文章编号: 1001- 2486(2009) 02- 0070- 06

基于优化初始聚类中心 K-Means 算法的跳频信号分选*

陈利虎[1], 张尔扬[1], 沈荣骏[2]
(1. 国防科技大学 电子科学与工程学院, 湖南 长沙 410073; 2. 解放军总装备部 科学技术委员会, 北京 100080)

摘 要: 提出了一种优化初始聚类中心的方法。方法通过搜索参数统计直方图峰值预估类数目, 并根据峰值位置确定聚类中心大概位置。由于优化的初始类心与实际类心相隔不远, 聚类迭代次数大为减少。与传统的优化聚类中心方法相比, 本方法计算量更少。最后将改进 K-Means 聚类算法应用于跳频信号分选, 仿真结果表明, 分选效果良好。

关键词: 聚类; K-Means 算法; 跳频; 信号分选
中图分类号: TN97 文献标识码: A

The Sorting of Frequency Hopping Signals Based on K-Means Algorithm with Optimal Initial Clustering Centers

CHEN Li-hu[1], ZHANG Er-yang[1], SHEN Rong-jun[2]
(1. College of Electronic Science and Engineering, National Univ. of Defense Technology, Changsha 410073, China;
2. General Equipment Department of PLA, Beijing 100080, China)

Abstract: A new method is proposed to select optimal initial cluster centers. By searching parameters' histogram peak values, the number of cluster centers can be estimated, and these optimal initial cluster centers are selected in the columns or cells where the histogram peaks exist. Because these optimal initial cluster centers are near to real cluster centers, the iterations of clustering are reduced efficiently. Theoretical analysis demonstrates that the compute complexity of new method is lower than some conventional techniques. The improved K-Means algorithm is applied to sort frequency-hopping signals, and the simulation results demonstrate that the algorithm is effective.

Key words: clustering; K-Means algorithm; frequency hopping; signal sorting

跳频通信具有良好的抗干扰性、低截获概率及组网能力, 在军事通信中得到了广泛的应用, 也向通信对抗提出了严峻的挑战。开展对跳频信号侦察如盲检测、盲分选等技术的研究, 对于当前军事通信对抗具有重大意义。

对跳频信号的侦察通常分为 4 个步骤: 信号截获、参数估计、信号分选以及分选后信号处理(解跳、解调等)。本文着重研究基于聚类分析的跳频信号分选技术, 即在已获得跳频信号每跳(hop) 的各参数(载频 RF、跳周期 T_h、跳时 T_0、功率 PA、到达方向 DOA) 的基础上, 利用这些参数将混叠在一起的 hop 进行分类, 要求分选后的一个类代表一个电台。设待分选 hop 集为 $X = [X_1, X_2, \cdots, X_N]$, 其中 $X_n = [RF, T_h, T_0, PA, DOA]'$, $n = 1, 2, \cdots, N$。

1 传统的 K-Means 聚类方法

K-Means 是理论上可靠、应用上高效的聚类算法, 文献[1] 运用随机过程的方法给出了 K-Means 聚类算法的收敛性证明。K-Means 算法的主要思想是试图对 N 个待分选对象给出 K 个划分($K \leqslant N$), 其中每个划分代表一个类[2]。具体步骤如表 1。

* 收稿日期: 2008- 09- 19
作者简介: 陈利虎(1980—), 男, 博士生。

表 1 传统 K-Means 算法步骤

Tab. 1 Process of conventional K-means algorithm

1. 任选 K 个待分选对象作为初始聚类中心: $z_1^{(0)}, z_2^{(0)}, \cdots, z_K^{(0)}$, 迭代次数 $s=0$。
2. 将待分选对象集 X 中的各 hop 逐个按最小距离原则分划给 K 类中的某一类, 即如果 $d_{ij}^{(s)}=\min[d_{ij}^{(s)}]$, $i=1,2,\cdots,N$, 则判 $X_i\in\omega_j^{(s)}$。式中 $d_{ij}^{(s)}$ 表示 X_i 和类 $\omega_j^{(s)}$ 的中心 $z_j^{(s)}$ 的距离, 上角标表示迭代次数。迭代一次后产生新的聚类 $\omega_j^{(s+1)}, j=1,2,\cdots,K$。
3. 计算重新分类后的各类心: $z_j^{(s+1)}=\frac{1}{n_j^{(s+1)}}\sum_{X_i\in\omega_j^{(s+1)}}X_i, j=1,2,\cdots,K$, 式中 $n_j^{(s+1)}$ 为类 $\omega_j^{(s+1)}$ 中所含对象的个数。
4. 如果 $z_j^{(s+1)}=z_j^{(s)}$, 结束迭代; 否则 $s=s+1$, 转至第 2 步。

K-Means 算法是基于局部最优的聚类算法, 通常对初始聚类中心的选取非常敏感。对于给定的聚类数目 K, 不同的初始聚类中心很可能会导致截然不同的聚类结果[3]。因此选取合适的初始聚类中心, 不仅能提高此类算法的精度, 而且可以减少算法收敛时所需要的迭代次数, 降低算法的运行时间。

2 初始聚类中心的优化选取和聚类数目确定

目前的聚类中心初始化方法大致分为 3 种: 任意样本法、密度评估法和距离优化法[4]。任意样本法从样本集合中任意选取 K 个样本作为初始聚类中心, 这样选取的聚类初始中心使得聚类结果进入全局最优的概率较小, 同时还容易出现死点问题。文献[5] 提出了一种称为爬山法的初始聚类方法, 该方法先在数据样本空间中构造一个网格, 然后根据待分选对象到每个网格的距离, 分别算出每个网格点的势函数, 对象越密集的地方势越高, 将最高的网格点作为第一个初始聚类中心, 继而重复此操作选出 K 个初始聚类中心。该算法的缺点是计算量随着维数指数级增长。文献[6–7] 改进了爬山法, 用密度函数代替势函数, 使计算量与对象维数无关。KR[8] 是一种密度评估的方法, 它通过两两距离的比较完成对密度的评估, 从输入样本中具有较高局部密度的区域中选取初始聚类中心。上面的算法选取聚类中心阶段均需计算对象两两之间距离, 算法复杂度为 $O(MN^2)$。文献[9] 提出了一种 psFCM 算法, 该算法第一步将原大量待分选对象分成多个网格, 求网格里面数据的质心, 再将各网格质心作为待分选数据进行第一步 K-Means 聚类, 得到初级的聚类中心; 第二步将第一步得到的聚类中心作为初始中心, 再进行所有数据的聚类。该方法需要计算质心, 并且需要两次聚类。文献[10] 同样将待分选对象分成多个网格, 计算各个网格的密度, 定义密度指针, 指针从密度较低的网格指向密度较高的网格, 继而利用密度指针找到初始聚类中心。

上面的算法在一定程度上优化了初始聚类中心, 减少了聚类迭代次数, 但都假设聚类数目 K 已知, 而在很多情况下 K 是未知的。本文提出了通过搜索参数统计直方图峰值的初始类心选取方法, 该方法能够预估聚类数目 K 值, 并且根据直方图峰值位置选取初始聚类中心。方法较为简单, 易于实现, 并且算法选取的中心与真实类心相隔不远, 能有效减少迭代次数。

2.1 距离测度定义

待分类对象 X_i, X_j 之间的加权欧氏距离定义如下:

$$d(X_i,X_j)=\|\boldsymbol{W}X_i-\boldsymbol{W}X_j\|=\left[\sum_{i=1}^{n}(X_i-X_j)'\boldsymbol{W}(X_i-X_j)\right]^{1/2} \tag{1}$$

其中加权矩阵 $\boldsymbol{W}=\mathrm{diag}(w_1,w_2,\cdots,w_M)$, diag 为对角阵, M 为对象的参数维数。权值反映对象中各参数对分选判断贡献大小。

2.2 搜寻直方图峰值确定初始聚类中心

直方图的传统定义为: 将一个矩阵 $\boldsymbol{X}_{(N\times M)}$ 的各元素值用直条描述。如果矩阵 $\boldsymbol{X}$ 的列数为 1, 其直方图为二维图像; 矩阵列数为 2, 其直方图为三维图像。将直方图推广到矩阵多列情况下, 即列数为 M ($M\geqslant 3$) 的直方图将在 $M+1$ 维空间显示。可知矩阵 $\boldsymbol{X}_{(N\times M)}$ 为二维空间, 其直方图表示即张成了 $M+1$

维空间。

待分选 hop 集 $\boldsymbol{X}$ 共有 N 个 hop, 每个 hop 参数为 M 维, 于是可以对 hop 集 $\boldsymbol{X}$ 进行直方图统计, 得到区间集合 $CELL=\{cell_{i_1 i_2 \cdots i_M}\}\big|_{i_1=1,2,\cdots,p_1;\cdots;i_M=1,2,\cdots,p_M}$, 区间对应的直条高度(即区间内所含对象数目)为 $Y=\{y_{i_1 i_2 \cdots i_M}\}\big|_{i_1=1,2,\cdots,p_1;\cdots;i_M=1,2,\cdots,p_M}$; 继而搜寻直方图的峰值数目和峰值位置, 峰值数目即为初始聚类中心数目, 峰值位置对应的区间质心作为聚类中心。步骤如表 2 所示。

表 2 基于直方图峰值确定初始聚类中心算法步骤

Tab. 2 Process of initializing clustering centers based on histogram peaks search

1. 确定待分选对象的每维参数预分箱数 p_m, $m=1,2\cdots,M$, 可分为 $P=\prod_{m=1}^{M}p_m$ 个区间。
2. 统计待分选对象集 $\boldsymbol{X}$ 落入各个区间的对象数目 $Y=\{y_{i_1 i_2 \cdots i_M}\}\big|_{i_1=1,2,\cdots,p_1;\cdots;i_M=1,2,\cdots,p_M}$。
3. 搜寻直方图峰值, 并求峰值位置对应区间的质心。
4. 峰值数目作为聚类数目, 峰值位置对应区间的质心作为初始聚类中心。

直方图峰值搜索基本原理是: 将每个区间的对象数目与相邻的区间比较, 如果某个区间的对象数目均大于各邻区间数目, 则该区间即为峰值所在区间。任何一个非边界区间的相邻区间为 $2M$ 个, 边界上的区间的相邻区间数目范围为 $[M, 2M-1]$, 定义区间 $cell_{i_1 i_2 \cdots i_M}$ 相邻区间的集合为 $Neignbor_{i_1 i_2 \cdots i_M}=\{cell_{(i_1\pm1) i_2 \cdots i_M}, \cdots, cell_{i_1 i_2 \cdots (i_M\pm1)}\}$。

直方图预分箱数 $p_m(m=1,2\cdots,M)$ 的选择非常重要。预分箱数过少, 则可能漏掉某些峰值的检测; 预分箱数过多, 又增加了算法复杂度, 并且会因为噪声的影响得到一些伪峰值, 虚警概率增加, 并易使分选陷入局部最小。根据经验得知, 预分箱数应为实际类数目的 3 倍左右。如果实际类数目未知, 则只有选取较多的预分箱数, 噪声引起的伪峰值通过设定峰值门限去掉。

与其他优化初始聚类中心方法相比, 本文方法计算量大为减少(尤其在参数维数较少情况下)。其运算量共分为两部分: 一是直方图统计步骤约 $N\times M$ 次判断运算; 二是直方图峰值搜索步骤, 约为 $P\times 2M$ 次比较运算。另外改进爬山法的算法复杂度为 $O(MN^2)$, 文献[9]的 psFCM 算法约有 $2N^2M/P$ 次加法和 $2N^2M/P$ 次乘法。表 3 为不同情况下改进爬山法[6-7]、psFCM 算法和本文方法的计算量比较。

表 3 与传统方法的计算量比较

Tab. 3 Compute complexity of new method compared with conventional methods

	改进爬山法	psFCM 算法	搜寻直方图峰值方法
$N=1000, M=2; p_1=\cdots=p_M=10;$	2×10^6	8×10^4(乘加法各半)	2.4×10^3
$N=2000, M=2; p_1=\cdots=p_M=20;$	8×10^6	4×10^4(乘加法各半)	5.6×10^3
$N=1000, M=3; p_1=\cdots=p_M=10;$	3×10^6	1.6×10^4(乘加法各半)	9×10^3

(注: 计算量均为约数)

3 基于改进 K-Means 算法的跳频网台分选

跳频通信同一网内跳周期(跳频电台载频转换时间很短, 跳周期可用每跳持续时间代替)是相同的, 所以一般先根据跳周期参数将待分选 hop 分成多个网络; 跳频组网多为异步组网, 同网内不同电台跳时不同, 可以根据跳时再对同网内电台进行分选。流程如图 1。

值得注意的是, 对于同步组网电台, 因为同网内各电台同时改变载频, 跳时基本相同, 所以同步组网电台的分选需依靠别的参数, 如 *DOA*。当然异步组网电台也可依靠 *DOA* 信息, 只是通常 *DOA* 信息的获取更为困难。利用跳周期和 *DOA* 的聚类流程与图 1 相似, 只是将第 3 步骤的"根据跳时聚类"改为"根据 *DOA* 聚类"即可。

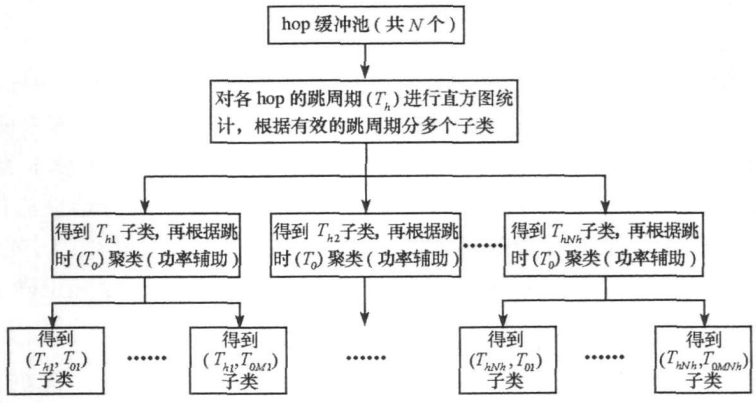


图 1 异步组网电台的跳周期- 跳时聚类过程

Fig. 1 Clustering process of non-orthogonal FH signals based on $T_h - T_0$

4 仿真结果

为验证改进 K-Means 算法, 分别利用 UCI 数据库的 Iris、Breast-w 数据[11] 和人造的跳频 HDW 数据进行了仿真。

4.1 真实数据集的聚类仿真

UCI 数据库的数据均为真实数据, 常用来检验聚类算法的优劣。Iris 数据为三类 Iris 叶片的统计数据, 每个数据包含四维属性; Breast-w 数据为美国威斯康星州乳腺癌肿瘤的统计数据, 分良性和恶性两类, 每个数据包含 10 维属性。对两种数据均利用第 3、4 维属性参数进行聚类, 直方图预分箱数 $p_1 = p_2 = 10$; 分别利用传统的 K-Means 算法和本文的优化初始聚类中心 K-Means 算法对上两种数据进行聚类, 聚类性能评价指标为聚类迭代次数和聚类正确率。因为传统 K-Means 算法随机选取初始聚类中心, 每次聚类的结果会有很大不同, 故仿真 100 次, 然后取迭代次数和聚类正确率的平均值。得到仿真结果如表 4 所示。

表 4 两种真实数据的聚类性能

Tab. 4 Clustering performance of two real databases

数据 / 算法	Iris 数据		Breast-w 数据	
	迭代次数	聚类正确率	迭代次数	聚类正确率
传统 K-Means 算法	2.03	[85.2% 86.3% 84.9%]	3.97	[98.3% 76.8%]
本文算法	1	[100% 90.0% 98.0%]	3	[98.3% 76.8%]

由表 4 可以看出, 对于 Iris 数据, 优化初始聚类中心后, 迭代次数减少, 聚类正确率也得到提高; 对 Breast-w 数据, 聚类正确率相同, 迭代次数稍有减少。

4.2 基于改进 K-Means 算法的跳频信号聚类分选

对改进 K-Means 算法在跳频网台分选的应用进行了仿真。仿真条件如表 5。

表 5 跳频信号聚类分选的仿真参数

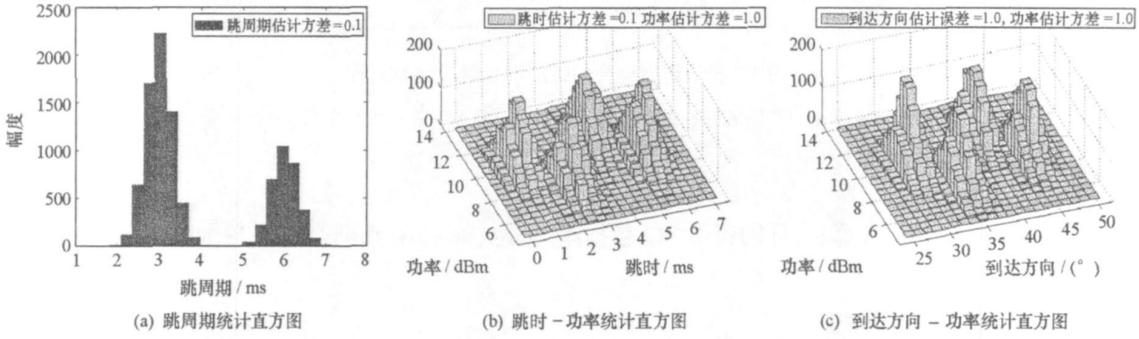
Tab. 5 Simulation parameters of FH signals clustering

仿真实验	仿真参数
基于跳周期聚类的网络分选	N= 10 000, M= 1, p_1= 20; η= N/P; T_h= [6 3] ms
基于跳时- 功率聚类的异步组网网台分选	N= 10 000, M= 2, p_1= p_2= 20; W= [1 0.1]; η= N/P; T_0= [1 2.5 4 5.5] ms; PA= [10 8 11 9] dBm
基于 DOA- 功率聚类的网台分选	N= 10 000, M= 2, p_1= p_2= 20; W= [1 1]; η= N/P; DOA= [30 35 40 45]°; PA= [10 8 11 9] dBm

将各维参数真值加上一定方差的高斯白噪声, 然后应用改进的 K-Means 聚类算法。

4.2.1 基于跳周期聚类的网络分选

网络分选仅仅利用跳周期一维参数, 其统计直方图为二维图像, 如图 2(a) 。分选性能见图 3, 其中左图为聚类迭代次数- 参数估计方差曲线, 右图为正确分选概率- 参数估计方差曲线, 实线为传统 K-Means 算法的性能, 虚线为本文算法性能。由左图可知, 优化初始聚类中心可使 K-Means 算法迭代次数大大减少, 跳周期的估计方差小于 10^{-1} 时只要一次迭代即可; 由右图可知分选性能良好, 方差小于 10^{-1} 时分选正确率为 100%, 并且两类算法分选性能相差不大。由此可见, 优化聚类中心的主要优点是降低分选迭代次数和避免分选进入局部最小, 减少计算量和提高算法鲁棒性, 分选正确率提高不大, 故下面仅比较两者的迭代次数。



(a) 跳周期统计直方图 (b) 跳时 - 功率统计直方图 (c) 到达方向 - 功率统计直方图

图 2 跳频参数的统计直方图

Fig. 2 The histograms of FH parameters

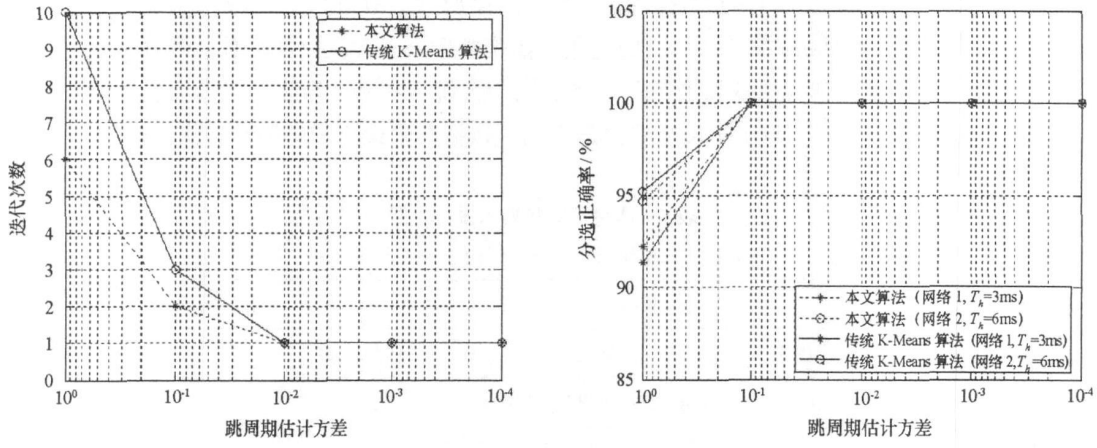


图 3 基于跳周期聚类的网络分选性能

Fig. 3 Networks clustering performance based on T_h

4.2.2 基于跳时- 功率聚类的异步组网网台分选

完成网络分选后, 对跳时- 功率参数进行聚类。二维参数的统计直方图为三维图像, 如图 2(b) , 图中 4 座山峰代表 4 个网台。

为避免误差传递, 假设第一步的网络分选完全正确。通常功率估计精度不会很高, 故假设功率估计方差始终是跳时估计方差的 10 倍, 权值矩阵 $\boldsymbol{W}$= [1 0.1]。其分选性能见图 4, 与传统 K-Means 算法相比, 采取优化初始聚类中心的本文算法迭代次数大为减少。

4.2.3 基于到达方向- 功率聚类的网台分选

基于到达方向- 功率聚类的统计直方图为图 2(c) ; 其网台分选性能如图 5, 由图可知两参数的方差小于 1 时分选正确率大于 99% , 分选效果良好。该仿真两参数的估计方差相当, 故权值矩阵取 $\boldsymbol{W}$= [1 1]。聚类过程中参数维数更多相当于对象间距离增大, 分选性能当然更好(与图 3、图 4 相比较可知) 。

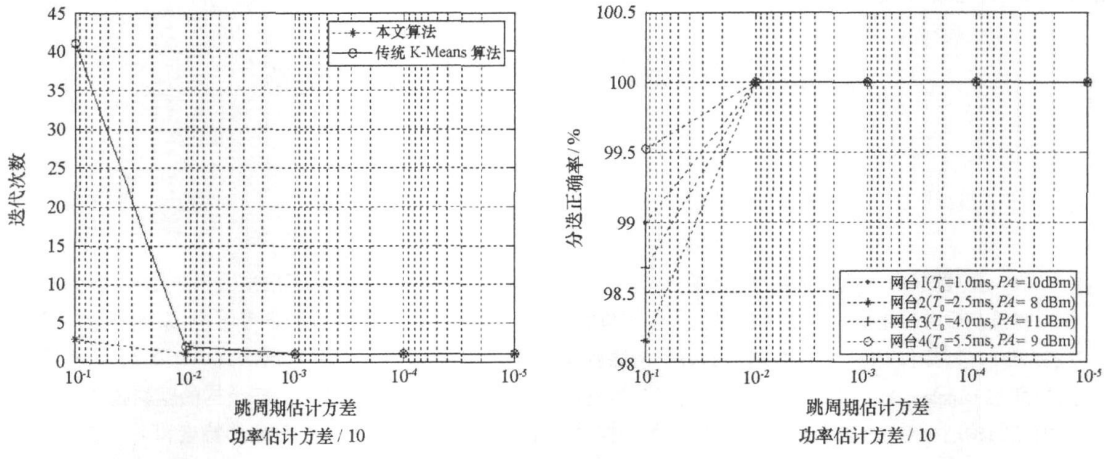


图 4 跳时- 功率聚类的网台分选性能

Fig. 4 Signals sorting performance of clustering based on T_0 - PA

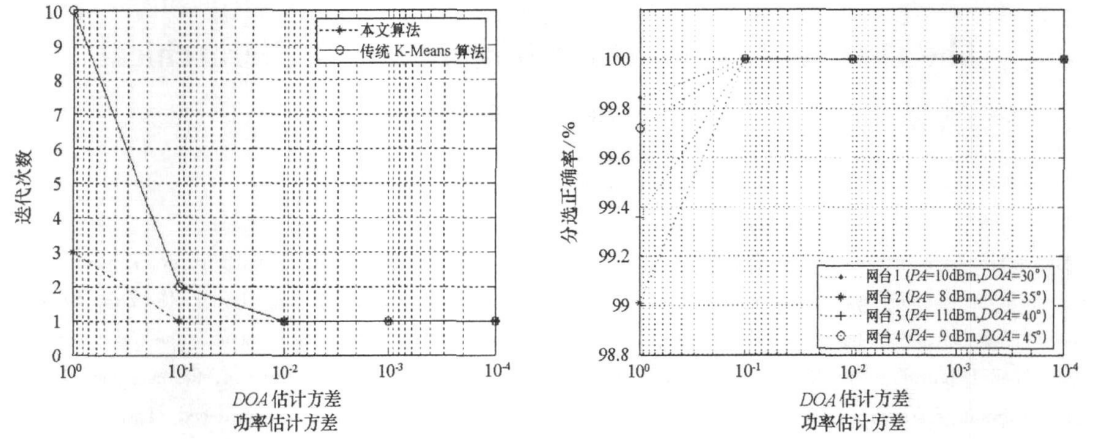


图 5 到达方向- 功率聚类的网台分选性能

Fig. 5 Signals sorting performance of clustering based on DOA- PA

5 结束语

本文提出的参数统计直方图峰值搜索方法能预估类数目 K, 并能快速得到初始聚类中心, 其值已非常接近实际类心, 能有效减少迭代次数。在较低维数情况下, 初始聚类中心获取过程的算法复杂度降低到 $O(N)$。将改进的 K-Means 聚类算法应用于跳频信号的网络分选、网台分选, 仿真结果表明算法非常高效。

参 考 文 献:

[1] Han J W, Kamber M. Data Mining Concepts and Techniques[M]. Singapore: Elesvier Inc., 2006.
[2] 孙即祥, 等. 现代模式识别[M]. 长沙: 国防科技大学出版社, 2002.
[3] Ye Y M, Huang Z X, Chen X J, et al. Neighborhood Density Method for Selecting Initial Cluster Centers in K-Means Clustering[C]// Proceedings of PAKDD'06: Advances in Knowledge Discovery and Data Mining, 10th Pacific-asia Conference, Singapore: Springer, 2006: 189- 198.
[4] He J, Lan M, Tan C L, et al. Initialization of Cluster Refinement Algorithms: A Review and Comparative Study[C]// Proceedings of International Joint Conference on Neural Networks, Budapest, 2004: 297- 302.
[5] Yager R R, Filev D P. Approximate Clustering Via the Mountain Method[J]. IEEE Trans. on SMC, 1994, 24(8): 1279- 1284.
[6] Chiu S L. Fuzzy Model Identification Based on Cluster Estimation[J]. Journal of Intelligent and Fuzzy Systems, 1994, 2(3): 267- 278.
[7] 裴继红, 范九伦, 谢维信. 聚类中心的初始化方法[J]. 电子科学学刊, 1999, 21(3): 320- 325.
[8] Kaufman L. Finding Groups in Data: An Introduction to Cluster Analysis[M]. New York: Wiley, 1990.
[9] Hung M C, Yang D L. An Efficient Fuzzy C-means Clustering Algorithm[C]// Proceedings of the 2001 IEEE International Conference on Data Mining, Washington: 2001: 225- 232.
[10] 牛琨, 张舒博, 陈俊亮. 融合网格密度的聚类中心初始化方案[J]. 北京邮电大学学报, 2007, 30(2): 6- 10.
[11] http://www.ics.uci.edu.