K-means算法初始聚类中心选择的优化

冯 波,郝文宁,陈 刚,占栋辉
FENG Bo, HAO Wenning, CHEN Gang, ZHAN Donghui

解放军理工大学 工程兵工程学院,南京 210007
Engineering Institute of Corps of Engineers, PLA University of Science & Technology, Nanjing 210007, China

FENG Bo, HAO Wenning, CHEN Gang, et al. Optimization to *K*-means initial cluster centers. Computer Engineering and Applications, 2013, 49(14):182-185.

Abstract: To solve this problems that the traditional *K*-means algorithm has sensitivity to the initial cluster centers, a new improved *K*-means algorithm is proposed. The algorithm builds minimum spanning tree and then splits it to get *K* initial clusters and the relevant initial cluster centers. The initial cluster centers are found to be very closed to the desired cluster centers for iterative clustering algorithms. Theory analysis and experimental results demonstrate that the improved algorithms can enhance the clustering performance, get stable clustering in a higher accuracy.

Key words: *K*-means algorithm; clustering; initial clustering centers; TDKM algorithm

摘 要: 针对传统*K*-means算法对初始聚类中心敏感的问题,提出了基于数据样本分布情况的动态选取初始聚类中心的改进*K*-means算法。该算法根据数据点的距离构造最小生成树,并对最小生成树进行剪枝得到*K*个初始数据集合,得到初始的聚类中心。由此得到的初始聚类中心非常地接近迭代聚类算法收敛的聚类中心。理论分析与实验表明,改进的*K*-means算法能改善算法的聚类性能,减少聚类的迭代次数,提高效率,并能得到稳定的聚类结果,取得较高的分类准确率。

关键词: *K*-means算法;聚类;初始聚类中心;TDKM算法
文献标志码: A **中图分类号:** TP181 doi:10.3778/j.issn.1002-8331.1111-0289

聚类分析是数据挖掘领域一个重要的研究课题[1]。聚类就是将物理或抽象对象的集合分成相似的对象类的过程,同一个簇中的对象之间具有较高的相似度,而不同簇中的对象差别较大。通过自动聚类能够识别对象空间中稠密和稀疏区域[2],从而发现全局分布模式和数据属性之间有趣的相关。同时聚类分析可以作为其他算法(例如特征化[3]、属性子集选择以及分类)的预处理步骤。*K*-means[4]即*K*均值是一种基于划分的聚类算法,它以误差平方和SSE(Sum of the Squared Error)作为度量聚类质量的目标函数。传统的*K*-means算法随机选取初始聚类中心,算法容易陷入局部最优。如何选取一组合理的初始聚类中心,在降低聚类结果波动性的同时,又能得到较高的聚类准确率具有较高的意义。

为了克服*K*均值算法的一些缺陷,很多学者都试图从不同的角度对*K*均值算法进行改进。总体上初始聚类中心的选择可以分为3种:随机抽样、距离优化和密度估计。文献[5]运用距离代价函数作为聚类有效性检验函数,当距离代价函数达到最小值时,空间聚类结果为最优。文献[6]提出基于最大最小距离法寻找初始聚类中心,算法能够找到最佳的聚类数目*K*以及合理的初始聚类中心。文献[7-8]根据聚类对象分布密度来确定初始聚类中心。文献[9]用密度函数法求得样本数据空间的多个聚类中心,并结合小类合并运算,能很好地避免局部最小。文献[10]提出了半监督*K*-means的*K*值全局寻优算法,利用少量数据来指导和规划大量无监督数据。文献[11]提出利用图论知识迭代得到稳定的聚类结果。文献[12]提出了一种基于密度估计选取初始聚类中心的方法KNN(*K*-nearest Neighborhood),算法迭代寻找*K*个数据作为初始聚类中心。文献[13]提出了初始化*K*-means的谱算法。文献[14]提出了一种密度敏感的相似性度量,将其引入谱聚类得到密度敏感的谱聚类算法。Leng等人在文献[15]中提出了一种基于影响因子的*K*-means算法。本文提出了基于最小生成树以及树的剪枝的方法将数据点划分为*K*个初始的聚类簇,并计算出初始的聚类中心。实验表明,改进的聚类算法得到了稳定的聚类结果,降低了聚类过程中迭代的次数,并取得了较高的准确率。

作者简介: 冯波(1987—),男,硕士研究生,研究方向:数据库、数据挖掘;郝文宁(1971—),男,博士,副教授,主要研究方向:军用数据工程、海量高维数据规约、作战效能评估;陈刚,男,副教授;占栋辉,男,硕士研究生。E-mail:fengbogjk@163.com

收稿日期: 2011-11-16 **修回日期:** 2012-03-05 **文章编号:** 1002-8331(2013)14-0182-04
CNKI出版日期:2012-04-25 http://www.cnki.net/kcms/detail/11.2127.TP.20120425.1720.050.html

1　TD(Tree Distribution)算法

K-means算法不同的初始值可能会导致不同的聚类结果。聚类结果对初始聚类中心的依赖性，导致了聚类结果的不稳定性。传统算法同时易受噪音数据的干扰，影响聚类结果的准确性。提出的TD算法，按照数据分布动态地选取初始聚类中心。由此得到的初始聚类中心符合数据的真实分布，更加具有现实意义。

1.1　TD算法的思想

假设样本集 U 含有 N 个样本数据，样本数据有 m 个属性，则数据 x 可以表示为 $x=(x_1, x_2, \cdots, x_m)$。

定义1　数据 x 和数据 y 之间的距离定义为欧氏距离：

$$dist[x,y]=\sqrt{(x_1-y_1)^2+(x_2-y_2)^2+\cdots+(x_m-y_m)^2} \quad (1)$$

定义2　数据 x_i 和其他每一个数据点的距离和定义为 $sum[x_i]$：

$$sum[x_i]=\sum_{j=1}^{N} dist[x_i,x_j] \quad (2)$$

定义3　集合 U 中的距离和均值定义为 $avg[U]$：

$$avg[U]=\sum_{i=1}^{N} sum[x_i]/N \quad (3)$$

在*K*-means算法中，数据之间的相似度是用欧式距离来衡量的，距离越小越相似，数据分布的密集区域是形成聚类簇的地方。如果选取这些密集区域数据的平均值作为初始的聚类中心，将有利于目标函数的收敛。

TD算法思想是首先排除集合中的孤立点和噪音数据，减少这些数据对聚类过程的影响。然后利用样本集合的距离矩阵构造最小生成树，最后是对树进行剪枝，划分成 k 棵子树，每棵子树代表一个初始聚类簇，而子树中数据的算术均值作为初始的聚类中心。算法的实现分为3个步骤。(1)去除噪音数据的过程。根据式(1)计算出两两数据的距离 $dist[i,j](1\leqslant i\leqslant N, 1\leqslant j\leqslant N)$，由此得到样本集的距离矩阵。根据式(2)计算出每个数据点的距离和，式(3)计算出样本集的距离和均值。利用噪音数据和孤立点数据的距离和大于样本集合的距离和均值的特性，将样本集合中的噪音数据和孤立数据删除。(2)多次扫描距离矩阵，生成样本集的最小生成树。最小生成树能够真实地反应出数据的分布，具有较强的代表性。(3)遍历最小生成树，将最小生成树中权值最大的 $K-1$ 个树枝进行剪枝，生成 K 棵子树。每棵子树代表一个初始的聚类簇，子树中的数据点的算术均值为初始的聚类中心点。

假设现有一个二维的数据样本集合，含有7个数据点，欲分成2个簇。样本集 $U=\{a=(1,3), b=(2,3), c=(2,4), d=(5,5), e=(6,6), f=(6,1), g=(1,1)\}$。按照TD算法的思想首先计算出这7个数据两两之间的距离，得到距离矩阵 $\boldsymbol{A}$ 以及各数据点的距离和样本距离和均值。由于 $sum[f]>avg[U]$，f 点被看作是孤立点而从样本集 U 中删除。扫描距离矩阵 $\boldsymbol{A}$，依次扫描得到矩阵中最小距离，根据最小生成树的算法得到样本数据的最小生成树。因为需要分成两个簇，遍历树将最大的距离 $dist[c,d]$ 剪枝，生成了两棵子树。根据子树得到初始的簇 $c_1=(a\ b\ c\ g)$、$c_2=(d\ e)$，计算出各簇的初始聚类中心 $C[i]$。样本的数据最小生成树如图1所示。

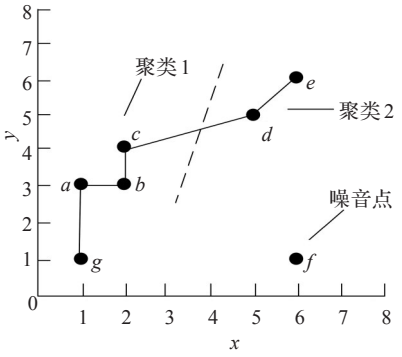

图1　二维数据最小生成树

TD算法的过程可以看出，最终形成的子树是以集合中各数据为中心。最终得到的簇是以彼此距离靠近的数据构成，处于数据的高密度区域。分别取2个簇中数据的算术均值，则确定了2个初始的聚类中心。这样得到的初始聚类中心符合数据的实际分布，从而能得到更好的聚类结果。

1.2　TD算法流程图及描述

TD算法能够找到数据在空间分布上相一致的且可代表各个簇的初始聚类中心。图2是算法的流程图，具体的算法描述如下：

输入：含有 N 个样本的数据集合 U，聚类个数 K

输出：K 个初始的聚类中心

(1)利用式(1)计算两两数据点的距离 $dist[i,j](1\leqslant i\leqslant N, 1\leqslant j\leqslant N)$，形成距离矩阵 $\boldsymbol{D}$；利用式(2)计算出各数据的距离和，式(3)计算样本距离和均值。

(2)样本集合中删除距离和大于样本距离和均值的孤立点。

(3)更新样本集和样本的距离矩阵；得到样本集 U_1 和矩阵 $\boldsymbol{D}_1$。

(4)调用 genMST (U_1, $\boldsymbol{D}_1$), 构造最小生成树 T。

(5)根据权值降序剪枝 $K-1$ 个树分枝，得到 K 棵子树。

(6)计算各子树中数据的算术均值记做数据中心

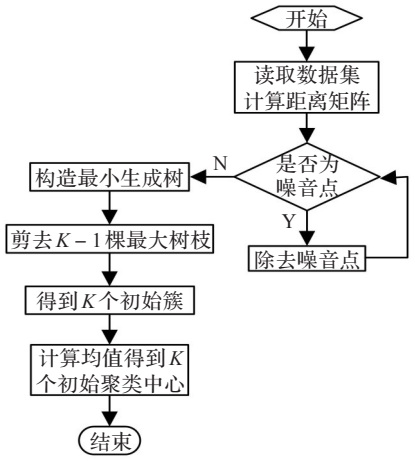

图2　TD算法流程图

$c[i](1\leqslant i\leqslant K)$,完成$K$个数据中心的选取。

最小生成树算法:

Procedure genMST(U,$\boldsymbol{D}$)

(1)T=null;n是U中对象数据点数

(2)Flag[]数组表示n个对象点的访问状态,置为false。Flag[0]=true

(3)Repeat

(4)　搜索出与所有已访问节点连接的权值最小的一条边$dist[i,j]$;

(5)　$dist[i,j]$加入树中;

(6)　将找出的新节点访问状态标记为true;

(7)Until 搜索出$n-1$条边;

(8)Return T

2　基于TD的改进*K*-means算法

K-means算法是一种被广泛使用的典型聚类算法,具有理论可靠,算法简洁和收敛速度快等优点,可以有效处理大数据集。该算法首先是随机选取K个数据点作为初始的聚类中心,每个数据点代表一个簇,根据剩下数据点到各聚类中心的距离将其分配到最近的簇,然后重新计算每个簇的平均值,作为新的聚类中心,以此迭代的方法不断改变聚类中心从而改变数据划分,直到准则函数收敛。随机的选取的初始聚类中心没有考虑到数据的实际分布,算法容易陷入局部最优。

将上述的TD算法得到的K个数据中心作为*K*-means算法的初始聚类中心,改进后的*K*-means算法记做TDKM算法。

输入:含有N个样本的数据集合U,聚类个数K

输出:K个簇的集合

(1)运用TA算法得到K个初始的聚类中心$c[i](1\leqslant i\leqslant K)$。

(2)Repeat

(3)根据簇中对象的均值,将每个对象指派到最相似的簇。

(4)更新簇均值,即计算每个簇中对象的均值。

(5)Until聚类中心不再发生变化,即准则函数收敛。

TDKM算法利用TD算法计算出K个初始聚类中心,这K个数据点尽可能分散,并且具有代表性,与数据的实际分布相一致。TDKM算法的时间复杂度是$O(m\times N\times K\times e)$,空间复杂度是$O((N+K)\times m)$,其中$N$表示所有样本数据个数,$K$是聚类数目,$e$是算法的迭代次数,$m$是样本的属性个数,通常$K<<N$和$e<<N$。

3　实验结果与分析

本文的实验环境为:Intel CPU,2 GB内存,500 GB硬盘,Windows XP 操作系统。利用Matlab进行编程验证算法的有效性,在UCI数据集Iris和Letter上进行了测试。Iris数据是国际公认的聚类算法比较好坏的典型数据。它包含150条带标签的4维样本点,分为三个类,每类各50条数据。Letter数据集上含有20 000条数据,每条数据包含16个属性,分为26类。分别在Iris数据集上用传统的*K*-means算法和TDKM算法进行10次测试,在Letter数据集上随机选择200,500,1 000,2 000,5 000,10 000,20 000共7个数据量的数据进行算法有效性验证实验。同时在Letter数据集上实验时加入了EM(期望最大化)算法作为对比算法。

3.1　实验结果

根据要求在Iris数据集上进行10次聚类实验,图3是传统的*K*-means算法和TDKM算法在数据集Iris中的实验迭代次数对比图,表1是TDKM算法和*K*-means算法在Iris上测试结果准确率情况。图4是传统的*K*-means算法,改进TDKM算法和EM算法在Letter数据集上进行实验的时间消耗情况,图5是上述三种算法在Letter数据集上进行聚类实验的准确率情况。图6,图7分别是TDKM算法和KM算法在Iris数据集上的聚类结果图。

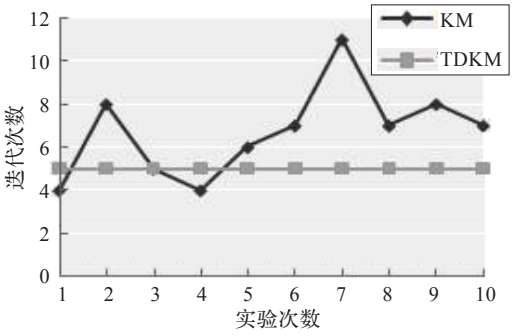

图3　Iris上的迭代次数

表1　算法的在Iris上测试结果

序号	Iris的准确率/(%)
1	89.33
2	88.67
3	56.00
4	89.33
5	56.67
6	88.67
7	51.33
8	56.00
9	56.00
10	89.33
KM算法平均值	72.13
TDKM算法	89.33

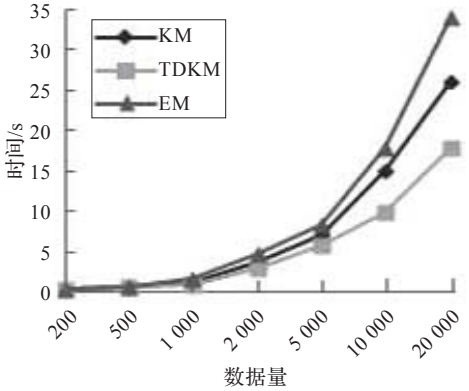

图4　Letter上聚类时间花费

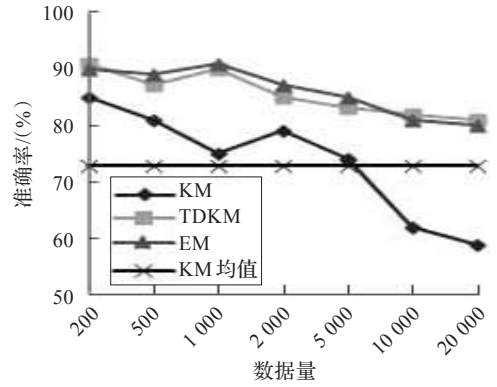

图5　Letter上聚类准确率

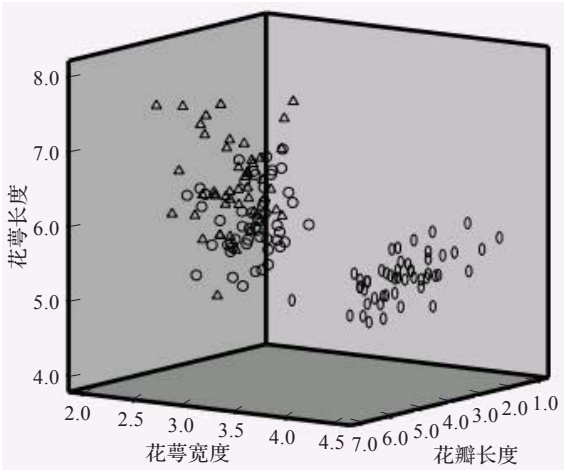

图6　TDKM算法在Iris聚类图

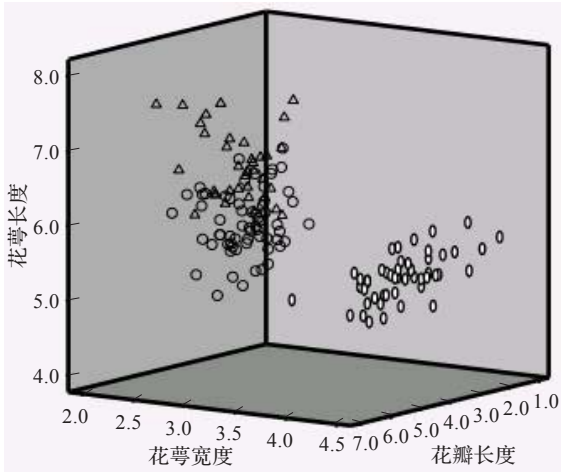

图7　KM算法在Iris聚类图

3.2　实验分析

传统*K*-means算法随机选取初始聚类中心，选择的初始聚类中心不同，算法迭代次数不同，得到的聚类结果也不同。如表1所示，传统*K*-means算法的准确率波动范围很大，Iris数据集上的平均值是72.13%，低于TDKM算法在测试数据集上的准确率。如果选取的初始聚类中心离实际的结果簇的中心比较远，将不利于目标函数的收敛。在Iris上的第7次测试结果表明，由于随机的选取的聚类中心与最终收敛中心较远，降低了聚类的准确率，增加了聚类过程中迭代的次数。图3可知，传统的*K*-means算法在实验中迭代次数上下波动范围大且迭代平均值大于TDKM算法的迭代次数。由于初始聚类中心的随机选取，没有依据数据的实际分布情况，造成算法迭代次数的增加，导致了传统的*K*-means算法的不稳定性，在实际运用中往往达不到预想的聚类效果。改进的TDKM算法在Iris数据集上的准确率为89.33%。较传统的*K*-means算法，准确率都有较大的提高。因为通过TD算法预处理得到了与数据的实际分布情况相一致的初始聚类中心，加速了聚类过程，收敛速度快，得到的聚类结果稳定。为了增加Letter数据集上的聚类实验的对比性，实验中加入了EM聚类算法。聚类效率方面由图4可以看出，当数据量较少时，三个算法的时间花费差距较少，随着测试数据量的增加，TDKM算法较传统*K*-means算法和EM算法有较大的提高；且随着数据量的增加，聚类效率的提高更加明显。在聚类准确率方面，TDKM算法和EM算法在各个数据量中的准确率都达到了80%以上，而*K*-means算法在数据量较小时准确率较高，但是随着数据量的增加，聚类的准确率呈现下降的趋势。当数据量达到20 000条时，*K*-means算法的聚类准确率只有55%左右。

综上所述，TDKM算法在传统*K*-means算法中增加了TD预处理算法。TD算法根据数据的实际分布情况动态选择初始聚类中心，得到的初始聚类中心与最终的收敛中心相一致。聚类过程中减少了迭代次数，提高了聚类准确率和聚类效率，增强了算法的稳定性。

为了更加直观地观察聚类结果，给出了在两种数据集上传统的*K*-means算法和TDKM算法的聚类结果图，如图6，图7所示。Iris数据集中第一类数据与其他数据离得较远，第二类数据和第三类数据由部分交叠。在TDKM算法的聚类结果图中可以看到，3个聚类簇的边界更加分明，聚类结果比较好。

4　结束语

本文根据传统的*K*-means算法中初始聚类中心的随机选取的缺陷提出了一种改进的算法。改进算法利用数据的分布动态选取初始聚类中心，找出数据对象中分布比较密集的区域，使得初始的聚类中心更具有代表性。由于初始的聚类中心接近结果聚类簇的中心，加速了聚类过程，因此收敛速度快。同时克服了孤立数据和噪音数据点对聚类结果带来的影响。实验结果证实改进后的算法能够得到较高且稳定的准确率，更适用于对实际数据的聚类。

参考文献：

[1] Treshansky A，McGraw R.An overview of clustering algorithms[C]//Proceedings of SPIE，The International Society for Optical Engineering，2001，4367：41-51.

[2] Clausi D A.*K*-means Iterative Fisher（KIF）unsupervised clustering algorithm applied to image texture segmentation[J]. Pattern Recognition，2002，35：1959-1972.

（下转192页）

5 结束语

行为序列是一种特殊的时序信号,每类行为往往包含若干帧典型样本。三维人体关节点序列是一种视角无关的表征行为序列姿势特征的有效方法之一,因此本文实验选择在三维人体关节点序列的数据集上进行。对于基于视频的行为识别,如果采用姿势序列或者包含姿势的特征序列(如前景序列)表征视频序列中的行为,则Adaboost-EHMM算法同样适用。

本文提出的AdaBoost-EHMM算法是一种典型样本选择和分类器学习同时进行的封装式方法:利用AdaBoost将典型样本选择出来作为HMM观测概率模型的均值。与之前提出的EHMM前向选择方法[6]相比,这种算法选择典型样本的评价准则不是分类器的平均识别率,而是侧重于易于混淆的样本。AdaBoost-EHMM算法还通过多级分类器的融合保证算法的收敛和识别率的提高。利用三维人体关节点序列包含的多尺度上的行为特性,本文提出了融合多特征的行为序列识别方法。全局运动特征序列、全身结构特征序列、手臂结构特征序列等从不同尺度、不同角度描述了行为的各种特性。对于全局运动特征序列,采用了传统的HMM方法进行建模;对于包含关键姿势的结构特征,利用本文提出的AdaBoost-EHMM算法进行建模。实验证明,融合多种特征的分类器有助于描述复杂行为各个层次的特性,提高了行为识别算法的性能。

参考文献:

[1] Rabiner L R.A tutorial on hidden Markov model and selected applications in speech recognition[J].Proceedings of the IEEE,1989,77(2):257-286.

[2] Rabiner L R,Juang B H.An Introduction to hidden Markov models[J].IEEE ASSP Magazine,1986,3(1):4-16.

[3] Yamato J,Ohya J,Ishii K.Recognizing human action in time-sequential images using hidden Markov model[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition,1992:379-385.

[4] Brand M,Oliver N,Pentland A.Coupled hidden Markov models for complex action recognition[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition,1997:994-999.

[5] Elgammal A M,Shet V D,Yacoob Y,et al.Learning dynamics for exemplar-based gesture recognition[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition,2003:571-578.

[6] Weinland D,Boyer E,Ronfard R.Action recognition from arbitrary views using 3D exemplars[C]//Proceedings of IEEE International Conference on Computer Vision,2007:1-7.

[7] 边肇祺,张学工.模式识别[M].2版.北京:清华大学出版社,2004.

[8] Viola P,Jones M.Rapid objects detection using a boosted cascade of simple features[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition,2001:511-518.

[9] Gu Junxia,Ding Xiaoqing,Wang shengjin,et al.Action and gait recognition from recovered 3-D human joints[J].IEEE Trans on Systems,Man,and Cybernetics,Part B,2010,40(4):1021-1033.

[10] Weinland D.The institut national de recherche en informatique et automatique xmas motion acquisition sequences data[EB/OL].[2006-12-01].https://charibdis.inrialpes.fr.

[11] Lv Fengjun,Nevatis R.Recognition and segmentation of 3-D human action using HMM and multi-class AdaBoost[C]//Proceedings of European Conference on Computer Vision,2006:359-372.

[12] Weinland D,Ronfard R,Boyer E.Free viewpoint action recognition using motion history volumes[J].Computer Vision and Image Understanding,2006,104(2/3):249-257.

(上接185页)

[3] Bezdek J C,Pal N R.Some new indexes of cluster validity[J].IEEE Transactions on Systems,Man,and Cybernetics Part B:Cybernetics,1998,28(3):301-315.

[4] Ramze R M,Lelieveldt B P F,Reiber J H C.A new cluster validity indexes for the fuzzy *c*-mean[J].Pattern Recognition Letters,1998,19:237-246.

[5] 杨善林,李永森,湖笑旋,等.*K*-means算法中的*k*值优化问题研究[J].系统工程理论与实践,2006(2):97-101.

[6] 冯超.*K*-means聚类算法的研究[D].大连:大连理工大学,2007:25-29.

[7] Lai Yuxia,Liu Jianping.Optimization study on initial center of *K*-means algorithm[J].Computer Engineering and Application,2008,44(10):147-149.

[8] 汪中,刘贵全,陈恩红.一种优化初始中心点的*K*-means算法[J].模式识别与人工智能,2009,22(2):299-304.

[9] Mao Shangyang,Li Kenli.Research of optimal *K*-means initial clustering center[J].Computer Engineering and Application,2007,43(22):179-181.

[10] 孙雪,李昆仑,胡夕坤,等.基于半监督*K*-means的*k*值全局寻优算法[J].北京交通大学学报,2009,33(6):106-109.

[11] 汪军,王传玉,周鸣争.半监督的改进*K*-均值聚类算法[J].计算机工程与应用,2009,45(28):137-139.

[12] Yang Shuzhong,Luo Siwei.A novel algorithm for initializing cluster[C]//Proceedings of the 4th International Conference on Machine Learning and Cybernetics,2005:5579-5583.

[13] 钱线,黄萱菁,吴立德.初始化*K*-means谱算法[J].自动化学报,2007,33(4):342-346.

[14] 王玲,薄列峰,焦李成.密度敏感的谱聚类[J].电子学报,2007,35(8):1577-1581.

[15] Leng Mingwei,Tang Haitao,Chen Xiaoyun.An efficient initialization scheme for *K*-means clustering[C]//8th ACIS International Conference on Software Engineering,Artificial Intelligence,Networking,and Parallel/Distributed Computing,2007:815-820.