] N R Pal, J C Bezdek. On cluster validity for the fuzzy c-means model [J]. IEEE Trans Fuzzy Systems, 1995, 3(3): 370-379.

[3] E H Ruspini. A new approach to clustering [J]. Inform and Control, 15: 22-32.

[4] J C Bezdek. Fuzzy Mathematics in Pattern Classification [D]. Ithaca, NY: Cornell University, 1973.

[5] 于剑.聚类有效性及其应用[D].北京:北京大学,2000.

[6] Jian Yu, Qiansheng Cheng, Houkuan Huang. Analysis of the weighting exponent in the FCM [J]. IEEE Trans Syst, Man, Cybern-Part B, 2002.

[7] J Yu, H Huang, S Tian. An efficient optimality test for the fuzzy c-means algorithm [A]. Proceedings of the 2002 IEEE International Conference on Fuzzy Systems [C]. Honolulu, HI, 2002. 98-103.

[8] 于剑.模糊指标与模糊 C 均值算法[A].第二届中国 ROUGH 集与软计算学术研讨会大会邀请论文[C].中国. P. 135.

作者简介:

于 剑 男,1969 年 12 月出生于山东,博士,副教授,现为北方交通大学计算机与信息技术学院人工智能研究室主任,主要研究兴趣包含模式识别,机器学习,模糊逻辑等方向.