

基于遗传算法优化支持向量机的网络流量预测

张颖璐

(中国船舶重工集团公司第 710 研究所 宜昌 443003)

摘要 介绍了支持向量机用于时间序列预测的理论基础和遗传算法优化支持向量机参数的方法,首次把遗传算法优化参数支持向量机应用于两组实际网络流量的预测,并与 BP 神经网络和 RBF 神经网络方法进行了比较。结果表明:支持向量机相比较 BP 神经网络和 RBF 神经网络对网络流量的预测结果精度更高、性能更好。利用支持向量机预测网络流量是一种可行、有效的方法。

关键词 遗传算法,支持向量机,网络流量,预测,神经网络

Internet Traffic Forecasting Based on Support Vector Machine Optimized by Genetic Algorithm

ZHANG Ying-lu

(No. 710 Research Institute, China Shipbuilding Industry Corporation, Yichang 443003, China)

Abstract The basic theory of time series forecasting based on Support Vector Machine (SVM) is introduced in this paper. And Genetic Algorithm (GA) optimizes the parameters of SVM. GA-SVM is firstly applied to forecast future Internet traffic including two sets of real data, and compared to the BP and RBF neural network. The results show that SVM is superior to these two kinds of neural network methods in prediction performance. And SVM is the suitable and effective method for forecasting Internet traffic.

Keywords Genetic algorithm, Support vector machine, Internet traffic, Forecasting, Neural network

1 引言

网络流量预测是规划、设计、控制、管理和优化网络上各种资源的一个重要前提,对于网络拥塞控制、路由选择等起着至关重要的作用^[1-4]。在过去的几十年里,网络流量预测算法一直是国内外众多研究学者关注的热点,先后提出了基于 ARMA^[1], FARIMA^[2] 模型的线性预测方法和基于 BP 神经网络^[3], RBF 神经网络^[4] 的非线性预测方法等多种预测算法。其中神经网络的预测方法由于能够刻画出网络流量的非线性特征而显示出比线性预测方法更优异的性能。但是,在实践中神经网络容易出现过学习,即网络的推广能力较差的缺陷。因此要求提出更先进的预测算法,进一步提高网络流量预测的精度和效率。近年来,针对网络流量混沌特性所取得的研究成果表明网络流量存在一定的混沌特性^[4,5],可以通过一个非线性动力学系统近似,这就启发我们应用更先进的机器学习算法——支持向量机进行网络流量的预测。

支持向量机是一种新型的机器学习方法^[6],是近年来机器学习领域的研究热点^[7,8]。该学习方法以结构风险最小化原则取代传统机器学习方法中的经验风险最小化原则,具有很好的泛化能力,在小样本的机器学习中显示出了优异的性能。支持向量机克服了神经网络方法固有的局部极小、过学习以及结构和类型的选择过分依赖经验等缺陷,已经被广泛应用于非线性时间序列预测中,并取得了良好的效果^[9,10]。但目前对时间序列预测的应用仍然存在诸多问题:其一是应用多集中于理想的混沌时间序列,对实际的时间序列(如网络流量预测)的研究仍然较少;其二是支持向量机预测时间序列的性能对参数的选择敏感,但目前对支持向量机参数优化问

题的研究也相对较少^[11],实际应用中大多凭经验确定参数。因此,结合优化算法(如遗传算法)确定支持向量机的参数^[12],应用遗传算法优化支持向量机对实际网络流量预测,不仅可以对网络流量预测的进一步发展提供新方法,而且对支持向量机在实际中的进一步应用具有很重要的参考价值。

本文首次把遗传算法优化支持向量机引入到网络流量预测的研究,应用支持向量机对两组网络流量实测数据进行了学习、预测,并与 BP 和 RBF 神经网络方法的预测结果进行了比较。

2 支持向量机回归

设一个给定长度为 n 的离散时间序列 $\{x_t, t=1, 2, \dots, n\}$ 是未知的非线性动力学系统,根据 Takens 嵌入定理,对满足一定条件的嵌入维 m 和几乎所有的延迟时间 τ 存在一个光滑映射: $f: R^m \rightarrow R$, 使得

$$x_{t+m} = f(x_t, x_{t+\tau}, \dots, x_{t+(m-1)\tau}) \quad (1)$$

其中,使式(1)成立的最小的 m 称为最小嵌入维。时间序列的预测问题就是通过对输入输出样本集 $\{X, Y\} \subset (R^m \times R)$ 的训练估计映射 f , 进而得出 x_{n+1} 的预测值。其中

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_{n-m\tau} \end{bmatrix} = \begin{bmatrix} x_1 & x_{1+\tau} & \cdots & x_{1+(m-1)\tau} \\ x_2 & x_{2+\tau} & \cdots & x_{2+(m-1)\tau} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n-m\tau} & x_{n-m\tau+\tau} & \cdots & x_{n-m\tau+(m-1)\tau} \end{bmatrix}$$

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_{n-m\tau} \end{bmatrix} = \begin{bmatrix} x_{1+m} \\ x_{2+m} \\ \vdots \\ x_{n-m\tau+m} \end{bmatrix} \quad (2)$$

支持向量机把上述预测问题表述成一个回归问题,其基本思想就是通过一个非线性映射 Φ 把输入样本集 X 映射到一个高维特征空间 F 进行线性回归,从而将低维特征空间的非线性回归问题转化成高维特征空间的线性问题解决。根据统计学习理论确定如下回归函数:

$$f(X) = [w \cdot \Phi(X)] + b \quad (3)$$

其中, $\langle \cdot \rangle$ 表示内积, Φ 表示从 R^m 空间到 F 空间的非线性映射, b 表示偏置量。根据结构风险最小化原则,上述函数回归问题就是寻求使下面风险函数最小的 f :

$$R_{\text{reg}} = \frac{1}{2} \|w\|^2 + CR_{\text{emp}}[f] \quad (4)$$

其中, $\|w\|^2 = \langle w, w \rangle$ 是描述函数 f 复杂度的项, $R_{\text{emp}}[f] = \frac{1}{l} \sum_{i=1}^l |Y_i - f(X_i)|$ 是经验风险, C 是惩罚因子,其作用是在经验风险和模型复杂度之间取一折中。 ϵ 是引入的不敏感损失函数,具体定义如下:

$$|Y - f(X)|_{\epsilon} = \begin{cases} |Y - f(X)| - \epsilon, & |Y - f(X)| \geq \epsilon \\ 0, & |Y - f(X)| \leq \epsilon \end{cases} \quad (5)$$

求式(4)的最小化等价于最小化泛函:

$$\min \frac{1}{2} \|w\|^2 + C \sum_{i=1}^l (\xi_i + \xi_i^*)$$

$$s. \begin{cases} Y_i - \langle w \cdot \Phi(X_i) \rangle - b \leq \epsilon + \xi_i \\ \langle w \cdot \Phi(X_i) \rangle + b - Y_i \leq \epsilon + \xi_i^* \\ \xi_i \geq 0, \xi_i^* \geq 0 \end{cases} \quad (6)$$

其中, $l = n - m\tau$ 是样本的个数, ξ_i 和 ξ_i^* 是引进的松弛变量。对于解上述二次优化问题寻找向量 w , 其核心思想是用拉格朗日算子转化为其对偶形式求解。原始的拉格朗日函数为

$$L = \frac{1}{2} \|w\|^2 + C \sum_{i=1}^l (\xi_i + \xi_i^*) - \sum_{i=1}^l \alpha_i [\epsilon + \xi_i + Y_i - (\langle w \cdot \Phi(X_i) \rangle + b)] - \sum_{i=1}^l \alpha_i^* [\epsilon + \xi_i^* - Y_i + (\langle w \cdot \Phi(X_i) \rangle + b)] - \sum_{i=1}^l (\eta_i \xi_i + \eta_i^* \xi_i^*) \quad (7)$$

其中, $\alpha_i, \alpha_i^*, \eta_i, \eta_i^*$ 是拉格朗日算子。对式(6)的最小化应对 w, b, ξ_i, ξ_i^* 的偏导为 0, 即

$$\begin{cases} \frac{\partial L}{\partial w} = w - \sum_{i=1}^l (\alpha_i - \alpha_i^*) \cdot \Phi(X_i) = 0 \\ \frac{\partial L}{\partial b} = \sum_{i=1}^l (\alpha_i - \alpha_i^*) = 0 \\ \frac{\partial L}{\partial \xi_i} = C - \alpha_i - \eta_i = 0 \\ \frac{\partial L}{\partial \xi_i^*} = C - \alpha_i^* - \eta_i^* = 0 \end{cases} \quad (8)$$

把式(8)带入式(7),就得到式(6)的对偶形式:

$$\max W(\alpha, \alpha^*) = -\frac{1}{2} \sum_{i,j=1}^l (\alpha_i - \alpha_i^*) (\alpha_j - \alpha_j^*) K(X_i, X_j) - \sum_{i=1}^l \epsilon (\alpha_i + \alpha_i^*) + \sum_{i=1}^l Y_i (\alpha_i - \alpha_i^*)$$

$$s. t. \sum_{i=1}^l (\alpha_i - \alpha_i^*) = 0, \alpha_i \in [0, C], \alpha_i^* \in [0, C] \quad (9)$$

其中, $K(x, y) = \langle \Phi(x) \cdot \Phi(y) \rangle$ 是支持向量机的核函数。求解式(9),可得

$$w = \sum_{i=1}^l (\alpha_i - \alpha_i^*) \Phi(X_i) \quad (10)$$

根据以上的推导可得回归函数

$$f(X) = \sum_{i=1}^l (\alpha_i - \alpha_i^*) K(X, X_i) + b \quad (11)$$

其中, 当 $\alpha_i - \alpha_i^*$ 非零时对应的训练样本就是支持向量, b 可

以利用 KKT(Karush-Kuhn-Tucker)条件计算出:

$$\begin{cases} b = Y_i - \epsilon - \langle w \cdot \Phi(X_i) \rangle, \alpha_i, \alpha_i^* \in [0, C] \\ b = Y_i + \epsilon - \langle w \cdot \Phi(X_i) \rangle, \alpha_i, \alpha_i^* \in [0, C] \end{cases} \quad (12)$$

核函数 $K(x, y)$ 可以是任何一个满足 Mercer 条件的正定函数, 几类常用的核函数有:

多项式核函数: $K(x, y) = (x \cdot y + b)^d, b \geq 0, d \in \mathbb{N}$

高斯核函数: $K(x, y) = e^{-\frac{\|x-y\|^2}{\sigma^2}}, \sigma > 0$

双曲核函数: $K(x, y) = \tanh[a(x \cdot y) + c], a, c$ 是常数。

3 遗传算法优化支持向量机参数

支持向量机时间序列预测结果的精度与嵌入维数 m 、惩罚因子 C 、损失函数参数 ϵ 、选择的核函数以及核函数参数之间存在一定的关系^[12]。因此,应用先进的优化算法在一定的先验区间内搜索支持向量机各参数的最优结果,进而获得具有最佳预测性能的支持向量机,对实际时间序列的预测具有重要意义。遗传算法是一种模拟生物进化的智能优化算法,是一种具有广泛适用性的搜索方法,具有很强的全局优化搜索能力。本文选择利用遗传算法实现支持向量机的参数优化。下面以高斯核函数为例简要说明参数优化的具体步骤。

Step 1: 确定参数 m, C, ϵ 和高斯核函数参数 σ 在实际应用中的可能取值范围;

Step 2: 定义适应度函数,本文选择时间序列预测值 $\bar{y}$ 与实测值 y 之间的相对均方误差 $[\sum_{i=1}^K (\bar{y}_i - y_i)^2] \times (\sum_{i=1}^K y_i^2)^{-1}$ 作为适应度函数。

Step 3: 对上述参数进行二进制编码,并随机产生初始群体;

Step 4: 对群体中的染色体解码,计算各个个体的适应度函数;

Step 5: 判断群体性能是否满足某一优化准则(本文选择完成最大遗传代数作为优化准则),若满足则输出最优参数,否则按照遗传策略,运用选择、交叉和变异算子作用于群体,产生下一代群体,并开始新一代的遗传。

这样,结合具体实际问题适当设置遗传算法的参数,通过上述步骤就可实现对支持向量机的参数优化。下面我们将遗传算法优化支持向量机应用到两组网络流量预测的实例,并与 BP 和 RBF 神经网络预测结果进行比较。

4 网络流量预测实例

本文采用贝尔实验室提供的两组网络流量数据作为预测实例,其中数据 A 长度为 2700 个,表示 2700s 的网络流量;数据 B 长度为 1700 个,表示 1700s 的网络流量,其原始序列见图 1。

具体的网络流量预测步骤如下。

Step 1: 分别归一化原始数据 A 和 B 到 $[0, 1]$ 的区间范围之内;

Step 2: 根据 Takens 定理对归一化数据重构相空间,即建立如式(2)的矩阵形式;

Step 3: 对数据 A 和 B 分别取前 2500 和 1500 个归一化数据构成的相空间相点为训练样本;

Step 4: 通过对所建训练样本的学习,获得网络流量序列的非线性关系;

Step 5: 截取后 200 个相点为预测样本,分别采用 BP 神经

网络、RBF 神经网络和支持向量机进行预测,并以预测值 $\bar{y}$ 与实测值 y 之间的相对均方误差 $[\sum_{i=1}^K (\bar{y}_i - y_i)^2] \times (\sum_{i=1}^K y_i^2)^{-1}$ 和相关系数 $[\sum_{i=1}^K (y_i - \bar{y}_i)^2] \times [\sum_{i=1}^K (y_i - \bar{y}_i)^2 \sum_{i=1}^K (\bar{y}_i - \bar{y}_i)^2]^{-1/2}$ 作为预测指标,比较 3 种情况下的预测结果。

根据上述预测步骤,下面分别将 BP 神经网络、RBF 神经网络和支持向量机应用到数据 A 和 B 的网络流量预测。其中 BP 神经网络隐含层神经元个数为 35,采取的是多次试验确定的方法。隐含层传递函数采用 tansig 函数,输出层传递函数采用线性函数,允许误差为 0.01,训练次数 2000 次,网络的初始化由计算机随机完成,学习函数选用 Levenbeger-Marquardt 法。该算法是梯度下降法的改进,具有收敛速度快的优点。RBF 神经网络采用 10×1 的输入输出层结构,隐含层神经元采用自适应调节,选择允许误差为 0.01。支持向量机选用高斯核函数,数据 A 对应的嵌入维数 $m=10$,核函数参数 $\sigma=1$,惩罚因子 $C=1$,损失函数参数 $\epsilon=0.001$;数据 B 对应的 $m=11$, $\sigma=1$, $C=1$, $\epsilon=0.01$ 。上述支持向量机参数的选择均是利用遗传算法通过对支持向量机的自适应优化的结果。遗传算法的参数分别设为:交叉和变异概率分别为 0.8 和 0.08,初始群体 20,进化代数为 100。各参数的搜索区间充分考虑了每一个参数在实际应用中的可能取值范围,选取 $m \in [1, 30]$, $\sigma \in [0.001, 1000]$, $C \in [1, 1000]$, $\epsilon \in [0.0001, 0.1]$ 。

表 1 BP 神经网络、RBF 神经网络和支持向量机预测网络流量结果比较

数 据	BP 神经网络		RBF 神经网络		支持向量机	
	相对 均方 误差	相关 系数	相对 均方 误差	相关 系数	相对 均方 误差	相关 系数
A	1.1825	0.8867	1.0618	0.8926	0.1271	0.9368
B	0.2288	0.9605	0.2352	0.9604	0.0404	0.9796

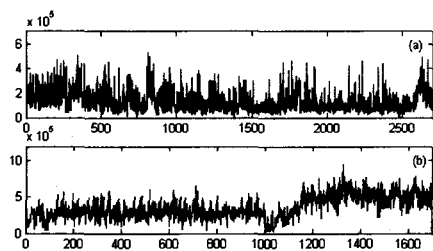


图 1 数据 A(a)和数据 B(b)的网络流量原始序列

应用上述 3 种方法预测网络流量数据 A 和 B 结果的性能指标见表 1,网络流量预测值与实测值的比较分别见图 2 和图 3。注意,由于 BP 神经网络初始值是随机的,因此每次运行的结果并不相同,本文选择的是多次运行以后的最优结果。从表 1 和图 2、3 中可以看出:支持向量机预测网络流量的结果相比较 BP 和 RBF 神经网络而言,其精度更高,效果更好。其中数据 A 预测结果的相对均方误差在支持向量机情况下约为 0.1271,而 BP 和 RBF 神经网络情况下相对均方误差分别为 1.1824 和 1.0618,几乎相差了一个量级。同时,支持向量机预测结果的相关系数也比两种神经网络预测结果的相关系数更大,说明预测的效果更好。同样地,支持向量机比 BP 和 RBF 神经网络对数据 B 预测结果的相对均方误差也有

约 5 倍的降低,相关系数也相对较大。

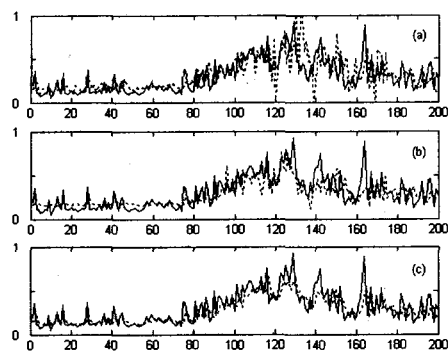


图 2 BP 神经网络(a)、RBF 神经网络(b)和支持向量机(c)的数据 A 网络流量预测值和实测值的比较
其中实线是实测值,虚线是预测值。

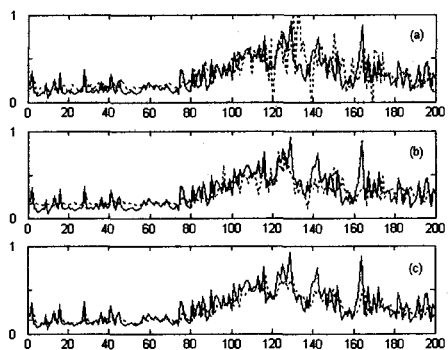


图 3 BP 神经网络(a)、RBF 神经网络(b)和支持向量机(c)的数据 B 网络流量预测值和实测值的比较
其中实线是实测值,虚线是预测值。

结束语 本文应用遗传算法优化支持向量机方法对网络流量数据进行了预测,并与 BP 和 RBF 神经网络方法相比较。结果表明:以结构风险最小化原则为基础的支持向量机克服了神经网络方法中易陷入局部极小值、精度和泛化之间不可调和的矛盾,在网络流量的预测中显示了优异的性能。相比较 BP 神经网络和 RBF 神经网络方法,支持向量机方法对两组网络流量实例的处理结果相对均方误差更小、相关系数更大、预测精度更高、效果更好,在网络流量的预测研究中具有很好的应用前景。

致谢 感谢为本文提供数据的贝尔实验室。

参考文献

- [1] 邹伯贤,刘强. 基于 ARMA 模型的网络流量预测[J]. 计算机研究与发展,2002,39(12): 1645-1652
- [2] Shu Y, Jin Z, Zhang L, et al. Traffic prediction using FARI-MA models[A]// IEEE International Conference on Communications[A]. Vancouver, Canada, 1999:891-895
- [3] 刘芹,余一娇,谭连生. 一种利用 BP 神经网络的 Internet 流量预测算法[A]//2003 中国计算机大会会议论文集. 北京:清华大学出版社,2003:1137-1142
- [4] 陆锦军,王执铨. 基于混沌理论的网络数据流 RBF 神经网络预测[J]. 计算机工程,2006,32(23):100-103

(下转第 197 页)

帧包含 5 个脉冲,需要 5bit 描述脉冲符号。

在实验中分别修改第一子帧和第二子帧中所有脉冲的极性,讨论语音质量受到影响的程度。

表 6 脉冲符号参数隐藏 1bit 数据的语音效果

第一子帧修改脉冲符号	SNR	SSNR	SD	第二子帧修改脉冲符号	SNR	SSNR	SD
第一位	5.239	7.952	0.200	第一位	7.141	7.360	0.214
第二位	6.008	7.888	0.210	第二位	6.194	7.366	0.221
第三位	8.334	8.474	0.207	第三位	7.758	7.824	0.213
第四位	8.538	8.686	0.196	第四位	9.350	8.463	0.186
第五位	9.865	9.199	0.190	第五位	10.287	9.009	0.170
第六位	10.073	9.403	0.184	第六位	10.073	9.403	0.184

实验结果如表 6 所示,可以发现数据存在一定的规律性。在同一子帧中,保密数据嵌入的符号位置越高,语音还原质量越好,而嵌入奇数子帧或偶数子帧同一位置的语音之间差别并不显著。主观感觉与之一致,在低位嵌入的语音无论噪声还是失真都比在高位嵌入的语音更严重。

出现这种现象的原因是,用来模拟残差信号的脉冲是依次单独进行计算的,首先寻找和原始残差信号相关性最高的第一个脉冲,并把相关信息保存在最低的位置,其次计算原始残差信号与第一个脉冲的区别,再寻找与差值相关性最高的第二个脉冲,并把相关信息保存在稍高的位置,然后计算原始残差信号与第一个脉冲、第二个脉冲的区别,如此反复直至得出算法规定的奇数子帧的 6 个脉冲、偶数子帧的 5 个脉冲为止。与残差信号相关性较低的脉冲信息放在高位,影响语音效果较小可以尝试进行信息隐藏,具体来说奇数子帧的第五位和第六位与偶数子帧的第四位和第五位。

按照从最高位开始、最多嵌入 2bit 的指导思想,每帧最多能隐藏 8bit 保密数据。比较利用不同子帧的脉冲符号高位隐藏不同信息量的保密数据对语音质量的影响。

表 7 脉冲符号参数隐藏 2-4bit 数据语音效果

子帧脉冲符号高位修改				SNR	SSNR	SD
一	二	三	四			
2	0	0	0	7.083	6.574	0.230
1	1	0	0	6.691	5.863	0.255
2	2	0	0	3.916	2.637	0.327
1	1	1	1	3.348	2.526	0.362

(上接第 179 页)

- [5] 陆锦军,王执铨. 基于混沌特性的网络流量预测[J]. 南京航空航天大学学报,2006,38(2):217-221
- [6] Vapnik V N. The Nature of Statistical Learning Theory [M]. New York: Springer-Verlag, 1995
- [7] 王晓丹,郑春颖,吴崇明. 一种新的 SVM 对等增量学习算法. 计算机应用,2006,26(10):2440-2443
- [8] 吴少雄,黄恩洲. 基于支持向量机的控制图模式识别. 计算机应用,2007,27(1):61-64

表 7 实验数据表明,无论数据分散隐藏到前半帧和后半帧还是集中隐藏到前半帧或后半帧,解码语音效果都不理想,隐藏信息量略微增加信噪比和分段信噪比迅速减小。收听感觉语音时常发生失真,强烈干扰使细节含混不清,机械感让听觉不舒适。

虽然在 G. 723.1 编码结果中脉冲符号占有超过 10% 的份额,但即使隐藏少量数据也会造成语音发生可感知的变化,令通信过程失去了最基本的保密性,因此不使用脉冲符号进行信息隐藏。

结束语 本文基于信息隐藏及其特点,尝试讨论一种在语音混沌保密通信过程中进行信息隐藏方案中,对 G. 723.1 高码率编码语音中适合进行压缩域信息隐藏的码位进行进一步的性能分析。提出了具体的实验方案。并综合实验结果,6.3kbps 算法中适合信息隐藏的位置有数据类型位、线谱对参数索引、脉冲平移、保留位、脉冲位置,最多分别可以隐藏 2bit,9 bit,4 bit,1 bit,8 bit 保密数据。

嵌入数据类型位、保留位无损语音质量,隐藏信息量虽小却能共同使用略作弥补,因容易被察觉,适用于隐藏非关键数据;嵌入线谱对参数索引、脉冲平移、脉冲位置通过牺牲部分效果提高安全性能,但由于对语音有破坏,不能叠加使用以免累积误差,适用于隐藏关键数据。实验结果进一步证明了文献[1]方案的有效性和安全性。

参考文献

- [1] 何招友,季晓勇. 基于混沌理论的语音保密通信技术. 信息安全与通信保密,2006(7):68-70
- [2] 倪皖荪,华一满. 混沌通讯. 物理学进展,1996,16(3-4):645-655
- [3] 代明,张宇,华一满,倪皖荪. 超混沌迭代同步在数字语音保密通讯中的应用. 南京大学学报(自然科学),1999,35:110-113
- [4] 陆佩芸,许留荣,季晓勇. 基于 DSP 的嵌入式语音保密通信模块. 微处理机,2003(10)
- [5] ITU-T Recommendation G. 723.1: DUAL RATE SPEECH CODER FOR MULTIMEDIA COMMUNICATIONS TRANSMITTING AT 5.3 AND 6.3 kbit/s
- [6] 张雄伟,曹铁勇. DSP 芯片的原理与开发应用. 北京:电子工业出版社,2000
- [7] 韩纪庆,张磊,郑铁然. 语音信号处理. 北京:清华大学出版社,2004
- [8] 蔡莲红,黄德智,蔡锐. 现代语音技术基础与应用. 北京:清华大学出版社,2003
- [9] 苏晓生. 掌握 MATLAB 6.0 及其工程应用. 北京:科学出版社,2002
- [9] 董辉,付鹤林,冷伍明. 基于支持向量机的时间序列回归与预测[J]. 系统仿真学报,2006,18(7):1785-1788
- [10] 刘涵,刘丁,李琦. 基于支持向量机的混沌时间序列非线性预测[J]. 系统工程理论与实践,2005,9:94-99
- [11] 周红刚,杨春德. 基于免疫算法与支持向量机的异常检测方法[J]. 计算机应用,2006,26(9):2145-2147
- [12] 杨虞微,左洪福,陈果. 支持向量机时间序列预测模型的参数影响分析与自适应优化[J]. 航空动力学报,2006,21(4):767-772