文章编号：2095-4980(2013)02-0260-06

基于高阶累积量的核 Logistic 回归调制分类算法

徐　闻，王　斌

(解放军信息工程大学 信息工程学院，河南 郑州 450002)

摘　要：针对现有数字信号调制识别的问题，提出了一种基于核 Logistic 回归(KLR)的自动分类方法。该方法提取了信号的高阶累积量参数用作训练与测试数据，采取常用的决策树分类构架的思想，仿真并比较已有的基于支撑向量机(SVM)的调制分类方法，结果表明，在低信噪比为 0 dB 时，分类性能一般高于 SVM；5 dB 时，采用 KLR 的分类识别率均达到 90%以上，有较为优越的分类性能。

关键词：调制识别；分类；高阶累积量；核 Logistic 回归；决策树

中图分类号：TN914　　**文献标识码：**A

A method of modulation classification of Kernel Logistic Regression based on high-order cumulants

XU Wen，WANG Bin

(Institute of Information Engineering，Information Engineering University，Zhengzhou Henan 450002，China)

Abstract：Aiming to the problem of automatic modulation classification of the existing digital signal，a classification method based on Kernel Logistic Regression(KLR) is developed.This method is primarily used in economic，medical science and speech process etc，while seldom applied in the field of communication signals. The characteristic parameter of high-order cumulants of the signal is used for training data and testing data.The classification is performed adopting the frequently-used decision tree method. The proposed method is compared to the modulation classification method based on Support Vector Machine(SVM) through simulation experiments. The results indicate that the proposed method is qualified to do the work.Under low SNR(0 dB)，the performance of classification is higher than that based on SVM; while under 5dB，the correct recognition rate is above 90% based on KLR.

Key words：modulation recognition；classification；high-order cumulants；Kernel Logistic Regression；decision tree

随着通信技术的发展、数字技术的广泛运用，无线通信环境日益复杂，通信信号在很宽的频带上采用不同的调制参数、不同的调制方式[1]。调制分类的目的是在未知信号调制信息的前提下，能够给出信号的调制方式和相应的调制参数，再根据特征量对信号划分类别，为进一步分析和后续的信号处理提供依据。高阶累积量对噪声的三阶及以上累积量为零，所以其对噪声的抑制比较好[2-3]。但是特征提取后还得人工设置门限进行分类，在工作量和准确度上有一定的代价，特别是低信噪比下识别率较低，因此使用机器分类的方法。本文采取一种基于核 Logistic 回归的分类方法，该方法较多应用在经济、医学、语音处理等领域。在由核函数扩展而成的希尔伯特空间(Reproducing Kernel Hilbert Space，RKHS)中，相比于 Logistic 回归不再需要样本数和样本维数相等这个条件，便于大规模数据处理[4]。并且相比于语音信号所需较多的特征值，数字信号通过特征提取方法的选取大都可以将特征值控制在低维度。同时通过仿真与支撑向量机(SVM)分类的算法相比较，在分类精确度不低于 SVM 的前提下该分类算法对 4ASK,2PSK,4PSK,2FSK,4FSK,16QAM 这 6 种信号都能较好地处理。

1　高阶累积量的参数提取

假设信号已经过下变频，且载波、相位、定时同步及波形恢复都已经完成。那么经过高斯白噪声污染的复基

收稿日期：2012-04-18；修回日期：2012-05-12

带信号模型为:

$$r(t)=s(t)+n(t)=A\sum_{n}a_n g(t-nT)\exp(\mathrm{j}\theta_c)+n(t) \tag{1}$$

式中: A 为载波幅度; a_n 为码元序列; $g(t)$ 为波形脉冲; θ_c 为载波相位; $n(t)$ 为高斯白噪声。

1.1 高阶累积量理论依据

信号 $r(t)$ 的 p 阶矩定义为:

$$M_{p,q}=E\left\{r^{p-q}(t)[r^{*}(t)]^{q}\right\} \tag{2}$$

式中 $r^{*}(t)$ 是 $r(t)$ 的共轭。信号 $r(t)$ 的 p 阶累积量定义为:

$$C_{p,q}(\tau_1,\tau_2,\cdots,\tau_{p-1})=cum[r(t),r(t+\tau_1),\cdots,r(t+\tau_{p-q-1}),r^{*}(t+\tau_{p-q}),\cdots,r^{*}(t+\tau_{p-1})] \tag{3}$$

对于实信号, 其 p 阶累积量与 q 的取值无关; 而对于复信号, 不同的 q 值可能有不同的 p 阶累积量。

下列矩与累积量的关系式省略累积量表示式中的 τ:

$$\begin{cases} C_{20}=M_{20} \\ C_{21}=M_{21} \\ C_{40}=M_{40}-3M_{20}^{2} \\ C_{41}=M_{41}-3M_{21}M_{20} \\ C_{60}=M_{60}-15M_{40}M_{20}+30(M_{20})^{3} \\ C_{63}=M_{63}-9C_{42}M_{21}-6M_{21}^{3} \end{cases} \tag{4}$$

1.2 调制信号的高阶累积量及特征参数提取

将各个信号经过零均值处理和功率归一化处理后代入各阶累积量的公式中, 计算得到各信号阶累积量的理论值见表 1。

表 1 各信号高阶累积量理论值

Table1 Theoretical value of high-order cumulants of the signals

modulation	$\lvert C_{21}\rvert$	$\lvert C_{40}\rvert$	$\lvert C_{41}\rvert$	$\lvert C_{42}\rvert$	$\lvert C_{60}\rvert$	$\lvert C_{63}\rvert$
4ASK	A^2	$1.36A^4$	$1.36A^4$	$1.36A^4$	$8.32A^6$	$9.16A^6$
2PSK	A^2	$2A^4$	$2A^4$	$2A^4$	0	$4A^6$
4PSK	A^2	A^4	0	A^4	0	$4A^6$
2FSK	A^2	0	0	A^4	0	$4A^6$
4FSK	A^2	0	0	A^4	0	$4A^6$
16QAM	A^2	$0.68A^4$	0	$0.68A^4$	0	$2.08A^6$

根据表 1 定义 2 个参数:

$$T_1=\frac{\lvert C_{40}\rvert}{\lvert C_{42}\rvert} \tag{5}$$

$$T_2=\frac{\lvert C_{63}\rvert^2}{\lvert C_{42}\rvert^3} \tag{6}$$

2 个参数的理论值见表 2。

表 2 两参数理论值

Table2 Theoretical value of the two parameters

modulation	4ASK	2PSK	4PSK	2FSK	4FSK	16QAM
T_1	1.00	1.000	1	0	0	1.00
T_2	33.36	21.125	16	16	16	13.76

经过观察, 2FSK,4FSK 的参数都一样。这里参考文献[5]的方法, 将 2FSK,4FSK 进行微分处理成具有幅度信息的函数, 而后计算 $\lvert C_{21}\rvert$,$\lvert C_{42}\rvert$ 的值[5], 此处将用 $\lvert C_{21\mathrm{dif}}\rvert$,$\lvert C_{42\mathrm{dif}}\rvert$ 来表示。参数 3 定义为:

$$T_3=\frac{\lvert C_{21\mathrm{dif}}\rvert^2}{\lvert C_{42\mathrm{dif}}\rvert} \tag{7}$$

2　核 Logistic 回归分类的设计

若将待分类信号提取出的参数用 $\left\{\left(\boldsymbol{X}_i,y_i\right)\right\}$ 来表示, (其中为了方便描述, 后文 $\boldsymbol{X}_i$ 或 $\boldsymbol{X}$ 均指待分类信号 N 维

特征值向量，y_i 表示所分类别)对于二分类问题的目标值 $y_i \in \{0,1\}$，那么样本 $\boldsymbol{X}$ 属于类别 C_I 的后验概率 $P\left(y=C_I \mid \boldsymbol{X}\right)$，属于 C_J 的后验概率 $P\left(y=C_J \mid \boldsymbol{X}\right)$，其中 $C_I, C_J \in \{0, C\}$。那么对于这两类的判别函数为[6]：

$$f\left(X\right)=\log\frac{P\left(y=C_I \mid \boldsymbol{X}\right)}{P\left(y=C_J \mid \boldsymbol{X}\right)}=\log\frac{P\left(\boldsymbol{X},\boldsymbol{\beta}\right)}{1-P\left(\boldsymbol{X},\boldsymbol{\beta}\right)} \tag{8}$$

式中：$\boldsymbol{\beta}$为样本 $\boldsymbol{X}$ 的权重向量；$P(\bullet)$ 为后验概率。

对上式进行处理可以得到 $y=C_i$ 的后验概率为：

$$\theta=P\left(y=C_I \mid \boldsymbol{X}\right)=\frac{1}{1+\exp\left(-f(\boldsymbol{X})\right)} \tag{9}$$

在 Logistic 回归中，可以用一个线性函数 $f\left(\boldsymbol{X}\right)=\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}+\boldsymbol{\beta}_0$ 对其进行估计。

现在考虑核 Logistic 回归，可以将这个线性模型扩展为一个广义非线性模型。设 $\boldsymbol{X}_i$ 和 $\boldsymbol{X}_j$ 为数据空间中的样本，数据空间到特征空间的映射函数为 $\boldsymbol{\Phi}$，那么实现内积变换表示为 $k\left(\boldsymbol{X}_i,\boldsymbol{X}_j\right)=\boldsymbol{\Phi}\left(\boldsymbol{X}_i\right)\boldsymbol{\Phi}\left(\boldsymbol{X}_j\right)$。所以有：

$$\boldsymbol{\beta}=\sum_{i=1}^{N}\alpha_i\boldsymbol{\Phi}\left(\boldsymbol{X}_i\right) \tag{10}$$

式中 α_i 为对应的参数，那么将线性函数变为：

$$f\left(\boldsymbol{X}\right)=\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Phi}\left(\boldsymbol{X}\right)=(\sum_{i=1}^{N}\alpha_i\boldsymbol{\Phi}(\boldsymbol{X}_i))\boldsymbol{\Phi}(\boldsymbol{X}_j)=\sum_{i=1}^{N}\boldsymbol{\alpha}\boldsymbol{K}\left(\boldsymbol{X}_i,\boldsymbol{X}_j\right) \tag{11}$$

所以 $y=C_I$ 的后验概率化为[4]：

$$\theta_j=P\left(y=C_I \mid \boldsymbol{X}\right)=\frac{1}{1+\exp\left(-\sum_{i=1}^{N}\boldsymbol{\alpha}\boldsymbol{K}\left(\boldsymbol{X}_i,\boldsymbol{X}_j\right)\right)} \tag{12}$$

现在只需要估计出初始 $\boldsymbol{\alpha}^0$ 在经过迭代运算后得到最终的 $\boldsymbol{\alpha}^t$，由经验 Logistic 变换可以得到[7]：

$$\begin{cases}\hat{\alpha}=\overline{z}-\gamma\overline{x}\\ \gamma=S_{xz}/S_{xx}\\ S_{xz}=\sum\left(x_i-\overline{x}\right)\left(z_i-\overline{z}\right)\\ S_{xx}=\sum\left(x_i-\overline{x}\right)^2\\ \overline{z}=\sum\left(In(y_i+\frac{1}{2})-In(n_i-y_i+\frac{1}{2})\right)/N\end{cases} \tag{13}$$

这里将 $\overline{\alpha}$ 进行处理，扩展为计算需要的 $\boldsymbol{\alpha}^0$。后验概率可以写为：

$$\theta_j^t=\frac{1}{1+\exp\left(-\sum_{i=1}^{N}\boldsymbol{\alpha}^t\boldsymbol{K}(\boldsymbol{X}_i,\boldsymbol{X}_j)\right)} \tag{14}$$

那么有：

$$\boldsymbol{\alpha}^{t+1}=\left(\boldsymbol{K}+\lambda(\boldsymbol{W}^t)^{-1}\right)^{-1}\left(\boldsymbol{K}\boldsymbol{\alpha}^t+(\boldsymbol{W}^t)^{-1}(\boldsymbol{y}-\boldsymbol{u}^t)\right) \tag{15}$$

这里 λ 是为了防止训练样本过拟合而加入的惩罚因子，所以核 Logistic 回归以负对数似然为指标，防止 $\boldsymbol{\alpha}$ 出现较大的波动。见式(16)：

$$L(\boldsymbol{\alpha})_{\mathrm{ridge}}=L(\boldsymbol{\alpha})+\frac{\lambda}{2}\|\boldsymbol{\alpha}\|^2 \tag{16}$$

其中
$$\boldsymbol{W}^t = diag\left\{n_1\theta_1^t(1-\theta_1^t), n_2\theta_2^t(1-\theta_2^t), \cdots, n_N\theta_N^t(1-\theta_N^t)\right\} \tag{17}$$

$$\boldsymbol{u}^t = \left\{n_1\theta_1^t, n_2\theta_2^t, \cdots, n_N\theta_N^t\right\}^{\mathrm{T}} \tag{18}$$

$$\boldsymbol{K} = \left\{k(\boldsymbol{X}_i, \boldsymbol{X}_j)\right\}_{N\times N} \tag{19}$$

式中 n_i 为每一对 θ_i 在序列中对应位置。最终在训练完所有训练样本并迭代完要求步数后，可以将后验概率写为：

$$P = \frac{1}{1+\exp\left(-\sum_{i=1}^{N}\boldsymbol{\alpha}^t\boldsymbol{K}(\boldsymbol{X}_i,\boldsymbol{X}_j)\right)} \tag{20}$$

根据后验概率的大小，可以对各待分类信号进行分类。

3　算法流程及仿真分析

3.1 算法流程

首先提取出基于高阶累积量的各信号的 3 个参数 T_1,T_2,T_3。而后将它们组合成为一个样本：

$$\boldsymbol{X} = (T_1, T_2, T_3) \tag{21}$$

因为除去 2FSK,4FSK 其 T_1,T_2 都一样，其他信号的这 2 个参数都不一样，所以除了 2FSK,4FSK 这 2 种信号，其他的信号参数 T_3 人为置为零。这样就构成了有差异且便于用分类器训练的样本。

这时将基于二分树的思想构造其分类顺序，每次分类出其中一类与其余的类，再对其余的类进行同样的步骤。

对于分类顺序进行如下确定，定义两类特征的距离：

$$d_{ij} = \sqrt{(\boldsymbol{X}_i - \boldsymbol{X}_j)^2} \tag{22}$$

将这几种信号的参数做内积处理得到 $e_i = \sqrt{(\boldsymbol{X}_i, \boldsymbol{X}_i)} = \sqrt{(T_1^2 + T_2^2 + T_3^2)}$，将 $e_i(i \leqslant 6)$ 从大到小地排列，第 1 步先取最大 e_i 所对应的 $\boldsymbol{X}_i$ 与其距离最近的 $\boldsymbol{X}_j$(由 d_{ij} 决定)。这样由距离很容易想到可由 $\boldsymbol{X}_i, \boldsymbol{X}_j$ 将整个类分为一类与其余类这 2 类。接着再将其余类看作一个整类，按照相同的思想进行下去，直到将这 6 类信号都分类出结果。等确定好顺序之后，先选取 2 类信号，分别将其状态值赋为 1 和 0。这里将要单独分类出的那类数据状态赋为 1；另外一组数据作为另一大类标准的状态赋为 0。再将状态为 1 的那一类数据取其一定量的样本作为训练数据代入核 Logistic 回归流程，训练出其最终的后验概率函数，而后再由 2 类样本取一定量的样本作为测试数据代入后验概率函数，得出 2 类测试样本各自的后验概率。定义分类的标准：

$$P(y=1 \mid \boldsymbol{X}) > P(y=0 \mid \boldsymbol{X}) \tag{23}$$

通过式(22)就可以分类出单独要分出的那一类数据的识别率。

具体流程如下：

1) 先对样本数据进行预处理，按照上文所述排列出分类顺序。

2) 将某一步骤要单独分类出的样本代入式(13)，通过结果构造出初始的 $\boldsymbol{\alpha}^0$。核函数选取高斯径向基核函数 $K(\boldsymbol{X},\boldsymbol{Y}) = \exp\left\{\dfrac{\left|\boldsymbol{X}-\boldsymbol{Y}\right|^2}{2\delta^2}\right\}$，再由式(14)确定出每一个初始 θ_j^0。

3) 联立式(15)、式(17)、式(18)、式(19)，由上一个状态 $\boldsymbol{\alpha}^t$ 求得下一状态 $\boldsymbol{\alpha}^{t+1}$。(对于初始状态 $t=0$)。

4) 再由下一状态 $\boldsymbol{\alpha}^{t+1}$ 代入式(14)确定出这一时刻状态的每个 θ_j^{t+1}。

5) 重复到步骤 3)，再重复到步骤 4)，依次迭代训练，直到训练完所有的训练样本。

6) 训练完后通过式(20)对测试数据进行分类，由式(23)作为标准来完成识别分类。

7) 完成一次分类后再重复步骤 2)，对下一要分样本进行同样步骤，直到分类识别出所有的类。

3.2 仿真分析

实验 1 待分类信号为：4ASK,2PSK,4PSK,2FSK,4FSK,16QAM。信号通过加性高斯白噪声。采样速率 f_s = 10 000 Hz，载频 f_c =1 000 Hz，符号速率 f_d =1 000 bps。高斯白噪声从 0 dB~25 dB，在每个噪声点上做 1 000 次

蒙特卡罗实验，再取均值作为观察分类标准。

对于本文中所选取的 3 个参数 T_1,T_2,T_3 分别进行仿真，见图 1~图 3。

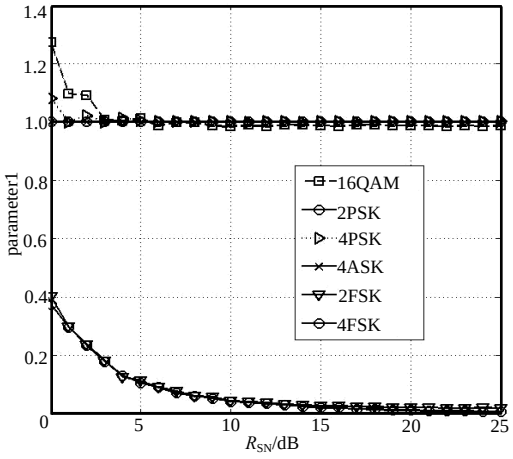

Fig.1 Process of parameter 1 with R_{SN}

图 1 参数 1 随信噪比变化趋势

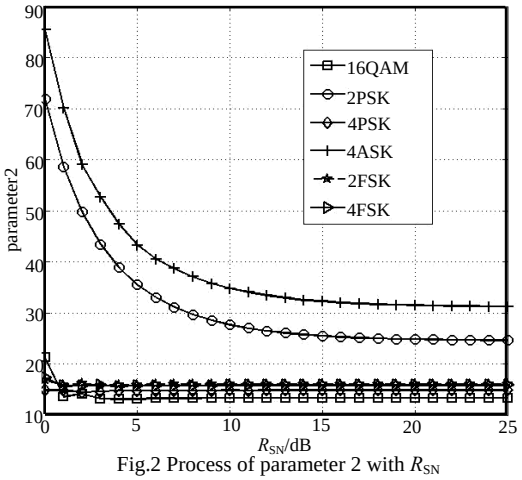

Fig.2 Process of parameter 2 with R_{SN}

图 2 参数 2 随信噪比变化趋势

通过上述 3 个参数构建上文所述的样本 $\boldsymbol{X}$，再进行下一步的运算。

设置核 Logistic 回归与 SVM 的各个参数，两者都采用高斯径向基核函数。这里 SVM 的惩罚因子 C 的选取要适中，它实际上控制对错分样本的惩罚程度，是在错分样本的比例与算法复杂度之间的折中[8]。由于样本的维度不高，所以这里不再用 SVM 探讨 C 与 δ^2 的最优取值问题，核 Logistic 回归也一样。C 选取 500，两者的核函数的 δ^2 选取 0.25。核 Logistic 回归里 λ 为平衡回归函数光滑性与损失函数的系数，选取 λ 为 0.5。

实验 2 用 100 个样本作为训练样本，200 个样本作为分类测试样本。从 0 到 25 dB，每隔 5 dB 做一次实验，SVM 参考文献[9]并采用本文的分类流程[9]。用 2 类方法的详细分类结果见表 3。

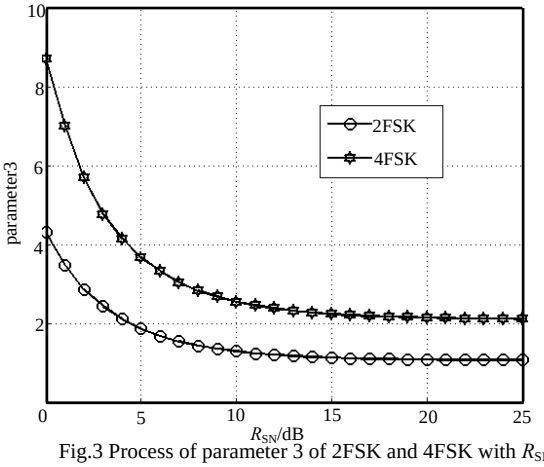

Fig.3 Process of parameter 3 of 2FSK and 4FSK with R_{SN}

图 3 2FSK 与 4FSK 信号参数 3 随信噪比变化趋势

表 3 采用 SVM 与 KLR 的识别率

Table3 Correct recognition rate based on SVM and KLR

signals		correct recognition rate/%				
		0 dB	5 dB	10 dB	15 dB	20 dB
4ASK	SVM	98.5	99	100	100	100
	KLR	100	100	100	100	100
2PSK	SVM	80	95.3	100	100	100
	KLR	88	97	100	100	100
16QAM	SVM	49	87.5	88	100	100
	KLR	52	92	100	100	100
4PSK	SVM	88	100	100	100	100
	KLR	94	99.7	100	100	100
2FSK	SVM	98	100	100	100	100
	KLR	100	99	100	100	100
4FSK	SVM	97.7	99.5	100	100	100
	KLR	100	99.8	100	100	100

定义总识别率为：同一信噪比时，6 种信号识别正确的样本数叠加除以总的样本数，如图 4 所示。

由表 3 和图 4 可得，在同样的训练样本与测试样本的情况下，采用核 Logistic 回归的分类算法的识别率要高于 SVM，在低信噪比 0 dB 时，核 Logistic 回归算法的性能更能够体现出来。

实验 3 本实验是为比较 2 种算法的计算量复杂度，进行仿真计算机配置为 AMD Althon 处理器，3.01 GHz 主频；内存 2 GB；操作系统 Windows XP；仿真工具 Matlab2010a。这里选取 0 dB 时对第 1 次分类的 2 类样本的时间进行运算，SVM 的运行时间为 0.076 9 s，核 Logistic 回归的运行时间为 0.120 4 s。由此得出，利用核 Logistic

回归算法的计算量要高于 SVM。这是由于核 Logistic 回归中其解的稀疏性不存在，当数据量较大时，限制了其应用，造成计算量较大[10]。

4　结论

本文提取了待分类信号的高阶累积量参数，由参数自行构造了一种便于计算与分类的向量，综合核 Logistic 回归(KLR)的思想设计了一种基于决策树思想的 KLR 分类器。仿真证明了 KLR 算法相比于 SVM 的有效性。但本文算法仅在高斯白噪声信道下进行，实际信道环境比较复杂，并且虽然 KLR 较 Logistic 回归(LR)对大样本数据处理起来较方便，但由于计算量的问题导致其处理时间过长，如何将本算法在实际环境推广及降低 KLR 的运算量是今后的一个研究方向。

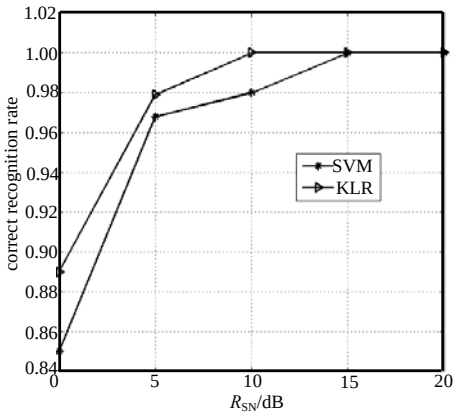

Fig.4 Correct recognition rate in all based on SVM and KLR

图 4 SVM 与 KLR 总的识别率

参考文献：

[1] 董彬虹,李少谦. 短波通信的现状及发展趋势[J]. 信息与电子工程, 2007,5(1):1–5. (DONG Binhong,LI Shaoqian. Current Status and Developing Tendency for High Frequency Communications[J]. Information and Electronic Engineering, 2007,5(1):1–5.)

[2] Ananthram Swami,Brian M Sadler. Hierarchical Digital Modulation Classification Using Cumulants[J]. IEEE Trans. on Communications, 2000,48(3):416–429.

[3] 王甲峰,岳旸,姚军. 基于高阶累积量的 MPSK 识别方法[J]. 通信技术, 2010,43(9):4–6. (WANG Jiafeng,YUE Yang, YAO Jun. A MPSK Recognition Method Based on High Order Cumulants[J]. Communications Technology, 2010,43(9):4–6.)

[4] Peter Karsmakers,Kristiaan Pelckmans,Johan Suykens A K. Multi-class Kernel Logistic Regression:a fixed-size implementation[C]// Proceedings of International Joint Conference on Neural Networks. Orlando,Florida:[s.n.], 2007:1751–1756.

[5] 吕正新,魏平,肖先赐. 利用高阶累积量实现数字调制信号的自动识别[J]. 电子对抗技术, 2004,19(6):3–6. (LV Zhengxin, WEI Ping,XIAO Xianci. Automatic Identification of Digital Modulation Signals Using High Order Cumulants[J]. Electronic Countermeasures, 2004,19(6):3–6.)

[6] Rahayu S P,Purnami S W,Embong A. Applying Kernel Logistic Regression in Data Mining to Classify Credit Risk[C]// Information Technology. Malaysia:[s.n.], 2008:1–6

[7] 史宁中. 统计检验的理论与方法[M]. 北京:科学出版社, 2008. (SHI Ningzhong. The Theory and Method of Statistics and Test[M]. Beijing:science press, 2008.)

[8] 崔和,龙玉峰. 支持向量机学习算法的研究现状与展望[J]. 信息与电子工程, 2008,6(5):328–332. (CUI He,LONG Yufeng. Status Quo and Expectation of Support Vector Machine Learning Algorithm[J]. Information and Electronic Engineering, 2008,6(5):328–332.)

[9] 闫朋展,王振宇. 运用高阶累积量和SVM的调制自动识别[J]. 电讯技术, 2010,50(10):36–40. (YAN Pengzhan,WANG Zhenyu. Automatic Recognition of Digital Modulation Signals by Applying High Order Cumulants and Support Vector Machine[J]. Telecommunication Engineering, 2010,50(10):36–40.)

[10] 刘遵雄,许金凤,曾丽辉. 基于核 Logistic 回归的乐器音乐辨识[J]. 华东交通大学学报, 2010,27(4):29–33. (LIU Zunxiong, XU Jinfeng,ZENG Lihui. Musical Instrument Audio Identification Based on Kernel Logistic Regression[J]. Journal of East China Jiaotong University, 2010,27(4):29–33.)

作者简介：

徐　闻(1987–)，男，成都市人，在读硕士研究生，主要研究方向为信号调制分析、模式识别.email:emonewxw@163.com.

王　斌(1969–)，男，河南省新乡市人，教授，主要研究方向为通信中的现代信号处理.