基于 UPGMA 的优化初始中心 K-means 算法研究

张 锐,王义武,朱啸龙,殷 俊,韩 晨,杨余旺
(南京理工大学 计算机科学与工程学院,江苏 南京 210000)

摘 要: 为了弥补传统 K-means 算法聚类效果严重依赖于初始聚类中心这一不足,提出了 OICC K-means 算法。将不加权算术平均组对法(UPGMA)进行改进,通过该算法将密集区域的数据合并得到可以反映数据分布的若干数据点,再由最大最小距离算法从中选出彼此相距较远的点,作为传统 K-means 算法的初始聚类中心,从而使 K-means 算法有一个可以反映数据分布特征的输入。在典型数据集上进行的实验发现,相较于传统 K-means 算法,OICC K-means 算法拥有更强的聚类能力,在准确率、召回率和 F-测量值方面均有明显提高。在 OICC K-means 算法的前两个阶段(即 UPGMA 算法和最大最小距离算法)产生了较理想的初始聚类中心,这些中心点选自于数据密集的区域,因此避免了噪声数据、边缘数据带来的不良影响,使得 K-means 算法没有陷入局部最优解而达到了整体良好的聚类效果,同时聚类中心的个数在算法中自动确定而不需要手动设置。

关键词: 聚类; 初始中心; 不加权算术平均组对法; 最大最小距离算法; K-means 算法
中图分类号: TP301.6 **文献标识码:** A **文章编号:** 1673-629X(2018)02-0050-04
doi: 10.3969/j.issn.1673-629X.2018.02.012

Research on K-means Algorithm for Optimizing Initial Center Based on UPGMA

ZHANG Rui, WANG Yi-wu, ZHU Xiao-long, YIN Jun, HAN Chen, YANG Yu-wang
(School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210000, China)

Abstract: In order to compensate for deficiency that the traditional K-means algorithm depends heavily on initial clustering centers in clustering effect, we propose the OICC K-means algorithm. By improved UPGMA, the data in dense area is combined to obtain a number of data points that can reflect the distribution of the data, from which the distant one from each other is chosen by the maximum and minimum distance algorithm as the initial clustering center of traditional K-means algorithm, so that it has an input that reflects the characteristics of the data distribution. It can be found in experiment on the typical data set that OICC K-means algorithm, with a stronger clustering, compared with the traditional K-means algorithm, is improved in accuracy, recall and F-measure obviously. The first two stages of the OICC K-means algorithm (the UPGMA and the maximum and minimum distance) produces ideal initial clustering centers which are selected from the data-intensive regions, thus avoiding the adverse effects caused by noise data and edge data. Therefore, the K-means algorithm does not fall into the local optimal solution and achieves the overall good clustering effect, and the number of clustering centers is automatically determined without manual setting.

Key words: clustering; initial centers; UPGMA; maximum and minimum distance algorithm; K-means algorithm

0 引 言

K-means 是一种经典的基于划分的聚类分析方法,具有简单、高效的特点,在众多领域得到了广泛应用。通过计算每个数据对象与 k 个聚类中心的距离,将数据对象划分到距离它最近的一个类,然后更新每个类的中心,这个过程反复迭代直到收敛,输出聚类结果。传统的 K-means 算法中,k 个初始聚类中心是在数据对象集中随机选取的,因此迭代过程从不同的初始聚类中心出发,得到的聚类结果也不同,并且聚类的迭代过程容易产生局部最优解[1]。同时,这种聚类结果波动在工程应用中也会带来许多技术问题[2]。

为了选择优化的初始聚类中心,已有研究主要从密度优化和距离估计两方面对 K-means 算法加以改进。文献[3]是一种典型的密度优化聚类中心选择算

收稿日期: 2017-02-23 修回日期: 2017-06-28 网络出版时间: 2017-10-19
基金项目: 国家自然科学基金(61640020);江苏省科技支撑计划(BE2012386,BE2011342);江苏省农业自主创新项目(CX(13)3054、CX(16)1006);江苏省重点研发计划(BE2016368-1)
作者简介: 张 锐(1977-),男,硕士生,研究方向为聚类分析;杨余旺,教授,研究方向为大数据系统、计算机网络、网络编码。
网络出版地址: http://kns.cnki.net/kcms/detail/61.1450.TP.20171019.1626.066.html

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

法,通过密度函数法求得样本数据空间的多个聚类中心,结合类的合并运算,避免局部最优问题。类似的,文献[4]提出按照数据集的数学分布动态地寻找 k 个数据点作为初始聚类中心。而进一步的理论和实验表明,在密度优化方法的基础上,采用距离估计(如最大最小距离法)可提升算法的收敛速度、准确率[5-6]。

文中引入了不加权算术平均组对法(UPGMA)[7]和最大最小距离算法,通过这两种算法的优化和筛选,选取优化的数据中心,为 K-means 算法提供准确反映数据分布的聚类中心,解决聚类中心高度依赖的问题。

1　优化候选中心 K-means 算法

在理想的聚类算法中,各聚类中心应该分散地取自密集区域,根据这一思想,提出优化候选中心(OICC) K-means 算法。该算法可以将改进的 UPGMA 算法和最大最小距离算法相融合,充分发挥各自的优点,得到经过优化的可以反映密集区数据分布的初始候选中心。如图 1 所示,这些点可以作为密集区域的代表,再把这些聚类中心应用到 K-means 算法中,在整体上提高聚类效果。

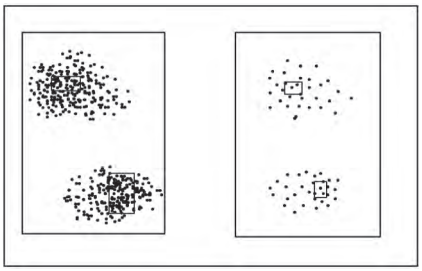


图 1　数据与候选中心分布

1.1　基本思想

定义 1: 数据点 X_i 与 X_j 之间的距离定义为:

$$\mathrm{Dist}(X_i,X_j)=\mathrm{sqrt}\left(\sum_{k=1}^{m}(X_{ik}-X_{jk})^2\right) \tag{1}$$

其中,X_i、X_j 表示数据集中 m 维的数据对象。

定义 2: 聚类 C 中心点的坐标为 $C=(C_1,C_2,\cdots,C_j)$,第 j 维的数据定义为:

$$C_j=\frac{\sum_{i=1}^{m}X_{ij}}{n} \tag{2}$$

其中,n 为子类包含的数据个数; X_i 为子类内的数据。

最大最小距离[8]算法其基本思想是使作为聚类中心的候选点尽可能相隔较远,得到经过优化的候选中心,更好地刻画数据集的整体分布。避免了在选取初始候选点时选点存在过近的情况,尽可能地在所有数据聚集区都能选到候选点,有效避免了初始中心的过于集中,很大程度上提高了聚类算法的效果[9-10]。但在这个过程中可能把噪声数据、边缘数据也加入到初始聚类中心内。同时,该算法需要两次扫描数据库,第一次是找到每个类到已有聚类中心的最短距离,第二次是扫描数据库得到最大的最小距离即 $\mathrm{Max}(\mathrm{Min}(D_t))$。算法完成时,时间复杂度为 $O(nk)$(k 为聚类中心的个数,n 为数据对象的个数)。可见,当数据规模很大时,计算量过大。K-means 算法是建立在数据间距离之上的,即它的类划分标准为数据间的相互距离,数据距离近说明数据间相似度大,就可把它们划分到同一个类中[11-12]。

定义 3: 聚类和数据集中剩余数据的最大最小距离定义为:

$$D_{\max}=\mathrm{Max}(d) \tag{3}$$

其中,d 为每个聚类与数据集中剩余各数据距离的最小值组成的集合。

定义 4: d 为每个聚类与数据集中剩余各数据距离的最小值,即

$$d=\mathrm{Min}\left(\sum_{k=1}^{m}(X_{ik}-X_{jk})^2\right) \tag{4}$$

其中,X_i 为聚类中心; X_j 为数据集中剩余的数据; m 数据的维度。

定义 5: 判断是否将初始候选中心选为优化后的候选中心。

$$\mathrm{Max}(\mathrm{Min}(D_t))>\theta\|v_1-v_2\| \tag{5}$$

其中,v_1、v_2 为最先成为优化后的候选中心的点; θ 为参数,常取 0.5。

定义 6: K-means 算法进行聚类的准则函数为误差平方和准则函数。

$$J_c=\sum_{i=1}^{k}\sum_{p\in C_i}(\|P-M_i\|)^2 \tag{6}$$

其中,M_i 为类 C_i 中所有数据的均值; p 为类 C_i 中的每个数据; J_c 为样本和聚类中心的函数。在已知样本集时,J_c 的值取决于 M_i。J_c 描述了 n 个样本数据得出 k 个分类产生的总的误差平方和。可以看出,J_c 越大,表明误差越大,效果越差。所以应力求 J_c 最小,从而得出最好的聚类效果。

初始随机给定 k 个簇中心,将数据划分到距自己最近的簇中,然后重新计算每个簇的中心,一直迭代,直到前后两次得出的簇的中心不再变化或者满足一定要求为止。每次迭代后,需要检验数据分类是否正确,如果错误,就要继续划分聚类,重新计算簇中心,进行下一次迭代。但 K-means 算法结束时找到的解常常为局部最优而非全局最优,如图 2 所示。

图中,p_1、p_2、p_3 的初值不同,目标函数分别顺着误差平方和准则函数逐渐减小的方向搜索,直到找到各自对应的最小值 A、B、C。其中,A、B 为局部极

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

小值，C 为全局最小值，但算法结束时经常只能收敛局部极小值[13-14]。

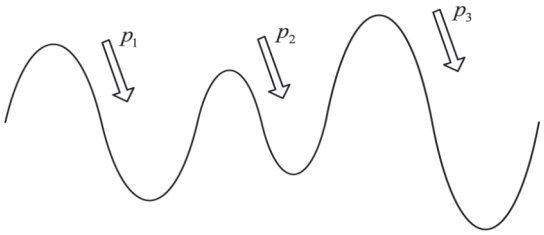


图 2 局部极小值和全局极小值

1.2 改进 UPGMA 算法

在 UPGMA 改进算法中，将每个数据都看作一个类，获得每个类之间的距离，将相互间距离最小的数据加入到同一个类中形成一个新类，重复此步骤，当满足算法停止的条件或者不再产生新类时，算法结束。改进的 UPGMA 运算过程如下所述：

算法 1：改进 UPGMA 算法。

Input：需要进行聚类的数据；聚类内数据数目占数据总数的百分比 $m\%$；序列 Q 的前 $p\%$；

Output：初始候选中心。

Step1：把所有数据都看成一个独立的类。

Step2：计算出任意两个类的距离，合并相距最近的两个类，同时判断剩下数据的总数是否大于等于总数的 $m\%$，若否，则转 Step4。

Step3：迭代次数 $i = 0$

Step4：do

Step5：$t = t + 1$；

Step6：for $j = i + 1, \cdots$，maxcluster do

当子类 i 和子类 j 内部的数据个数少于等于总数的 $m\%$ 时，计算两个类之间的距离，并加入到距离矩阵中。

Step7：endfor

Step8：在距离矩阵中找到距离最小的对应的两个类，将它们合并为新类，同时加入到序列 Q 的结尾，转至 Step2。

Step9：while $t >$maxcluster

Step10：将序列 Q 前 $p\%$ 个子类作为候选子类，通过计算得出它们的中心点，作为初始候选中心。

其中，m 为筛选条件，只有当类中数据的数目大于数据总数的 $m\%$，才将此类加入到序列 Q 中；取序列 Q 的前 $q\%$ 计算它们的中心，作为初始候选中心；maxcluster 为聚类的总数，测量两个数据之间的距离采用欧氏距离。

在这个迭代过程中可以发现，位于数据密集区的数据最先聚集在一起。这一过程选择出来的初始候选中心是经过优化的，能够很好地反映数据的分布情况，提高了精确性。

2 算法流程

将最大最小距离算法的输出作为 K-means 算法的输入，使得聚类中心点能够充分反映数据分布情况，很大程度上弥补了 K-means 算法在初始聚类中心选择上的不足之处。优化候选中心(OICC) K-means 算法的框架大致分为三个阶段：第一阶段是产生候选中心及其优化，获得最佳的候选中心；第二阶段是算法执行，主要是 K-means 算法在已有初始聚类中心上进行聚类操作；第三阶段主要是实验和评估结果，验证 OICC K-means 的实际效果。

算法流程如图 3 所示。

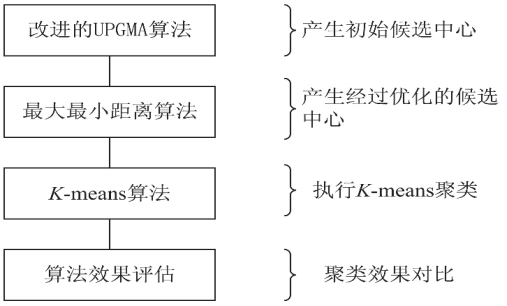


图 3 OICC K-means 算法框架

OICCK-means 算法有效解决了如何选择候选中心的问题，既保证了聚类中心来自于数据密集区域，减小噪声数据和边缘数据的影响，同时也确保了被选中的聚类中心之间有足够远的距离，真实地反映了整体数据的分布，使得算法更加稳定和有效。

具体操作过程如下所述：

算法 2：OICC K-means 算法。

Input：需要进行聚类的数据；聚类内数据数目占数据总数的百分比 $m\%$；序列 Q 的前 $P\%$，θ；

Output：经过聚类的数据。

Step1：UPGMA 算法：得到初始聚类候选中心；

Step2：最大最小距离算法：得到优化的初始聚类中心；

Step3：K-means 算法迭代；

Step4：结果评定。

3 实验结果与分析

3.1 实验描述

对传统的 K-means 算法和提出的 OICC K-means 算法在 UCI 的三个标准数据集上进行聚类效果对比。使用 F-measure 来衡量算法效果，包括准确率(precision)和召回率(recall)。

准确率定义为：

$$P(i,j) = \text{precision}(i,j) = N_{ij}/N_i \tag{7}$$

召回率定义为：

$$R(i,j) = \text{recall}(i,j) = N_{ij}/N_j \tag{8}$$

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

其中，N_{ij} 为聚类 j 中分类 i 的数目; N_i 为分类 i 中所有的对象数; N_j 为聚类 j 中所有的对象数。

F-measure 表示为:

$$F(i) = 2P/(P + R) \tag{9}$$

在输出的结果中，对分类 i 来说，F-measure 高的聚类即对应分类 i。在实验中，文中运用 UCI 中的数据集作为测试集，由于库中的数据有明确的类别，可以直观准确地测试算法的聚类效果。选择其中的 Iris、Tae 和 Hayes-roth 三个库(见表 1)，比较了传统 K-means 算法和 OICC K-means 算法在准确率、召回率和 F-measure 上的实际数据。

表 1　测试数据集

数据集	数据大小	分类数	属性数
Iris	150	3	4
Tae	151	3	5
Hayes-roth	132	3	5

3.2　实验结果

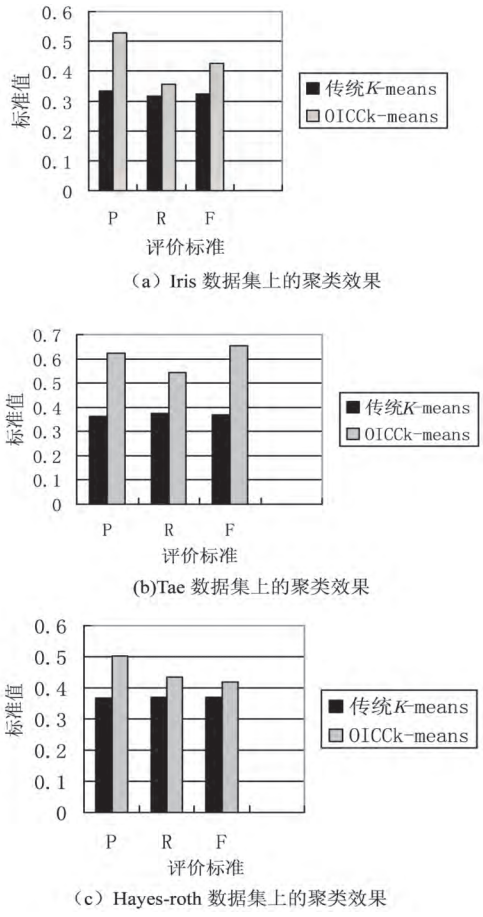
Iris、Tae、Hayes-roth 在 OICC K-means 算法中 (m,q,θ) 分别取(0.08,0.8,0.5)、(0.09,0.8,0.5)、(0.06,0.8,0.5)，传统 K-means 算法中 k 的取值均取 3。通过对这些结果的对比来说明 OICC K-means 算法在聚类效果方面的有效性。评价标准的值越大，说明聚类效果越好。对比实验结果如表 2 所示。

表 2　算法对比

数据集	评价标准	传统 K-means 算法	OICC K-means 算法
Iris	准确率	0.333 33	0.526 67
	召回率	0.316 60	0.356 67
	F-measure	0.324 75	0.425 31
Tae	准确率	0.362 00	0.624 40
	召回率	0.374 63	0.542 64
	F-measure	0.368 20	0.654 48
Hayes-roth	准确率	0.367 65	0.501 42
	召回率	0.370 36	0.435 37
	F-measure	0.369 00	0.417 98

通过图 4 可以更加直观地看出两个算法在效果上的差别。

由图 4 可见，OICC 算法相比于传统 K-means，在 Iris 数据集上，准确率提高了 58.0%，召回率提高了 12.7%，F-measure 提高了 31.0%; 在 Tae 数据集上，准确率提高了 72.5%，召回率提高了 44.9%，F-measure 提高了 77.7%; 在 Hayes-roth 数据集上，准确率提高了 36.4%，召回率提高了 17.8%，F-measure 提高了 13.3%。原因在于 OICC K-means 算法运用改进 UPGMA 算法和最大最小距离算法得到经过优化的聚类候选中心，这些聚类中心可以真实有效地代表实际数据的分布[13]，大大增强了算法的聚类效果，同时算法的计算量明显减少。该算法不仅可以自动确定初始候选中心 k 的值，也避免了噪声数据和边缘数据对实验的影响，比传统 K-means 算法在聚类方面具有更高的准确性和实用性。



(a) Iris 数据集上的聚类效果

(b)Tae 数据集上的聚类效果

(c) Hayes-roth 数据集上的聚类效果

图 4　数据集实验结果

4　结束语

为了弥补传统 K-means 算法聚类效果严重依赖于初始聚类中心这一不足，提出了 OICC K-means 算法。实验结果表明，该算法在聚类效果方面要好于传统 K-means 算法，准确率、召回率、F-measure 三项指标都有明显提高，并且对不同的数据集都有比较好的效果。OICC 算法是根据真实数据分布情况，通过对比数据间的距离得出较理想的候选中心，这些数据中心在一定程度上成为了数据密集区的代表，降低了噪声数据、边缘数据和不恰当的初始聚类中心对实验的影响，使得 K-means 算法有一个较理想的聚类中心，具有较高的可行性。

参考文献:

[1]　谢娟英，高红超.基于统计相关性与 K-means 的区分基因子集选择算法 [J] .软件学报，2014，25(9)：2050-2075.

(下转第 58 页)

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

[2] BRANCHINI R M,ARMENTANO V A,LOKKETANGEN A.Adaptive granular local search heuristic for a dynamic vehicle routing problem [J].Computers & Operations Research,2009,36(11):2955-2968.
[3] CREPUT J C,HAJJAM A,KOUKAM A,et al.Self-organizing maps in population based metaheuristic to the dynamic vehicle routing problem [J].Journal of Combinatorial Optimization,2012,24(4):437-458.
[4] KHOUADJIA M R,SARASOLA B,ALBA E,et al.A comparative study between dynamic adapted PSO and VNS for the vehicle routing problem with dynamic requests [J].Applied Soft Computing,2012,12(4):1426-1439.
[5] ELHASSANIA M,JAOUAD B,AHMED E A.A new hybrid algorithm to solve the vehicle routing problem in the dynamic environment [J].International Journal of Soft Computing,2013,8(5):327-334.
[6] HOUSROUM H,HSU T,DUPAS R,et al.A hybrid GA approach for solving the dynamic vehicle routing problem with time windows [C]//Information and communication technologies.[s.l.]:IEEE,2006:787-792.
[7] HONG L.An improved LNS algorithm for real-time vehicle routing problem with time windows [J].Computers & Operations Research,2012,39(2):151-163.
[8] ATTANASIO A,CORDEAU J F,GHIANI G,et al.Parallel Tabu search heuristics for the dynamic multi-vehicle dial-a-ride problem [J].Parallel Computing,2004,30(3):377-387.
[9] GENDREAU M,GUERTIN F,POTVIN J Y,et al.Neighborhood search heuristics for a dynamic vehicle dispatching problem with pick-ups and deliveries [J].Transportation Research Part C:Emerging Technologies,2006,14(3):157-174.
[10] 张 迅,刘海东,李 丹,等.基于遗传算法的快递配送车辆路径问题研究 [J].物流技术,2013,32(3):263-267.
[11] 王振锋,王 旭,葛显龙.基于遗传算法的不同约束条件车辆调度问题研究 [J].计算机应用研究,2010,27(10):3673-3675.
[12] 袁麟博,章卫国,李广文.一种基于遗传算法-模式搜索法的无人机路径规划 [J].弹箭与制导学报,2009,29(3):279-282.
[13] 邝航宇,金 晶,苏 勇.自适应遗传算法交叉变异算子的改进 [J].计算机工程与应用,2006,42(12):93-96.
[14] 卢月品,赵 阳,孟跃强,等.基于改进遗传算法的狭窄空间路径规划 [J].计算机应用研究,2015,32(2):413-418.
[15] 雷伟军,程筱胜,戴 宁,等.基于改进遗传算法的多模型加工路径规划 [J].机械工程学报,2014,50(11):153-161.
[16] 庄 健,杨清宇,杜海峰,等.一种高效的复杂系统遗传算法 [J].软件学报,2010,21(11):2790-2801.
[17] 薛 明,许德刚.基于云网格集成调度的防拥堵车辆路径规划算法 [J].计算机科学,2015,42(7):295-299.
[18] 侯占亭.基于分解和决策空间相似性度量的进化多目标车辆路径规划算法研究 [D].西安:西安电子科技大学,2014.
[19] 于 锐,曹介南,朱培栋.车辆运输路径规划问题研究 [J].计算机技术与发展,2011,21(1):5-8.

(上接第53页)

[2] JAIN A K.Data clustering:50 years beyond K-means [J].Pattern Recognition Letters,2010,31(8):651-666.
[3] SHI N,LIU X,GUAN Y.Research on k-means clustering algorithm:an improved k-means clustering algorithm [C]//Third international symposium on intelligent information technology and security informatics.[s.l.]:IEEE,2010:63-67.
[4] 仝雪姣,孟凡荣,王志晓.对k-means初始聚类中心的优化 [J].计算机工程与设计,2011,32(8):2721-2723.
[5] 张文明,吴 江,袁小蛟.基于密度和最近邻的k-means文本聚类算法 [J].计算机应用,2010,30(7):1933-1935.
[6] 熊忠阳,陈若田,张玉芳.一种有效的k-means聚类中心初始化方法 [J].计算机应用研究,2011,28(11):4188-4190.
[7] MILLIGAN G W.Cluster analysis for researchers [J].Journal of Marketing Research,1985,22(2):224-225.
[8] 赖玉霞,刘建平.K-means算法的初始聚类中心的优化 [J].计算机工程与应用,2008,44(10):147-149.
[9] 王赛芳,戴 芳,王万斌,等.基于初始聚类中心优化的K-均值算法 [J].计算机工程与科学,2010,32(10):105-107.
[10] SAMBASIVAM S,THEODOSOPOULOS N.Advanced data clustering methods of mining Web documents [J].Issues in Informing Science and Information Technology,2006,8(3):563-579.
[11] 傅德胜,周 辰.基于密度的改进K均值算法及实现 [J].计算机应用,2011,31(2):432-434.
[12] DEHARIYA V K,SHRIVASTAVA S K,JAIN R C.Clustering of image data set using k-means and fuzzy k-means algorithms [C]//International conference on computational intelligence and communication networks.[s.l.]:IEEE Computer Society,2010:386-391.
[13] ERISOGLU M,CALIS N,SAKALLIOGLU S.A new algorithm for initial cluster centers in k-means algorithm [J].Pattern Recognition Letters,2011,32(14):1701-1705.
[14] 张宜浩,金 澎,孙 锐.基于改进k-means算法的中文词义归纳 [J].计算机应用,2012,32(5):1332-1334.
[15] 唐贤伦,仇国庆,庄 陵.一种面向非规则非致密空间分布数据的聚类方法 [J].计算机科学,2009,36(3):167-169.

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net