异构网络中丢包隶属度函数的构建方法

甄雁翔，苏 放，寇明延，徐惠民

(北京邮电大学信息与通信工程学院 北京 海淀区 100876)

【摘要】异构网络中TCP丢包区分机制对无线网络的稳定性起着重要的作用，包对探测包的单向传输时延(ROD)作为区分参数对不同丢包类型进行区分，算法的准确度依赖于ROD样本的隶属度函数及其参数估算。为了更加准确地区分丢包类型，通过对传统高斯混合模型的EM算法进行分析，提出了基于势函数的初始化方法，并且在网络拥塞和无线误码同时存在的情况下，将改进的EM算法(PEM)应用于不同丢包模式下隶属度函数的构建中。仿真验证了该算法具有较好的收敛特性和稳定性，并且对不同丢包模式隶属度函数的确定达到了很好的构建效果。

关 键 词 EM算法； 高斯分布；异构网络； 隶属度函数； 势函数

中图分类号 TN913 **文献标识码** A doi:10.3969/j.issn.1001-0548.2010.06.009

Construction Method of Packet Loss Membership Function in Heterogeneous Networks

ZHEN Yan-xiang, SU Fang, KOU Ming-yan, and XU Hui-min

(School of Information and Communication Engineering, Beijing University of Posts and Telecommunications Haidian Beijing 100876)

Abstract The packet loss differentiating mechanism of TCP for heterogeneous networks plays an important role in the stability of wireless networks, relative one-way delay (ROD) of packet pair is used as the differentiating parameter to distinguish the loss type. The accuracy of this algorithm depends on ROD samples membership functions and parameters estimation. In order to differentiate the packet loss pattern more accurately, the initialization method based on potential functions is proposed by analyzing the traditional expectation maximization (EM) algorithm in Gaussian mixture model. Then the improved EM (PEM) algorithm is applied in the construction of different packet loss membership functions in the situation when the network congestion and wireless error coexist. The simulation results indicate that this algorithm has better convergence characteristics and stability, and has well building effect in the construction of different packet loss pattern membership functions.

Key words EM algorithm; Gaussian distribution; heterogeneous networks; membership functions; potential functions

随着无线应用技术的快速发展，无线广域网(GPRS、UMTS等)、无线局域网(IEEE802.11)、卫星通信网、蓝牙网络等多种无线网络系统正逐步代替传统有线网络成为互联网接入的最后一跳。TCP是目前Internet中广泛使用的端到端传输控制协议，在有线/无线融合的异构网络环境中，由于无线网络中存在高误码率、信号衰落、切换等原因，网络拥塞不是引起TCP分组丢失的唯一原因，无线误码也会引起分组的丢失[1-2]。而传统的TCP协议把所有的分组丢失简单地归因于网络拥塞，就会造成盲目减小发送速率，降低带宽利用率，导致TCP性能的无谓恶化[3-4]。因此，只有正确区分无线误码所造成的丢包和网络拥塞所造成的丢包，分别采取不同的控制策略，才能提高网络的效率。

文献[5]采用包对(packet pair)探测包的ROD作为区分参数对不同丢包类型进行区分，采用条件概率构造不同丢包模式下的隶属度函数，从而按照最大隶属度原则进行丢包区分，提出了异构网络下基于Fuzzy模式识别的丢包区分算法，该算法的准确度依赖于ROD样本的隶属度函数及其参数估算。

传统的参数估计方法如最大似然估计、最小二乘法等是在先验知识和类标号已知的条件下进行的。但在异构网络丢包区分的应用领域里，会出现对丢包分类不了解，即没有先验知识的情况，因此需要借助无监督的分类技术。聚类分析就是在没有任何关于分类先验知识的情况下，依据数据的相似

收稿日期：2009 – 06 – 10； 修回日期：2009 – 10 – 14
基金项目：国家自然科学基金(60572122)
作者简介：甄雁翔(1982 –)，女，博士，主要从事无线多媒体应用方面的研究.

性划分数据的类[6]。目前主要的聚类技术有以下几类：划分方法、层次方法、基于密度的方法、基于网格的方法、基于模型的方法等[7]。这些方法采用不同的方式确定聚类中心，但都有各自的局限性。如划分方法算法简单，但只适用于识别中小规模具有相近尺寸和密度的球状类数据集，对大规模的数据集以及复杂形状的聚类，需要进一步地扩展。层次方法的局限性在于合并点或分类点的选择，如果在某一步没有很好地做出合并或分裂的决定，可能会导致低质量的聚类结果。基于密度的方法能够在有噪声的空间数据库中发现任意形状的聚类，但对于密度分布不均的数据集得不到满意的聚类效果；此外，其对于密度函数参数的设置也比较敏感。基于网格的方法处理速度较快，但该算法精确性较低。基于模型的聚类方法是为每一类假定一个模型，然后寻找数据对给定模型的最佳拟合。因此，聚类算法的选择依赖于可用数据的类型和应用的特定目标，有些情况可以根据聚类标准综合多种聚类技术。

由文献[5]和大量仿真实验可知，包对探测包的ROD样本在不同丢包模式下的隶属度函数符合高斯混合分布，每一类都可以用参数概率分布数学描述。对不同丢包模式隶属度函数的构建即确定隶属度函数的参数，基于模型的聚类方法正是试图优化给定数据和某数学模型之间的拟合，因此选取基于模型的聚类方法即可描述不同丢包模式下的隶属度函数。传统的EM算法[8]是一种基于统计模型进行期望最大化分析的算法，对初始值的选择具有很强的依赖性。本文通过对传统高斯混合模型的EM算法进行分析，提出了基于势函数[9]的方法确定聚类中心，并且在网络拥塞和无线误码同时存在的情况下，将改进后的PEM算法运用到不同丢包模式下隶属度函数的构建中。仿真验证PEM算法比传统EM算法具有较好的稳定性和收敛特性，对不同丢包模式的隶属度函数可以达到很好的区分效果，具有很好的实际应用价值。

1 高斯混合模型的EM算法

1.1 算法描述

高斯混合分布的概率密度模型[8]为：

$$P(x_i)=\sum_{k=1}^{M}\pi_k\frac{1}{\sqrt{2\pi}\sigma_k}\exp\left(-\frac{(x_i-\mu_k)^2}{2\sigma_k{}^2}\right) \tag{1}$$

式中 M 是类的个数；π_k 是每类分布的权重；$\sum_{k=1}^{M}\pi_k=1$；$N(\mu_k,\sigma_k)$ 表示每类高斯分布。

算法步骤如下：

(1) 对混合模型的参数进行初始化。

(2) 期望步(E-step)。用 $P(x_i\in C_k)$ 概率将每个样本 x_i 指派到类 C_k 中：

$$P(x_i\in C_k)=P(C_k\mid x_i)=\frac{P(C_k)P(x_i\mid C_k)}{P(x_i)} \tag{2}$$

其中：

$$\begin{cases}P(x_i\mid C_k)=N(x_i;\mu_k,\sigma_k)=\dfrac{1}{\sqrt{2\pi}\sigma_k}\exp\left(-\dfrac{(x_i-\mu_k)^2}{2\sigma_k{}^2}\right)\\ P(C_k)=\pi_k\\ P(x_i)=\sum_{k=1}^{M}\pi_k\dfrac{1}{\sqrt{2\pi}\sigma_k}\exp\left(-\dfrac{(x_i-\mu_k)^2}{2\sigma_k{}^2}\right)\end{cases}$$

(3) 最大化(M-step)。该步对给定样本的分布似然“最大化”，利用每一类的隶属度概率之和 $\alpha_k{}^{(0)}=\sum_{i=1}^{n}P(x_i\in C_k)$（$n$ 为样本的个数)重新估计(求精)模型参数，迭代公式为：

$$\begin{cases}\pi_k{}^{(j+1)}=\dfrac{\alpha_k{}^{(j)}}{n}\\ \mu_k{}^{(j+1)}=\dfrac{\sum_{i=1}^{n}x_iP(x_i\in C_k)}{\alpha_k{}^{(j)}}\\ \sigma_k{}^{(j+1)}=\dfrac{\sum_{i=1}^{n}P(x_i\in C_k)(x_i-\mu_k^{(j+1)})^2}{\alpha_k{}^{(j)}}\end{cases} \tag{3}$$

1.2 算法分析

由以上算法描述可知，EM算法的核心是根据一个代表隶属度概率的权值将每个对象指派到类中，并不断迭代使之收敛于某个最优值。运用EM算法实现高斯混合模型聚类，如何初始化参数是一个关键问题，EM算法收敛的优劣很大程度上取决于其初始参数[10]。

对于高斯混合模型，参数的初始化主要包括每一类的均值 μ_k、方差 σ_k^2、权重 π_k 和类的个数 M。对于中心值的选取，目前常用的方法是随机选取，然后通过不断迭代调整达到最优。该种随机选取的方法虽然操作简单，但对于算法的收敛稳定性有一定的影响，有时会收敛于局部最小值，而不能得到全局最优解，使得聚类效果受到影响；在数据规模较大的情况下，效率也很低，耗时长且需要较大的存储空间。本文提出了基于势函数的方法选取聚类中心，仿真验证，在达到同样精度的条件下，该算法具有较好的收敛速度和稳定性。

1.3 PEM算法

基于势函数的方法确定聚类中心值，算法步骤如下。

(1) 计算每个样本点的初始势值，定义初始势函数为：

$$P_i^{(1)}=\sum_{j=1}^{n}\exp\left(-\frac{4}{r_a^{\,2}}(\exp(-N_i)\cdot\left\|x_i-x_j\right\|^2)\right) \quad (4)$$

式中　N_i 为该样本点发生的次数；r_a 为一个正常数，表示领域半径[8]。

取 $P_1^*=\max\{P_i^{(1)},i=1,2,\cdots,n\}$，$x_1^*$ 为第一个聚类中心值。

(2) 计算剩余样本点的更新势值，定义更新势函数公式为：

$$P_i^{(k+1)}=P_i^{(k)}-P_k^*\exp\left(-\frac{4}{r_b^{\,2}}(\exp(-N_i)\cdot\left\|x_i-x_k^*\right\|^2)\right) \quad (5)$$

式中　r_b 为一个正常数，表示领域半径，取 $r_b=1.5r_a$[8]；依次取 $P_k^*=\max\left\{P_i^{(k)},i=1,2,\cdots,n\right\}$；$x_k^*$ 为第 k 个聚类中心值，$k=1,2,\cdots,M$，M 是类的个数。

由上述势函数公式可知，样本点的势值不仅与样本点间的距离有关，还与样本点发生的次数有关。当样本点发生的概率比较大，且具有足够高的密度，其势值就会相对较大。利用上述迭代公式，可以依次找到样本的最密集点 μ_1，次密集点 μ_2、μ_2、…、μ_M，这些点可以作为高斯混合模型中每个高斯分布的初始均值。

1.4 仿真验证

在Matlab7仿真平台上，随机产生几类高斯混合分布，由势函数的方法确定出聚类中心，如表1所示。由势函数确定的聚类中心可以近似表示为每个高斯分布的均值，从而验证了算法的有效性。

表1　由势函数确定的聚类中心值

类数M	理论产生的高斯混合分布的均值	势函数确定的聚类中心值
2	−5.816 4	−6.512 1
	−8.391 9	−9.526 2
3	−3.460 5	−2.213 5
	−13.324 7	−13.245 0
	1.002 7	0.505 0
4	−6.563 5	−6.923 3
	−2.205 5	−2.971 5
	−12.165 6	−13.660 8
	−10.363 2	−10.425 1

对理论构造的高斯混合分布数据，选取不同类数的样本集合，采用不同的初始化方法进行聚类仿真。通过大量仿真实验可知，达到相同的精度，随机选取聚类中心迭代次数不稳定，具有很大的随机性。基于势函数的方法选取聚类中心迭代次数稳定且收敛速度快，表2为不同初始化方法的平均迭代次数比较。

表2　理论样本不同初始化方法平均迭代次数比较

类数M	EM迭代次数	PEM迭代次数
2	922	18
3	2 155	40
4	3 109	58

2 不同丢包模式隶属度函数的构建

2.1 不同丢包模式隶属度函数的特点

在网络拥塞和无线误码同时存在的情况下，通过对文献[5]和大量包对探测包的ROD样本进行分析可知，拥塞丢包模式下ROD的隶属度函数为高斯混合分布，瓶颈的个数即为高斯混合分布类的个数。无线误码丢包模式下ROD的隶属度函数也为高斯混合分布，但与误码率有关。误码率较大时，由误码引起的丢包概率增大，由TCP协议[11]可知，拥塞窗口会不断进行减半调整，发生拥塞的概率减小；当误码率较小时，发生拥塞的概率增大，拥塞丢包的特征变得明显。因此，无线误码模式下丢包隶属度函数随误码率的变化而变化[12]。

2.2 不同丢包模式下隶属度函数的确定

拥塞丢包模式下，由于网络中可能会有多个瓶颈，通过大量仿真实验可知，每个瓶颈处包对探测包的ROD样本可以用一个高斯分布来描述。当瓶颈数大于两个时，瓶颈处的ROD样本的次数很少，在高斯混合分布中的权重也很小，与瓶颈为两个的ROD分布特征近似。因此，拥塞模式的隶属度函数可以由一个类数为2的高斯混合分布来描述。

无线误码丢包模式下，通过大量仿真实验可知，当误码率较大时，隶属度函数的均值和方差都比较小，当误码率较小的时候，隶属度函数的方差比较大。由于无线误码丢包具有很大的随机性，所以ROD分布具有重尾现象[13]。高斯混合模型的EM算法对重尾数据的描述是用一个方差比较大的高斯分布进行拟合。因此，无线误码模式的隶属度函数可以由一个类数为2的高斯混合分布来描述。

由以上分析可知，在网络拥塞和无线误码同时存在的情况下，类的个数为4。因此，在利用势函数的方法确定聚类中心时，选取4类进行迭代。

类个数确定后，定义由PEM算法描述不同丢包模式的隶属度函数为：

$$\begin{cases} P_{\mathrm{C}}=\sum_{k=1}^{2}\pi_{\mathrm{C}k}\dfrac{1}{\sqrt{2\pi}\sigma_{\mathrm{C}k}}\exp\left(-\dfrac{(x_i-\mu_{\mathrm{C}k})^2}{2\sigma_{\mathrm{C}k}^{\ 2}}\right) \\ P_{\mathrm{W}}=\sum_{k=1}^{2}\pi_{\mathrm{W}k}\dfrac{1}{\sqrt{2\pi}\sigma_{\mathrm{W}k}}\exp\left(-\dfrac{(x_i-\mu_{\mathrm{W}k})^2}{2\sigma_{\mathrm{W}k}^{\ 2}}\right) \end{cases} \tag{6}$$

式中 P_{C} 为拥塞丢包模式的隶属度函数；P_{W} 为无线误码丢包模式的隶属度函数。

采用PEM算法估算出4个高斯分布的参数，然后根据不同丢包模式下隶属度函数的特点，将估算出的高斯分布的均值 μ_k 和方差 σ_k^2 作为不同丢包模式的区分参数进行分类。

(1) 比较方差：选取方差最大的一类高斯分布作为无线误码丢包的隶属度函数。

(2) 比较均值：对其余3类高斯分布比较均值，选取均值最小的作为无线误码丢包的隶属度函数，其余两类即为网络拥塞丢包的隶属度函数。

由以上步骤即可确定式(6)不同丢包模式的隶属度函数。

2.3 仿真验证

利用Matlab7进行仿真，表3为不同网络条件下丢包隶属度函数统计参数表，表4为不同初始化方法平均迭代次数的比较。

表3 不同网络条件下丢包隶属度函数统计参数表

a. 瓶颈数为2，误码率为1‰

参数	P_{C_1}	P_{C_2}	P_{W_1}	P_{W_2}
势函数确定的中心值	0.213 0	0.444 0	0.214 0	2.054 0
μ_k	0.214 9	0.338 5	0.131 0	2.049 8
σ_k	0.004 2	0.161 7	0.032 9	1.849 8
π_k	0.868 4	0.037 2	0.091 0	0.003 3

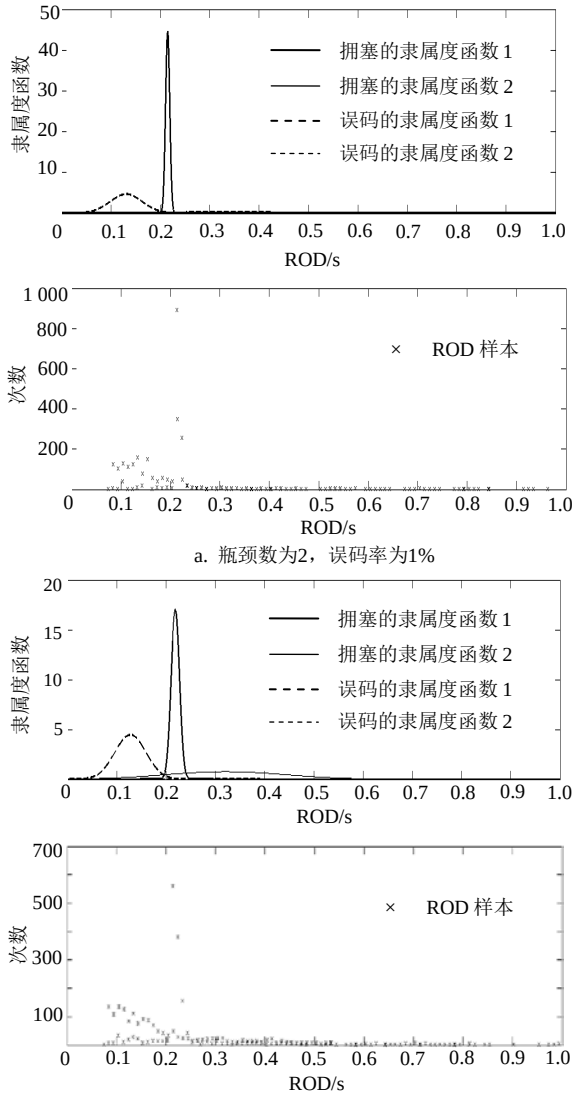
b. 瓶颈数为2，误码率为1%

参数	P_{C_1}	P_{C_2}	P_{W_1}	P_{W_2}
势函数确定的中心值	0.213 0	0.223 0	0.074 0	1.203 0
μ_k	0.219 0	0.320 6	0.127 8	2.788 8
σ_k	0.008 4	0.122 3	0.031 3	2.750 9
π_k	0.710 6	0.096 4	0.187 7	0.005 3

表4 ROD样本不同初始化方法平均迭代次数比较

网络条件	EM迭代次数	PEM迭代次数
瓶颈数为2，误码率为1‰	285	53
瓶颈数为2，误码率为1%	188	62

可以看出，对于ROD样本，PEM算法比传统EM算法具有较好的稳定性和收敛特性。图1为不同网络条件下ROD样本分布及隶属度函数对比图。比较ROD样本分布图和由PEM算法确定的隶属度函数，可以看出隶属度函数能近似表示实际样本的概率分布，从而验证了该方法模型的有效性。



a. 瓶颈数为2，误码率为1%

b. 瓶颈数为2，误码率为1%

图1 不同网络条件下ROD样本分布与隶属度函数对比图

3 结 束 语

本文通过对传统高斯混合模型的EM算法进行分析，针对传统EM算法在收敛速度和稳定性上的问题，在保持算法迭代简单的前提下，提出了基于势函数的方法来确定初始聚类中心，从而达到更稳定的聚类效果。并在网络拥塞和无线误码同时存在的情况下，将PEM算法运用到异构网络条件下不同丢包模式隶属度函数参数的估算中。仿真验证该算法具有很好的收敛速度和稳定性，对不同丢包模式的隶属度函数可以达到很好的区分效果，进而可以更加准确地区分丢包类型，提高网络的效率。

本文的研究工作得到了华为高校科技基金(YJCB2005055WL)的资助，在此表示感谢！

参 考 文 献

[1] 范英磊, 郑培超, 苏　放, 等. 异构网络中的视频传输服务质量框架[J]. 电子科技大学学报, 2008, 37(1): 90-93.
FAN Ying-lei, ZHENG Pei-chao, SU Fang, et al. An QoS framework for video delivery over heterogeneous Networks[J]. Journal of University of Electronic Science and Technology of China, 2008, 37(1): 90-93.

[2] CHEN Ming-hua, ZAKHOR A. Rate control for streaming video over wireless[C]//INFOCOM 2004. Hong Kong, China: IEEE, 2004: 2(2): 1181-1190.

[3] TODOROVIC M, Lopez-Benitez N. Efficiency study of TCP protocols in infrastructured wireless networks[C]// ICNS'06: International Conference on Networking and Services. Slicon Valley CA: IEEE Computer Society Press, 2006: 103-108.

[4] SONG C, COSMAN P C, VOELKER G M. End-to-end differentiation of congestion and wireless losses[J]. Networking, IEEE/ACM Transactions on networking, 2003, 11(5): 703-717.

[5] 苏　放, 范英磊. 一种基于Fuzzy丢包区分的TCP拥塞控制算法[J]. 系统仿真学报, 2008, 20(7): 1904-1908, 1911.
SU Fang, FAN Ying-lei. TCP congestion control algorithm based on fuzzy loss differentiating[J]. Journal of System Simulation, 2008, 20(7): 1904-1908, 1911.

[6] 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: 37-60.
GAO Xin-bo. Fuzzy cluster analysis and its applications[M]. Xian: Xidian University Press, 2004: 37-60.

[7] 韩家炜, 堪　博. 数据挖掘概念与技术[M]. 范　明, 孟小峰, 译. 第2版. 北京: 机械工业出版社, 2008.
HAN Jia-wei, KAMBE Micheline. Data mining concepts and techniques[M]. Translated by FAN Ming, MENG Xiao-feng. 2nd ed. Beijing: China Machine Press, 2008.

[8] BILMES J A. A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models[DB/OL]. [2009-01-08]. http://ssli.ee.washington.edu/people/bulyko/papers/em.pdf.

[9] CHIU S L. A cluster estimation method with extension to fuzzy model identification[C]//Proceedings of the Third IEEE Conference on Fuzzy Systems. Orlando FL: IEEE Congress on Computational Intelligence, 1994, 2: 1240-1245.

[10] REDMOND S J, HENEGHAN C. A method for initializing the K-means clustering algorithm using kd-trees[J]. Pattern Recognition Letters, 2007, 28(8): 965-973.

[11] 史蒂文斯. TCP/IP详解卷1: 协议[M]. 范建华, 胥光辉, 张　涛, 等, 译. 北京: 机械工业出版社, 2000: 209-244.
STEVENS W R. TCP/IP illstrated volume 1: the protocols[M]. Translated by FAN Jian-hua, XU Guang-hui, ZHANG Tao, et al. Beijing: China Machine Press, 2000: 209-244.

[12] BI Jing-ping, WU Qi, LI Zhong-cheng. Packet delay and packet loss in the Internet[C]//Proceedings of Seventh International Symposium on Computers and Communications. Taormina Italy: IEEE Computer Society Press, 2002: 3-8.

[13] RIZVI A A, HUSSAIN A. Analysis of single server queueing systems with heavy tail distributions[C]//7th International Multi Topic Conference. Islamabad: IEEE INMIC, 2003: 176-181.

编　辑　漆　蓉

(上接第844页)

[3] CHOI J D, STARK W E. Performance of ultra-wideband communications with suboptimal receivers in multipath channels[J]. IEEE J Sel Areas Commun, 2002, 20(9): 1754-1766.

[4] CASSIOLI D, WIN M Z, VATALARO F, et al. Low complexity rake receivers in ultra-wideband channels[J]. IEEE Trans on Wireless Commun, 2007, 6(4): 1265-1275.

[5] 岳光荣, 李少谦. 超宽带冲激无线电性能比较[J]. 电子科技大学学报, 2003, 32(5): 477-480.
YUE G R, LI S Q. Performance comparison of Ultra wideband impulse radio[J]. Journal of University of Electronic Science and Technology of China, 2003, 32(5): 477-480.

[6] WALDEN R H. Analog-to-digital converter survey and analysis[J]. IEEE J Sel Areas Commun, 1999, 17(4): 539-550.

[7] LE B, RONDEAU T W, REED J H,et al. Analog-to-digital converters[J]. IEEE Sig Proc Mag, 2005, 22(6): 69-77.

[8] VETTERLI M, MARZILIANO P, BLU T. Sampling signals with finite rate of innovation[J]. IEEE Trans on Sig Proc, 2002, 20(6): 1417-1428.

[9] BLU T, DRAGOTTI P L, VETTERLI M, et al. Sparse sampling of signal innovations[J]. IEEE Signal Processing Magazine, 2008, 25(2): 31-40.

[10] KUSUMA J, RIDOLFI A, VETTERLI M. Sampling of communication systems with bandwidth expansion[C]// The 2002 IEEE International Conference on Communications. New York: IEEE, 2002, 3: 1601-1605.

[11] KUSUMA J, MARAVIC I, VETTERLI M. Sampling with finite rate of innovation: channel and timing estimation for UWB and GPS[C]//The 2002 IEEE International Conference on Communications. Anchorage, AK, USA: IEEE, 2003, 5: 3540-3544.

[12] HAYES M H. Statistical digital signal processing and modeling[M]. New York: John Wiley and Sons, 1996.

[13] 张贤达. 现代信号处理[M]. 北京: 清华大学出版社, 2002.
ZHANG X D. Modern signal processing[M]. Beijing: Tsinghua University Press, 2002.

编　辑　张　俊