◎数据库与信息处理◎

改进的快速模糊 C-均值聚类算法

陈松生,王 蔚
CHEN Song-sheng,WANG Wei

南京师范大学 教育科学学院 机器学习与认知实验室,南京 210097
ML&C Lab,School of Education Science,Nanjing Normal University,Nanjing 210097,China
E-mail:njuwangwei@hotmail.com

CHEN Song-sheng,WANG Wei.Modified fast fuzzy C-means clustering algorithm.Computer Engineering and Applications, 2007,43(10):167-169.

Abstract: The Fuzzy C-Means (FCM) clustering algorithm requires a long time,due to processing the large data set.This paper presents a method to speed up the FCM algorithm using cluster centers obtained by the multi-times random sampling clustering as the initial cluster centers for the FCM algorithm to reduce the number of iterations required for convergence,and for optimization of the data set to reduce the time for each iteration.This method enormously accelerates the FCM algorithm while maintaining the clustering accuracy.

Key words: fuzzy clustering analysis;fuzzy c-means;multi-times random sampling;data reduction

摘 要:为解决模糊 C-均值(FCM)聚类算法在大数据量中存在的计算量大、运行时间过长的问题,提出了一种改进方法:先用多次随机取样聚类得到的类中心作为 FCM 算法的初始类中心,以减少 FCM 算法收敛所需的迭代次数;接着通过数据约减,压缩参与迭代运算的数据集,减少每次迭代过程的运算时间。该方法使 FCM 算法运算速度大大提高,且不影响算法的聚类效果。

关键词:模糊聚类分析;模糊 C-均值;多次随机取样;数据约减
文章编号:1002-8331(2007)10-0167-03 **文献标识码**:A **中图分类号**:TP391

1 引言

模糊聚类分析作为非监督机器学习的主要技术之一,建立了样本类属的不确定性的描述,能够比较客观地反映现实世界[1],在数据挖掘、图像分割、矢量量化、模式识别、模糊逻辑等诸多领域有着广泛地应用。在众多的模糊聚类算法中,应用最广泛且较成功的是 1974 年由 Dunn 提出并由 Bezdek 加以推广的模糊 C-均值(Fuzzy C-Means,简称 FCM)算法[2]。但其有一些自身的缺点:(1)聚类的类数不能自动确定,使用时必须确定聚类的有效性准则;(2)类中心的位置和特性不一定事先知道,必须由随机初始化产生;(3)对大的数据集进行聚类时,运算的开销太大;(4)在很多情况下,算法对噪音数据比较敏感。

针对 FCM 算法在大数据量中存在的计算量大、运行时间过长的问题,许多人对算法进行改进,并提出了新的算法。文献[3,4]采用网格将原数据集划分成大量子集,求出每个子集的质心,用所有子集的质心构成的子集作为原数据集的近似,然后对该子集采用算法进行聚类,把求得的最终类中心作为原数据集的初始化值。文献[5-7]通过对数据集的特征值的量化、合并、聚合,使数据集压缩、减少,从而把对数据集的聚类转化为对特征集的聚类。

本文结合两者优势,提出一种改进算法:先用多次随机取样聚类得到的类中心作为 FCM 算法的初始类中心,以减少 FCM 算法收敛所需的迭代次数;接着通过数据约减,压缩参与迭代运算的数据集,减少每次迭代过程的运算时间;最后用 FCM 算法进行聚类。该方法使 FCM 算法运算速度大大提高,且不影响算法的聚类效果。

2 FCM 算法

FCM 算法是把 n 个数据 $x_i(i=1,2,\cdots,n)$ 分成 c 个模糊簇,并求得每个簇的类中心,使目标函数达到最小。FCM 算法目标函数为:

$$J_m(U,V)=\sum_{i=1}^{n}\sum_{j=1}^{c}(u_{ij})^m(d_{ij})^2 \tag{1}$$

这里 $\sum_{j=1}^{c}u_{ij}=1,u_{ij}\in(0,1),\forall i,d_{ij}=\|x_i-v_j\|$。其中:$X=\{x_1,x_2,\cdots,x_n\}$ 为数据集,m 为模糊加权指数且 $\infty>m\geqslant 1$,c 为聚类的类别数且 $c\geqslant 2$,$\boldsymbol{U}=\{u_{ij}\}$ 表示隶属度矩阵,u_{ij} 是第 j 类中样本 x_i 的隶属度,$\boldsymbol{V}=\{v_j\}$ 表示类中心矩阵。为使目标函数 J_m 达到最小,类中心和隶属度的更新如下:

$$v_j=\frac{\sum_{i=1}^{n}(u_{ij})^m\cdot x_i}{\sum_{i=1}^{n}(u_{ij})^m},j=1,2,\cdots,c \tag{2}$$

基金项目:教育部留学回国人员科研启动基金(The Project-sponsored by SRF for ROCS, SEM);教育部"十五"规划重点项目(No.DCA050056)。
作者简介:陈松生(1975-),男,硕士研究生,主要研究方向:机器学习、数据挖掘;王蔚(1966-),女,教授,博士,主要研究方向:模式识别、机器学习。

$$u_{ij}=\left[\sum_{k=1}^{c}\left(\frac{\|x_i-v_j\|}{\|x_i-v_k\|}\right)^{2/(m-1)}\right]^{-1},i=1,2,\cdots,n \tag{3}$$

当 $d_{ij}=0$ 时，则 $u_{ij}=1,u_{ik}=0,k\neq j,i=1,2,\cdots,n$。

3 FCM 的改进

FCM 算法的性能强烈地依赖类中心的初始化，但其初始类中心又是随机选取的。如果能选择与实际类中心近似的初始类中心，将减少算法的迭代次数，缩减聚类时间，并很快收敛于实际的类中心。同时，由于在 FCM 算法中，每一步迭代都要对整个数据集进行计算，运算量很大。通过数据约减，可减少参与计算的数据，降低运算量，大大加速运算时间。通过优势整合，优化设计，提出改进的 FCM 算法。

3.1 多次随机取样聚类，减少迭代次数

FCM 算法利用随机函数对类中心初始化。这种初始化的随意性，使 FCM 需要很长的迭代过程才能达到收敛结果。而 FCM 算法是一个从初始值到最优解迭代寻优的过程，所以当所赋类中心初值非常接近实际类中心时，FCM 的迭代次数将明显减少。在文献[4,8]中，基于这样的假设：通过对一个小的原始数据子集聚类可以得到接近整个数据的类中心。经多次随机抽样聚类来寻找接近实际类中心的初始化类中心。面对大数据量，特别是多维数据，这种方法极为有效地减少迭代次数，缩短运算时间，提高运算速度。

多次随机取样聚类是在原数据集中随机选取一个合适大小的数据子集，用 FCM 算法聚类得到类中心。然后从剩下的未经聚类的原数据集中随机抽取一个小的数据子集，并与以前抽取到的数据子集合并，形成一个新的数据子集，用以前得到的类中心为初始中心，对新的数据子集进行 FCM 算法聚类，得到一个新的类中心。上面的步骤重复多次，直到合并形成的数据子集足够大，使得到的类中心接近整个数据集实际的聚类中心。其中，有 4 个变量的选取是非常关键的：一是在原始数据集 X 中选取的 n 个样本所占的比重 $\Delta\%$；二是取样的次数 r；最后两个是第一次和最后一次聚类的阈值 $\varepsilon_{firststage}$、$\varepsilon_{laststage}$。

将最后得到的类中心作为整个数据集的初始类中心，用 FCM 算法聚类。这种改进的算法称为 RFCM 算法。

3.2 采用数据约减，加速迭代时间

算法中常常假定：一些小的不为人察觉的数值的变化不会影响到对象的分类[10]。在 FCM 算法中，每一步迭代都要对整个数据集 X 进行计算，运算量很大。文献[9,10]采用对数据压缩、合并的方法，减少参与计算的数据量，降低运算开销，极大地减少 FCM 算法每次迭代过程的时间，提高计算速度。

数据约减是对特征向量进行压缩，使聚类的不同的样本数目从 n 减少到 p，且 p 远远小于 n，并保持良好的划分，从而减少 FCM 每次迭代过程的时间。

显然，在数据集 X 中，具有相同特征的向量应该属于同一类别。故数据约减形式化描述为：设数据集 X 中有 p 种特征的向量，则可以得到新的数据集 $X'=\{x'_1,x'_2,\cdots,x'_p\}$，每种特征对应的向量个数为 $H=\{h_1,h_2,\cdots,h_p\}$，即 h_i 为聚合到中的特征向量的数目。例如对于大小为 256×256 像素、灰度级为 0~255 的图像，X 有 65 536 个向量，而约减后的最多有 256 个元素。

在将数据集 X 约减为 X' 后，用 FCM 算法对数据集 X' 聚类。这种改进的算法称为 DFCM算法。

3.3 改进的 FCM 算法

本文从多次随机取样聚类减少迭代次数、采用数据约减节省迭代时间两个方面对 FCM 算法进行了改进。改进算法主要思想是：先对整个数据集采取多次随机取样聚类，得到类中心；接着利用数据约减，压缩数据集；最后以前面得到的类中心为初始类中心，用 FCM 算法对压缩后数据集进行聚类，这种改进的 FCM 算法记为 MFCM 算法。其实现过程为：

(1)对整个数据集 X 采取多次随机取样聚类，得到的类中心，记为 V_{first}。

(2)通过数据约减，将数据集 X 压缩为 X'，得到 p 和 H。

(3)以 V_{first} 为初始类中心，对数据集 X' 进行 FCM 算法，隶属度和类中心的更新如下：

$$u_{ij}=\left[\sum_{k=1}^{c}\left(\frac{\|x'_i-v_j\|}{\|x'_i-v_k\|}\right)^{2/(m-1)}\right]^{-1},i=1,2,\cdots,p \tag{4}$$

$$v_j=\frac{\sum_{i=1}^{p}h_i(u_{ij})^m\cdot x'_i}{\sum_{i=1}^{p}h_i(u_{ij})^m},j=1,2,\cdots,c \tag{5}$$

4 实验结果

实验对两类数据进行测试。一类实验数据来自 IRIS 数据(http://www.ics.uci.edu/~mlearn/databases/)。它由 150 个样本点组成，其实际类中心为：$v_1=(5.00,3.42,1.46,0.24)$，$v_2=(5.93,2.77,4.26,1.32)$，$v_3=(6.58,2.97,5.55,2.02)$。为了测试本文提出的改进算法对大数据量的聚类性能，将著名的 IRIS 数据扩大 400 倍至 60 000 个数据作为测试样本集。取 $c=3,m=2$，$\varepsilon_{firststage}=1e-4,\varepsilon_{laststage}=1e-6,r=7,\Delta\%=4\%,p=150$，采取相同的随机初始化类中心，用 4 种算法分别重复 10 次实验。实验在 Intel(R) 4、CPU2.93 GHz、512 M 内存的微机上进行。表 1 给出了聚类后的类中心、正确率及运算时间的平均值。

表 1 4 种算法对 IRIS 数据聚类结果的比较

算法	类中心	正确率/%	运算时间/s
FCM	v_1=(5.003 6,3.403 0,1.485 0,0.251 5) v_2=(5.889 2,2.761 2,4.364 3,1.397 4) v_3=(6.775 1,3.052 4,5.646 9,2.053 6)	89.33	4.208 33
RFCM	v_1=(5.003 6,3.403 0,1.485 0,0.251 5) v_2=(5.889 2,2.761 2,4.364 3,1.397 4) v_3=(6.775 1,3.052 4,5.646 9,2.053 6)	89.33	1.381 06
DFCM	v_1=(5.002 8,3.403 2,1.484 8,0.251 3) v_2=(5.889 1,2.760 9,4.363 5,1.398 0) v_3=(6.775 3,3.053 0,5.646 8,2.052 9)	89.33	0.046 875
MFCM	v_1=(5.002 6,3.403 8,1.485 1,0.251 4) v_2=(5.889 0,2.760 7,4.363 2,1.398 1) v_3=(6.775 2,3.052 8,5.646 5,2.053 2)	89.33	0.025 625

另一类数据是人脑 MR 图像(http://www.cma.mgh.harvard.edu/ibsr/)。因为人脑 MR 图像主要包括脑白质、脑灰质、脑脊液和背景 4 部分。取 $c=4,m=2,\varepsilon_{firststage}=1e-2,\varepsilon_{laststage}=1e-5,r=7,\Delta\%=4.5\%$，对一副大小为 256×256 像素、灰度级为 0-255 的人脑 MR 图像，用以上 4 种算法以灰度为特征分别进行 10 次聚类分割。表 2 给出了 10 次聚类后的类中心和运行时间的平均值。

从表 1、表 2 都可以看出，用多次随机取样聚类得到类中心作为 FCM 的初始类中心可以减少 FCM 的迭代次数，RFCM

算法运算速度提高到 3 倍以上;采用数据约减,缩小数据集,可以减少 FCM 算法每次迭代的运算时间,DFCM 算法将运算时间缩短了 80-90 倍。而 MFCM 算法使运算速度提高了 100 多倍。同时改进的 FCM 算法相对于 FCM 算法,聚类效果基本一致。

表 2　4 种算法对灰度图像分割实验结果

算法	类中心				运算时间/s
	脑白质	脑灰质	脑脊液	背景	
FCM	147.784	109.449	67.133	0.045	4.962
RFCM	147.784	109.449	67.133	0.045	1.546
DFCM	147.778	109.437	66.988	0.044	0.057
MFCM	147.780	109.441	67.095	0.044	0.038

5　结语

FCM 算法是一种经典的模糊聚类分析方法。本文针对 FCM 算法面对大数据集,运算量过大、时间过长的缺点,采用多次随机取样聚类得到初始化类中心和数据约减的方法来减少数据量,提出改进的 FCM 算法。它在大大提高计算速度的同时,也保证聚类效果的一致。这个将多种方法和手段相结合的改进思路也为以后的研究工作提供了更为广阔的视野和丰富的理念。

当然,在使用多次随机取样聚类时,参数的选取有一定的困难,且难以把握;同时文中的数据约减方法只是对特征值进行聚合,如何更有效地用于高维空间数据的处理,并使数据压缩量进一步提高,值得深入研究。(收稿日期:2006 年 12 月)

参考文献:

[1] 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.

[2] 张敏,于剑.基于划分的模糊聚类算法[J].软件学报,2004,15(6):858-868.

[3] Wang W,Zhang J,Wang H.Grid-ODF:detecting outliers effectively and efficiently in large multi-dimensional databases[C]//Lecture Notes in Artificial Intelligence,Springer Inc,2005.

[4] Hung M C,Yang D L.An efficient fuzzy c-means clustering algorithm[C]//IEEE International Conference on Data Mining(ICDM2001),California,USA,2001:225-232.

[5] Eschrich S,Ke J W,Hall L O.Fast fuzzy clustering of infrared images[C]//IFSA World Congress and 20th NAFIPS International Conference,Joint 9th,Vancouver,BC,Canada,2001,2:1145-1150.

[6] Zhang J,Wang W.A color indexing scheme using two-level clustering processing for effective and efficient image retrieval[C]//The 2005 International Conference on Data Mining(DMIN'05),Las Vegas,USA,2005:20-23.

[7] Ke J W.Fast accurate fuzzy clustering through reduced precision[D].South Florida:University of South Florida,1999.

[8] Cang T W,Goldof D B.Fast fuzzy clustering[J].Fuzzy Sets and Systems,1998,93(1):49-56.

[9] Pal N R,Bezdek J C.Complexity reduction for"large image"processing[J].IEEE Trans on Systems,Man and Cybernetics-Part B,2002,32(5):598-611.

[10] Eschrich S,Hall L O.Fast accurate fuzzy clustering through data reduction[J].IEEE Trans on Fuzzy System,2003,11(2):262-270.

(上接 155 页)

图 3 为第 48 天 14:00-14:15 的负载预测值与实际值的比较及预测误差条形图。

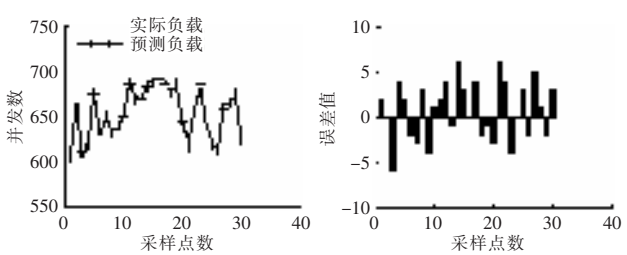

图 3　预测结果示例及误差

为评估本算法的性能,基于同一组数据分别建立 BP 预测模型和 AR(6)模型,进行单步预测,并分别计算其均方误差(式(9)),结果见表 2。

$$\sigma=\sum_{t=1}^{N}\frac{(x_t+\hat{x}_t)^2}{N} \tag{9}$$

其中,x_t 是流量真实值,$\hat{x}_t$ 是预测值,N 是样本数。

表 2　三种预测方法的均方误差

预测方法	本文方法	BP 模型	AR(6)模型
均方误差	10.51	13.060 2	19.162 7

显然,本文方法明显优于其它两种预测方法。

5　结论

本文将小波引入服务器负载预测,并针对小波分解后各信号不同特点建立不同预测模型,取得了较好的预测效果。但是,在实验中发现:小波函数的选择是影响模型预测精度的关键因素之一。在今后的研究工作中将对该问题作进一步研究。(收稿日期:2006 年 8 月)

参考文献:

[1] Dinda P.Online prediction of the running time of tasks[C]//Proceedings of 10th IEEE International Symposium on High Performance Distributed Computing.San Francisco,CA,USA:IEEE Press,2001:383-394.

[2] Khotanzad A,Sadek N.Multi-scale high-speed network traffic prediction using combination of neural networks[C]//Proceedings of the International Joint Conference on Neural Networks.Portland,OR,USA:IEEE Press,2003:1071-1075.

[3] 汤勇平.Java 并行计算环境中的负载监测与预报系统[D].上海:上海交通大学,2002.

[4] 冉启文,单永正,王骐,等.电力系统短期负荷预测的小波-神经网络-PARIMA 方法[J].中国电机工程学报,2003,23(3):38-42.

[5] Mallat S G.A theory for multiresolution signal decomposition:the wavelet representation[J].IEEE Tran on Pattern Analysis and Machine Intelligence,1989,11(7):674-693.

[6] 张树京,齐立心.时间序列分析简明教程[M].北京:北方交通大学出版社,2003.