

文章编号:1001-9081(2010)12-3388-03

# 基于模糊相关度的模糊 C 均值聚类加权指数研究

肖满生, 阳娣兰, 张居武, 唐文评

(湖南工业大学 科技学院, 湖南 株洲 412008)

(xiaomansheng@tom.com)

**摘要:** 在极小化模糊 C 均值(FCM)聚类目标函数的过程中, 针对目前模糊加权指数  $m$  的确定缺乏理论依据和有效评价方法的问题, 提出了一种基于模糊相关度的模糊加权指数计算方法。首先定义模糊相关度的聚类有效性函数, 然后通过 Gauss 迭代计算 FCM 聚类有效性并将其反馈到模糊加权指数的变化中, 从而使  $m$  收敛到一个稳定的最优解。理论分析和实验结果表明, 该算法是有效的, 所得到加权指数  $m$  符合预期的结果。

**关键词:** 模糊加权指数; 模糊 C 均值; 聚类有效性; 模糊相关度

中图分类号: TP391 文献标志码: A

## Research of weighting exponent of fuzzy C-means algorithm based on fuzzy relevance

XIAO Man-sheng, YANG Di-lan, ZHANG Ju-wu, TANG Wen-ping

(College of Science and Technology, Hunan University of Technology, Zhuzhou Hunan 412008, China)

**Abstract:** In the process of minimization Fuzzy C-Means (FCM) clustering objective function, to solve the problem of lacking theoretical foundation and effective evaluation methodology in determining fuzzy weighted exponent "m" at present, a fuzzy weighted exponent algorithm based on fuzzy relevance was put forward. Firstly, valid function was defined based on Fuzzy relevance, then the validity of FCM clustering was calculated by Gauss iteration and its result was returned to the change of fuzzy weighted exponent, the fuzzy weighted exponent "m" will be converged to a stable optimum resolution. This algorithm is proved to be effective by theoretical analysis and experiments, and the weighted exponent "m" got from this algorithm conforms to prospective result.

**Key words:** fuzzy weighting exponent; Fuzzy C-Means (FCM); clustering validity; fuzzy relevance

## 0 引言

模糊 C 均值(Fuzzy C-Means, FCM)聚类的目标函数为:

$J_m(U, P) = \sum_{k=1}^n \sum_{i=1}^c (u_{ik})^m d_{ik}^2$ 。基于该目标函数的模糊聚类过程是一个带约束的非线性规划问题, 通过优化求解获得样本集的模糊划分, 该方法在信息技术和控制决策领域有着广泛的应用<sup>[1]</sup>。在目标函数的求解过程中, 模糊加权指数  $m$  的选择严重影响着算法的表现, 正如文献[2]指出: 当  $m = 1$  时, FCM 算法退化为 HCM 算法, 当  $m \rightarrow \infty$  时, FCM 算法得到的各个类的中心几乎都退化成数据的重心。因此, 如何选择合适的模糊加权指数  $m$  是 FCM 算法要解决的关键问题之一。

然而基于模糊集理论<sup>[3]</sup>的 FCM 算法由于发展时间短, 相关理论有待进一步完善, 如何选取最佳  $m$  尚缺乏理论指导, 目前国内外有部分文献涉及该问题, 但没有统一的标准。如文献[2]中指出指数  $m$  控制着模糊类间的分享程度, 并给出其经验取值范围为[1.1, 5], 但没有给出严格的证明; 文献[4]中分析了 FCM 算法与聚类有效性之间的联系, 指出  $m$  的区间应为[1.2, 2.5], 但没有给出最佳取值大小, 也没有对取值区间进行检验; 文献[5]中通过融合  $k$ -means 和  $k$ -modes 的模糊 FKP 算法, 得出加权指数  $m$  应小于 1.5, 由于其在文中提出的聚类有效性判断标准是一个主观标准, 不同分析者针对

不同的数据对象可能会有不同的标准要求, 因此得出的  $m$  值较小; 文献[6]从基于模糊决策的角度提出了一种  $m$  的优选方法, 该方法既考虑了样本结构的紧凑性, 也兼顾了样本的分离性, 具有隶属度简单、不含参数等优点, 但其得出的  $m$  值不稳定, 不同类间相差太大; 文献[7]中通过引入一个新的隶属度约束函数, 解决了 IFP-FCM 算法中  $m$  的一般化问题, 但该方法只是解决模糊加权指数的通用取值范围, 并指出在图像分割领域  $m$  无论取何值对聚类结果都无影响, 而在其他应用领域  $m$  值的影响需进一步验证; 文献[8]中通过面向制造单元构建的数据仿真实验, 得到  $m$  的最佳取值为 2, 并且得出, 随着  $m$  的增加, 成组效果降低, 聚类时间减少。本文提出了一种基于模糊相关度的方法来研究  $m$  的最佳取值, 在计算过程中, 基于模糊相关度的聚类有效性变化反馈到  $m$  的变化中, 通过多次递归迭代使  $m$  最终收敛到一个稳定值。该方法不但理论可靠, 而且计算简明, 易于在计算机上实现和应用。

## 1 模糊相关度及聚类有效性函数定义

在最佳模糊加权指数  $m$  的求解过程中, 聚类有效性是判断聚类结果好坏的标准之一, 目前聚类有效性测度方法包括划分系数(划分熵)法、基于贴近度理论的子集测度法。划分系数是基于数据集的模糊划分模式引入的, 没有考虑到数据集的几何结构, 子集测度法描述了一个模糊子集包含在另一个模糊集的

收稿日期: 2010-06-07; 修回日期: 2010-07-15。

基金项目: 湖南省教育厅科研项目(09C339); 湖南省科技厅科技计划项目(2008CK3083)。

作者简介: 肖满生(1968-), 男, 湖南邵东人, 副教授, 主要研究方向: 智能计算、数据挖掘; 阳娣兰(1973-), 女, 湖南株洲人, 讲师, 硕士, 主要研究方向: 模糊数学理论; 张居武(1977-), 男, 湖南邵阳人, 讲师, 硕士, 主要研究方向: 图像处理; 唐文评(1969-), 男, 湖南株洲人, 副教授, 硕士, 主要研究方向: 智能控制。

程度,但在应用过程中发现它对模糊聚类有效性的判决并不十分理想。本节基于数据集的几何结构,引入了模糊聚类的类间相关度<sup>[9]</sup>,并在此基础上定义了一个有效性函数。

### 1.1 类间模糊相关度

**模糊相关** 对于给定的数据集  $X = \{X_1, X_2, \dots, X_n\}$ , 采用 FCM 聚类后, 第  $i$  类模糊子集  $V_i$  与第  $j$  类模糊子集  $V_j$  的模糊相关定义为:

$$R(V_i, V_j) = \sum_{k=1}^n \mu_{ik}^{m/2} \mu_{jk}^{m/2} d_{ik} d_{jk}; i, j = 1, 2, \dots, c \quad (1)$$

其中:  $\mu_{ik}, \mu_{jk}$  分别为样本  $X_k$  相对于子集  $V_i, V_j$  的隶属度,  $n$  为样本数,  $c$  为聚类数,  $d_{ik} = \|X_k - P_i\|$ ,  $d_{jk} = \|X_k - P_j\|$  分别表示样本  $X_k$  到聚类中心  $P_i$  与  $P_j$  的距离, 与定义的  $V_i, V_j$  子集的模糊相关, 它只考虑了两个可能性分布子集的取值关系, 没有顾及样本点的位置和聚类中心对相关程度的影响。

**模糊相关度** 在上述定义的基础上, 第  $i$  类模糊子集  $V_i$  与第  $j$  类模糊子集  $V_j$  的模糊相关度定义为:

$$\rho_{ij} = \frac{\sum_{k=1}^n \mu_{ik}^{m/2} \mu_{jk}^{m/2} d_{ik} d_{jk}}{\sqrt{\sum_{k=1}^n \mu_{ik}^m d_{ik}^2} \cdot \sqrt{\sum_{k=1}^n \mu_{jk}^m d_{jk}^2}} = \frac{R(V_i, V_j)}{\sigma_i \cdot \sigma_j} \quad (2)$$

其中:  $\sigma_i^2 = \sum_{k=1}^n \mu_{ik}^m d_{ik}^2$  为  $V_i$  的模糊偏差, 它反映了模糊子集  $V_i$  中样本点的分布情况。

由定义可知,  $0 \leq \rho_{ij} \leq 1$ , 当  $i = j$  时,  $\rho_{ii} = 1$ , 即每一类与它本身的相关度最大。另外, 当  $m$  趋向于某一定值时, 若  $\mu_{ik} \rightarrow 1$ , 则  $\mu_{jk} \rightarrow 0$ , 若  $\mu_{jk} \rightarrow 1$ , 则  $\mu_{ik} \rightarrow 0$ , 因此  $\rho_{ij} \rightarrow 0$ ; 当  $m \rightarrow \infty$  时, 由于  $\mu_{ik}, \mu_{jk} \rightarrow [1/c]$ , 故  $\rho_{ij}$  趋向于一个常数  $\rho(0 \leq \rho \leq 1)$ 。这一结果表明, 基于模糊相关度判断聚类有效性是可行的。

### 1.2 基于模糊相关度的聚类有效性函数

对于一个好的聚类, 子类之间划分越分明越好, 因此类与类间的模糊相关度应尽可能小, 基于该思想, 定义了一个聚类有效性函数, 记为:

$$\rho(\mathbf{U}; c) = \frac{\sum_{i=1}^c \sum_{j=i+1}^c \rho_{ij}}{c(c-1)/2} - \min_{i=1}^c \left\{ \min_{j=1}^c \rho_{ij}(\mathbf{U}; c) \right\} \quad (3)$$

其中有效性函数表示  $c$  个类中所有类间相关度的平均值与  $c$  个类间最小相关度之差, 为了去掉重复计算值(即  $\rho_{ij} = \rho_{ji}$ ),  $\sum_{i=1}^c \sum_{j=i+1}^c \rho_{ij}$  中  $j$  从  $i \sim c$  范围求和, 显然对于给定的聚类数  $c$  和最佳模糊划分矩阵  $\mathbf{U}$ , 该函数可以用来判断模糊划分的结果好坏。

## 2 模糊加权指数计算

基于目标函数的 FCM 聚类中, 加权指数  $m$  对聚类过程和聚类结果有重要影响, 理论分析和应用表明, 随着  $m$  的增加, 目标函数值单调下降, 而且较大的  $m$  还有抑制噪声的功能, 因此  $m$  取值越大越好; 但另一方面, 参数  $m$  控制着 FCM 聚类结果的模糊性,  $m$  越大聚类结果越模糊, 从这个角度又希望  $m$  的取值不要太大。本文利用聚类有效性函数, 提出了一种基于 Gauss 迭代的最佳  $m$  值的计算方法, 其实现算法如图 1 所示。

#### 算法分析:

1) 模糊聚类初始化: 初始化主要包括初始聚类中心的选取和指数  $m$  的初始化。初始中心的选取直接影响 FCM 聚类结果的质量, 因此也决定着算法的成功与否, 本文采用文献[10]中提出的基于最大最小距离法来确定数据集的初始聚类中心; 模糊加权指数  $m$  初始值的选取直接影响到迭代计算效率,  $m$  值过大或过小, 都会使算法要经过更多次的迭代才能

收敛到一个稳定解, 根据文献[2]中所确定的  $m$  取值区间, 本文取  $m$  的初始值为  $m = 1.1$ , 目的是使  $m$  向递增方向收敛到一个稳定解。

2) 模糊 C 均值聚类: 即用传统的 FCM 算法对样本进行聚类, 计算在给定初始条件下最优模糊划分  $\mathbf{U}$ , 该聚类过程也是一个循环爬山迭代过程。

3) 计算聚类有效性: 在步骤 2) 得到了  $c$  个聚类的模糊划分  $\mathbf{U}$  和聚类中心  $P$  后, 采用式(3)定义函数  $\rho(\mathbf{U}; c)$  计算聚类结果的有效性, 由于类间划分越分明越好, 因此  $\rho(\mathbf{U}; c)$  越小越好, 但一般情况下  $\rho(\mathbf{U}; c) \neq 0$ ,  $\rho(\mathbf{U}; c) = 0$  代表类间完全不相关, 即硬聚类。

4) 更新模糊加权指数  $m$ :  $m$  的更新可用式(4)中的 Gauss 噪声来扰动:

$$m_i = m_{i-1} + N(0, r\sigma), i = 1, 2, \dots, t \quad (4)$$

其中:  $r$  为用户定义的比例常数,  $\sigma = \rho(\mathbf{U}; c)$  由聚类有效性确定,  $N(0, r\sigma)$  为一个 Gauss 随机变量, 代表着指数  $m$  的变化。由前面的分析可知, 分类越分明, 即类间越不相关, 则聚类有效性值越小,  $\sigma$  也就越小, 因而指数  $m$  的变化  $N(0, r\sigma)$  越小, 当聚类有效性趋近于 0 时,  $\sigma$  趋近于 0,  $N(0, r\sigma)$  也趋近于 0,  $m$  值则趋于稳定或收敛。

5) 设定一迭代次数  $t$  或一个可接受的聚类有效性值  $\rho^*(\mathbf{U}; c)$ , 重复步骤 2) ~ 4), 直到达到指定的迭代次数或  $\rho(\mathbf{U}; c) \leq \rho^*(\mathbf{U}; c)$  止, 此时的  $m$  值即为最佳模糊加权指数值。



图 1 基于 Gauss 迭代的模糊加权指数计算

## 3 实验结果与分析

为了验证上述算法的正确性并确定最佳模糊加权指数  $m$  值, 选取了 IRIS 数据集和纹理图像两类样本分别进行了实验, 实验中初始聚类中心的选取采用前面提及的最大最小距离法确定,  $m$  的初值取为 1.1, FCM 聚类算法终止条件为两次迭代之后隶属度  $|u^{(t+1)} - u^{(t)}| \leq 0.01$ , 式(4)中的比例常数  $r$  为 1。

#### 实验 1 IRIS 数据实验。

IRIS 数据集是一个著名标准测试数据集, 常用来检验聚类算法, 它由 4 维空间的 150 个样本组成, 每一个样本的 4 个分量分别表示 IRIS 数据的 petal length, petal width, sepal length, sepal width。IRIS 数据共有 3 个种类 ( $c = 3$ ) setosa、versicolor、virginica, 每一个种类有 50 个样本。实验中迭代计算的次数为 20, 图 2 是迭代计算所得到的模糊加权指数  $m$  和聚类有效性  $\rho(\mathbf{U}; c)$  之间的关系。

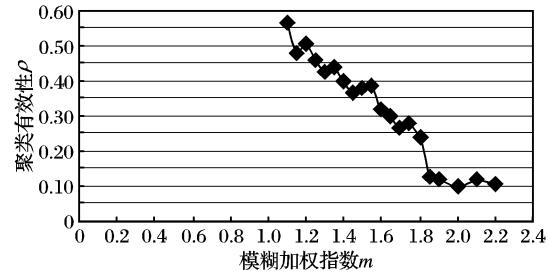


图 2 IRIS 模糊加权指数与聚类有效性关系

从图 2 中可以看出, 随着  $m$  的增加, 有效性值  $\rho$  不断减小, 当  $\rho$  趋近于 0 时,  $m$  收敛到 2.2 附近, 据此, 可取  $m = 2.2$  为最佳加权指数, 这与文献[2]指出的  $m$  应在 [1.1, 5] 相一致。

#### 实验 2 纹理图像分割。

实验采用 Brodatz 纹理库<sup>[11]</sup>中的纹理合成图像进行纹理分割测试, 图 3(a)是 Brodatz 纹理库中 D29 和 D68 合成的二

维纹理图像,其标准分割结果如图3(b)所示。

为了定量确定 $m$ 最佳取值并检验算法的稳定性,在利用Gabor滤波器对纹理图像进行特征提取后,采用FCM算法对图像进行分割测试,并依据分割测试所得的模糊划分 $U$ 和聚类中心 $P$ 分别用Gauss迭代计算方法进行了20次迭代实验,实验得到的聚类有效性与模糊加权指数的关系曲线如图4所示。

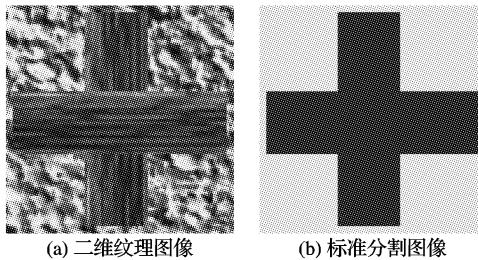


图3 纹理分割测试

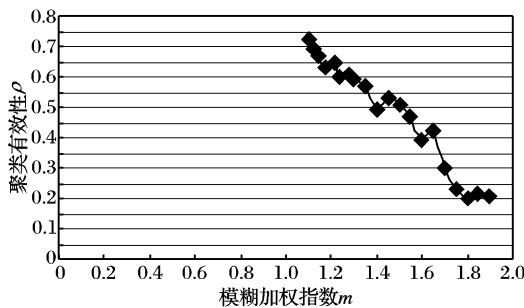


图4 纹理图像模糊加权指数与聚类有效性关系

从图4中可以看出,在 $m = 1.9$ 附近范围内,聚类有效性趋近于0且取得了一个较小值,在实验误差许可的范围内,可以认定 $m = 1.9$ 为最佳 $m$ 值,这也与文献[2]提出的 $m$ 在[1.1,5],最佳 $m$ 值应在2左右相一致。

#### 4 结语

本文通过引入基于模糊相关度的聚类有效性函数,提出

了一种新的基于模糊相关度的模糊加权指数 $m$ 最佳取值计算算法,得到了模糊加权指数最佳取值在 $m = 2$ 附近,与文献[2]分析的数据一致。该算法逻辑性强、推导合理,计算效率高,从而为模糊加权指数 $m$ 的探讨研究提供了一种新的思路和途径。但是,本文的研究还有待进一步完善,具体包括:1)实验中 $m$ 的初值从 $m = 1.1$ 开始,逐渐收敛到一稳定值, $m$ 能否从一个较大的值开始,收敛到该稳定值尚需继续实验验证;2)实验中参数 $r = 1$ 取值人为凭经验设定,缺乏理论依据,能否探讨一个有效的方法确定它,也值得继续研究。

#### 参考文献:

- [1] 修宇,王士同,吴锡生,等.方向相似性聚类方法DSCM[J].计算机研究与发展,2006,43(8):1425-1431.
- [2] BEZDEK J C. Pattern recognition with fuzzy objective function algorithm[M]. New York: Plenum Press, 1981.
- [3] ZADEH L A. Fuzzy sets [EB/OL]. [2010-04-04]. <http://www-bisc.cs.berkeley.edu/Zadeh1965.pdf>.
- [4] YU J, CHENG Q S, HUANG H K. Analysis of the weighting exponent in the FCM[J]. IEEE Transaction on System, Man and Cybernetics-Part B: Cybernetics, 2004, 34(1):634-639.
- [5] 汪加才,朱艺华.模糊K-Prototypes算法中的加权指数研究[J].计算机应用,2005,25(2):348-351.
- [6] 宫改云,高新波,伍忠东.FCM聚类算法中模糊加权指数 $m$ 的优选方法[J].模糊系统与数学,2005,19(1):143-147.
- [7] 朱林,王士同,邓赵红.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.
- [8] 李杰,徐勇,朱昭贤.模糊C均值算法参数仿真研究[J].系统仿真学报,2008,20(2):509-513.
- [9] 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,2004:112-113.
- [10] 周清,熊忠阳,张玉芳,等.基于最大最小距离法的多中心聚类算法[J].计算机应用,2006,26(6):1425-1427.
- [11] Trygve randem, brodatz textures [EB/OL]. [2007-08-01]. <http://www.ux.uis.no/~tranden/Brodatz.html>.

(上接第3387页)

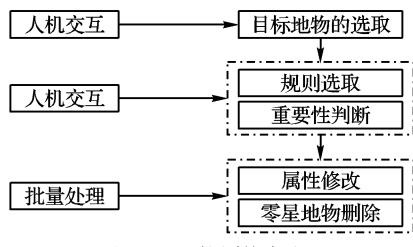


图8 人机协同综合流程

#### 4 结语

本文的研究能解决零星地物、线状地物、地类图斑三大要素的自动综合问题,但是还有待于从以下方面进一步完善。

1)本文的自动综合过程是制定好各类要素的综合规则和对应的知识库,在此基础之上对不同的土地利用数据进行自动综合操作,但是由于不同的土地利用数据存在不确定性,综合规则不可能适应所有土地利用数据。因此应该进一步研究知识库的自动学习和更新能力规则,并提供相应的算法,使知识库在每一次综合操作工程中能提高自身的健全性,提高整个系统的鲁棒性。

2)在综合操作中,对于每类要素数据的综合时,现在提供的综合算法是通过大量的循环和判断来寻求最佳解,达到自动综合的目的。这种算法效率较低,应进一步研究出效率更高的算法来适应综合操作。

#### 参考文献:

- [1] 武芳,钱海忠,邓红艳,等.面向地图自动综合的空间信息智能处理[M].北京:科学出版社,2008:175-246.
- [2] 贾泽露,刘耀林.ALCGEIS知识获取与推理机设计[J].地球信息科学,2006,8(1):67-72.
- [3] 岳河海.地图综合基础理论与技术方法研究[M].北京:测绘出版社,2004:34-58.
- [4] 王家耀,武芳.数字地图自动综合原理与方法[M].北京:解放军出版社,1998:40-52.
- [5] 陈伟伟.土地利用数据库综合的结构化模型和算法研究[D].武汉:武汉大学,2005.
- [6] 刘颖.空间图形的表达、识别与综合[D].郑州:信息工程大学,2005.
- [7] 高文秀.基于知识的GIS专题数据综合的研究[D].武汉:武汉大学,2002.
- [8] 国土资源部.县(市)级土地利用数据库标准(试行)[S],2002.
- [9] TURNER B L, SKOLE D L, FISCHER S S G, et al. Land-use and land-cover change: Science/research plan, IGBP Report No. 35. HDP Report No. 7[R]. Stockholm and Geneva, 1995: 132.
- [10] DUTON G. Scale, sinuosity, and point selection in digital line generalization [J]. Cartography and Geographic Information Science, 1999, 26(1): 33-53.
- [11] 牛方勇,甘国辉,程昌秀,等.矢量数据综合规则表达与实现方法——以土地利用数据综合为例[J].地球信息科学,2009,11(4): 421-427.