基于选择性支持向量机集成的海杂波背景中的微弱信号检测*

行鸿彦† 祁峥东 徐伟

(南京信息工程大学, 江苏省气象探测与信息处理重点实验室, 南京 210044)

(南京信息工程大学电子与信息工程学院, 南京 210044)

(2012 年 5 月 23 日收到; 2012 年 6 月 23 日收到修改稿)

基于复杂非线性系统相空间重构理论, 提出了一种混沌背景中微弱信号检测的选择性支持向量机集成的方法, 为了提高支持向量机集成的泛化能力, 采用 K 均值聚类算法选择每簇中精度最高的子支持向量机进行集成, 建立了混沌背景噪声的一步预测模型, 从预测误差中检测湮没在混沌背景噪声中的微弱目标信号 (包括周期信号和瞬态信号), 最后分别以 Lorenz 系统和实测的 IPIX 雷达数据作为混沌背景噪声进行实验研究, 结果表明该方法能够有效地将混沌背景噪声中极其微弱的信号检测出来, 抑制噪声对混沌背景信号的影响, 与神经网络和传统支持向量机方法相比, 预测精度和检测门限方面的性能有显著提高.

关键词: 支持向量机, 集成, 海杂波, 微弱信号检测

PACS: 05.45.Pq

1 引 言

混沌是由确定模型产生的不确定现象, 混沌系统在生活中普遍存在, 如雷达海杂波信号、舰船辐射信号等, 而海杂波背景中的微弱信号检测一直是信号检测的难点之一.

文献 [1] 提出利用信号的混沌背景这一先验知识, 建立非线性预测模型, 在预测误差中检测信号. 国内外已将多种方法应用于预测混沌时间序列, 如自适应非线性滤波预测法 [2]、神经网络法 [3,4]、支持向量机法 [5] 等, 其中人工神经网络因其较强的非线性映射能力被广泛应用于预测混沌时间序列, 但神经网络存在过学习, 容易陷入局部最优, 隐层和隐层节点数的选择过分依赖于经验以及维数灾难等固有缺陷, 导致学习精度与可靠性无法得到保证. 支持向量机 (SVM) 理论基础是 Vapnik 提出的统计学习理论, 采用结构风险最小化原则, 在最小化样本点误差的同时缩小模型泛化误差的上界, 即最小化模型的结构风险, 具有维数不敏感、泛化能力好、全局最优等优点, 在小样本学习中更为突出. SVM 集成学习方法属于同质类型的学习算法, 利用多个子 SVM 解决同一个问题, 传统的 SVM 集成学习方法将所有子 SVM 集成进行平均或加权输出, 存在计算量大, 效率低等缺点. 2002 年, Zhou 等 [6] 提出了选择性集成学习的方法, 剔出作用不大, 性能不好的网络, 选择部分精度高, 差异度大的子网络进行集成, 可以使集成在任何情况下达到或超过组成它的各个子网络的平均性能, 因此将支持向量机集成模型应用于混沌背景下的微弱信号检测, 并利用成熟的信号检测技术去除背景信号, 完成对微弱信号的提取.

2 混沌时间序列的相空间重构理论

在实际非线性系统中得到一个时间间隔为 Δt

* 国家自然科学基金 (批准号: 61072133) 和江苏省 "传感网与现代气象装备" 优势学科平台资助的课题.

† E-mail: xinghy@nuist.edu.cn

的单变量时间序列,

$$\{x_1, x_2, \cdots, x_N\}, \quad \text{其中} x_j = x(t_j),$$
$$t_j = t_0 + j\Delta t, \quad j = 1, 2, \cdots, N. \tag{1}$$

混沌动力学系统中任一分量的演化都是由与之相互作用着的其他分量所决定, 因此任意分量的发展过程都包含其他分量的发展信息, 基于相空间重构的时间延迟坐标法与支持向量机的基本思想相似, 将输入空间的向量扩展到高维空间, 发掘系统蕴藏的信息与规律, 重构出原动力系统模型, 对于单一时间序列 (1), 如果其嵌入维为 $D_{\rm E}$, 重构时间延迟为 τ, 则重构出的 $N_{\rm m}$ 个 $D_{\rm E}$ 维矢量为

$$Y_{\rm J} = [x_j, x_{j+\tau}, x_{j+2\tau}, x_{j+(D_{\rm E}-1)\tau}],$$
$$j = 1, 2, \cdots, N_{\rm m}, \quad N_{\rm m} = N - (D_{\rm E} - 1)\tau. \tag{2}$$

这 $N_{\rm m}$ 个 $D_{\rm E}$ 维矢量在 $D_{\rm E}$ 维描述出的轨迹可将混沌吸引子完全展开, 在拓扑等价的意义下恢复原来系统的动力学性质关于嵌入维数 $D_{\rm E}$ 和时延 τ 的选取方法目前存在两种观点, 一种认为两个参数互不相关, 如求 $D_{\rm E}$ 的 G-P 算法 [7], 伪最近领域法 [8], 真实矢量场法 [9], 求 τ 的改进自相关法 [10], 另一种观点认为两个参数的选取是相关的, 如 C-C 方法 [11]. Takens 定理指出 [12], 相空间中每一点存在映射关系

$$Y_{j+\tau} = \varphi(Y_j), \tag{3}$$

相对于 $Y_j, Y_{j+\tau}$ 中只有 $x_{j+D_{\rm E}\tau}$ 是新信息, 所以 (3) 式可以改写为

$$x_{j+D_{\rm E}\tau} = F([x_j, x_{j+\tau}, x_{j+2\tau}, x_{j+(D_{\rm E}-1)\tau}]). \tag{4}$$

3 基于聚类的选择性 SVM 集成算法

训练集的差异是影响子支持向量机之间差异度的重要条件之一, 在实际海杂波检测中, 在获得较少海杂波数据的情况下, 以很小的代价搭建出泛化性尽可能优秀的预测模型实际应用价值明显, 在训练集有限的情况下, 训练样本之间的差异度可以通过重复取样技术获得, 在子训练样本集中, 原始数据可能出现一次, 也可能一次都不出现, 利用 (bootstrap 技术)[13] 生成 N 组规模为 n 的训练样本集 (n 通常与原始训练集相当), 其中,

$$S = \{x_{i,k}, y_{i,k}\}, \quad x_{i,k} \in R^d,$$
$$y_{i,k} \in R, \quad i = 1, 2, \cdots, N, \quad k = 1, 2, \cdots, n.$$

每组训练样本集训练一个子 SVM, $x_{i,k}$ 与 $y_{i,k}$ 存在映射关系 $F = \{f|f{:}R^d \to R\}$, 支持向量机回归算法的基本思想是通过一个非线性映射把输入样本 X 映射到一个高维特征空间 F 进行线性回归, 即把低维特征空间的非线性回归问题转换为高维特征空间的线性回归问题. 回归函数为 $f(x) = (w \cdot \varphi(x)) + b$, 其中 b 是阈值, φ 是一个非线性映射, 把输入样本集 X 映射到高维特征空间 F, 根据 Vapnik 结构风险最小化原则并考虑函数的复杂度和拟合误差, 引入风险函数 $R(w)$, 子 SVM 的训练算法等价于最小化泛函

$$\min R(w) = \frac{1}{2}\|w_i\|^2 + c_i \sum_{k=1}^{n}(\xi_{i,k} + \xi_{i,k}^*), \tag{5}$$

约束条件为

$$y_{i,k} - w_i\varphi_i(x_{i,k}) - b_i \leqslant \varepsilon_i + \xi_{i,k},$$
$$w_i\varphi_i(x_{i,k}) - y_{i,k} + b_i \leqslant \varepsilon_i + \xi_{i,k}^*,$$
$$\xi_{i,k} \geqslant 0, \quad \xi_{i,k}^* \geqslant 0,$$

其中, $\|w_i\|^2$ 是描述函数 f 复杂度的项, c_i 是惩罚因子, $c_i > 0$, 作用是在经验风险与模型复杂度之间取折中, $\xi_{i,k}$ 与 $\xi_{i,k}^*$ 是引进的松弛变量; ε_i 是引入的不明感损失函数, 具体定义如下;

$$|y_{i,k} - f(x_{i,k})|_{\varepsilon_i} = \begin{cases} |y_{i,k} - f(x_{i,k})| - \varepsilon_i, \\ 0, \end{cases}$$
$$\begin{aligned} &|y_{i,k} - f(x_{i,k})| \geqslant \varepsilon_i, \\ &|y_{i,k} - f(x_{i,k})| \leqslant \varepsilon_i. \end{aligned} \tag{6}$$

对于解上述约束二次优化问题寻找向量 w_i, 其核心是用拉格朗日算子转化为对偶形式求解, 原始的拉格朗日函数为

$$\begin{aligned} L_i =& \frac{1}{2}\|w_i\|^2 + c_i \sum_{k=1}^{n}(\xi_{i,k} + \xi_{i,k}^*) \\ &- \sum_{k=1}^{n}\alpha_{i,k}[\varepsilon_i + \xi_{i,k} + y_{i,k} \\ &- (w_i\varphi_i(x_{i,k}) + b_i)] \\ &- \sum_{k=1}^{n}\alpha_{i,k}^*[\varepsilon_i + \xi_{i,k}^* - y_{i,k} \\ &+ (w_i\varphi_i(x_{i,k}) + b_i)] \\ &- \sum_{k=1}^{n}(n_{i,k}\xi_{i,k} + n_{i,k}^*\xi_{i,k}^*), \end{aligned} \tag{7}$$

其中 $\alpha_{i,k}$, $\alpha^*_{i,k}$, $\eta_{i,k}$, $\eta^*_{i,k}$ 是 Lagrange 乘子, (7) 式的最小化应对 w_i, b_i, $\xi_{i,k}$, $\xi^*_{i,k}$ 求偏倒为 0, 同时根据 Karush-Kuhn-Tucker(KKT) 互补条件, 可得如下关系;

$$\begin{aligned} w(\alpha_i, \alpha_i^*) = &-\frac{1}{2}\sum_{k,g=1}^{n}(\alpha_{i,k}-\alpha_{i,k}^*) \\ &\times(\alpha_{g,k}-\alpha_{g,k}^*)K(x_{i,k}, y_{g,k}) \\ &-\sum_{k=1}^{n}\alpha_{i,k}(\varepsilon_i - y_{i,k}) \\ &-\sum_{k=1}^{n}\alpha_{i,k}^*(\varepsilon_i + y_{i,k}), \end{aligned} \tag{8}$$

约束条件为

$$\begin{aligned} &\sum_{k=1}^{n}(\alpha_{i,k}-\alpha_{i,k}^*)=0, \\ &0 \leqslant \alpha_{i,k} \leqslant c_i, \\ &0 \leqslant \alpha_{i,k}^* \leqslant c_i, \end{aligned}$$

式中, $\alpha_{i,k}$, $\alpha^*_{i,k}$ 为对偶变量, $K(x_i, x_j)$ 为核函数, 求解上述二次规划问题就得到 $\alpha_{i,k}$, $\alpha^*_{i,k}$ 的全局最优解, 从而可以得到

$$w_i = \sum_{k=1}^{n}(\alpha_{i,k}-\alpha_{i,k}^*)\varphi_i(x_{i,k}), \tag{9}$$

因此, f 可以表示为

$$\begin{aligned} y_i(x) = &\sum_{k=1}^{n}(a_{i,k}-a_{i,k}^*)K(x_{i,k}, x_i) \\ &+ b_i, \quad (i=1,2,\cdots,N), \end{aligned} \tag{10}$$

在本模型中采用径向基核函数

$$K(x_i, x_j) = \exp(-\|x_i - x_j\|^2/(2pl^2)), \tag{11}$$

惩罚因子 C_i, 不敏感系数 ε_i, 利用交叉确认算法 (cross-validation)[14] 最优选择.

传统集成方式使用平均或加权平均进行集成, 可能导致过配 (overfitting), 将差异度较小的模型进行集成, 不一定能够降低集成的泛化误差, 还有可能起反作用, 组成集成学习模型的子 SVM 的精度越高, 成员之间的差异度较大, 越有利于集成学习泛化误差的降低 [15] 子 SVM 之间的差异度通过对同一问题的学习结果来衡量, 在相同输入下学习输出值相似度越大证明模型差异度越小, 反之模型差异度越大. 将模型差异度小的子 SVM 归为一簇, 获得数量有限个簇群, 同一簇中子 SVM 的差异度较小, 不同簇的子 SVM 之间的差异度比同一簇中的子 SVM 之间的差异度大, 选择性集成即选每一簇中精度最高的子 SVM 作为成员进行集成. 对 N 组子 SVM 进行聚类时保证聚类精度的同时提高聚类的速度在实际应用中意义明显, K 均值聚类算法 [16] 拥有收敛速度快算法简单等优点, 对子 SVM 进行聚类相当于对各子 SVM 的输出结果进行聚类, 比较在验证集上按不同聚类数 C 得到的集成学习模型的预测精度, 由 1 起始逐个增加类别数 C, 选择集成精度最高时的 C^* 作为最优类别数, 最终集成输出

$$y_{\text{ensemble}}(x) = \frac{1}{C^*}\sum_{i=1}^{C^*} y_i(x). \tag{12}$$

4 仿真实验

实验 1 选用 Lorenz 系统产生的混沌时间序列进行仿真, Lorenz 迭代方程为

$$\begin{aligned} \dot{x} &= \sigma(y-x), \\ \dot{y} &= \rho x - y - xz, \\ \dot{z} &= -\beta z + xy, \end{aligned} \tag{13}$$

式中 x, y, z 为时间函数, $\dot{x}$, $\dot{y}$, $\dot{z}$ 分别为它们对时间的微分, 其中 $\sigma = 10$, $\rho = 28$, $\beta = 8/3$, 初值 $x = 8$, $y = 5$, $z = 10$, 利用四阶龙格库塔 (Runger-Kutta) 法求得步长为 0.01, 其中 x 分量作为实验用时间序列, 取长度为 2000 的实验数据用于训练混沌预测模型, 将后续 1000 个数据中前 100 个数据作为验证集, 后 900 个数据验证模型的准确性. 根据经典 G-P 算法求得嵌入维为 5, 重构时延为 1, 信噪比定义为信号的幅值 A 与噪声的标准差之比, 即 $\text{SNR} = 20\lg\dfrac{A}{\sigma_N}$.

在验证集的第 405—454 点处加入一个幅值为 0.0001 的瞬态微弱信号 $s(n)$, 将 $s(n)$ 叠加到混沌背景中构成观测时间序列 $x(n)$, 信噪比 SNR 达 −92.8139 dB, 先对数据进行归一化处理, 然后进行相空间重构和选择性 SVM 集成单步预测, 结果进行反归一化处理, 预测结果均方根误差 (RMSE) 为 0.000015425, 图 1(c) 为检测结果 (最开始用来预测的 4 个点没有画出), 从图中可以看出 405—458 点的误差明显偏大, 这是由于嵌入维为 5, 时间延迟为 1, 454 点作为夹杂微弱目标信号时间序列的末点, 最后一次用这个点进行单步预测时, 必将引起

预测值 $\hat{x}(458)$ 存在较大的误差, 由此可以判断瞬态微弱信号的存在. 比较文献 [17] 中的 -77.3301 dB 门限和 0.0080 均方根误差, 本方法对于瞬态信号的检测能力相对于传统支持向量回归方法有了很大的加强, 包括误差的减小和检测门限的下降.

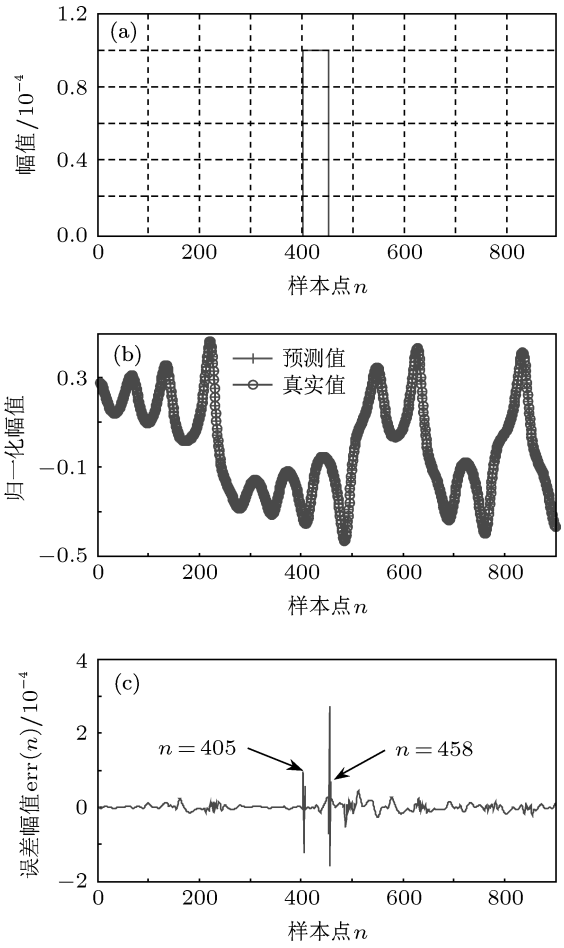

图 1 瞬态信号的检测结果 (a) 瞬态信号; (b) 含有瞬态信号的混沌信号预测值和真实值; (c) 单步预测误差

实验 2 设目标信号为

$$s(n) = 0.00025\sin(2\pi f n),$$

归一化频率为 0.04, 混沌背景噪声信号为 $c(n)$, 观测信号为 $x(n) = c(n) + s(n)$, 信噪比达 -90.2679 dB. 按照实验 1 步骤进行检测, 得到单步预测误差 $\mathrm{err}(n)$, $\mathrm{err}(n)$ 中主要包含两部分内容, 一是选择性 SVM 集成预测模型本身的预测误差, 另一个是周期信号 $s(n)$ 因此, 现在把检测湮没在强混沌背景噪声中的周期信号问题转化为检测湮没在单步预测误差 $\mathrm{err}(n)$ 中的周期信号问题把单步误差进行 FFT 变换, 可以明显看出预测误差的频谱在 4 处出现峰值, 如图 2 采用选择性 SVM 集成预测模型可以明显检测出预测误差中所含的微弱谐波信号的频率特性, 实验结果进行反归一化处理, RMSE 为 0.00000679, 比较文献 [18] 中支持向量回归的 -8965 dB 门限和 0.022 的均方根误差以及神经网络的 -62.82 dB 门限和 0.1824 的均方根误差均有大幅提高

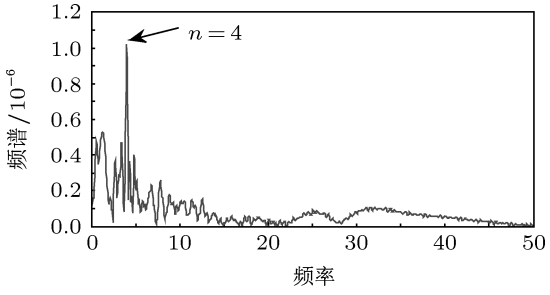

图 2 周期信号单步预测误差频谱

实验 3 本文采用的海杂波数据是加拿大 McMaster 大学的 IPIX 雷达海杂波数据, 该雷达发射频率为 9.39 GHz, 天线高度 30 m, 极化方式分为 HH, VV, HV, VH, 每个距离单元的回波数据包含 131072 个采样点, 天线增益为 45.7 dB.

实验选用雷达数据集的第 269# 海杂波数据, 采集距离单元 3 (不含小目标) VV 极化方式的 1700 个点, 对数据进行归一化处理, 为保证数据的实时性, 前 1000 个采样点作为训练样本, 后 700 个采样点中前 100 个作为验证集, 后 600 个作为预测对比, 对应时间约为 0.6 s, 按照前面所述方法使用选择性 SVM 集成, 图 3(b) 为其预测误差幅值, 可以看出预测误差平滑, 无明显尖峰, 经计算得其均方误差为 0.0011, 参考文献 [19], 对于径向基函数 (RBF) 为 0.0153 的均方误差和最小二乘支持向量机 (LSSVM) 为 0.0137 的均方误差, 本实验在回归精度上有很大提高.

表 1 海杂波序列预测模型性能对比

	选择性 SVM 集成	LS-SVM	RBF 神经网络	Volterra 预测器
均方误差	0.0011	0.0137	0.0153	0.0223

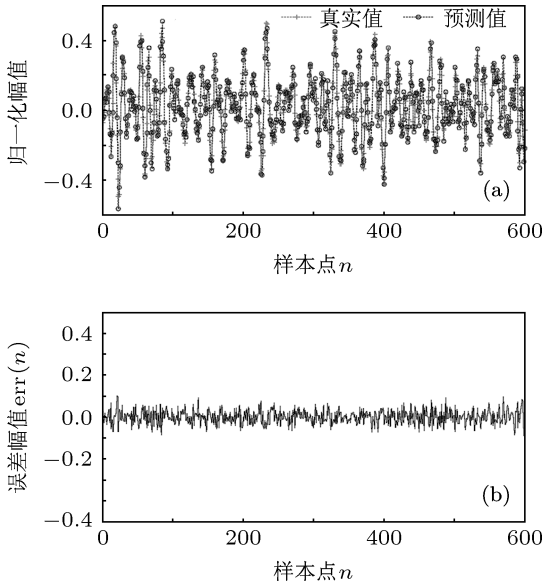

图 3 269# 数据海杂波单元的预测结果 (a) 预测值和真实值; (b) 单步预测误差

选用雷达数据集的第 54# 海杂波数据, 选取第 1 海杂波距离单元和第 8 目标距离单元的数据, 两组距离单元的数据均采样 VV 极化方式的 2300 个点, 1500 个采样点作为训练样本, 后续 800 个采样点中前 100 个作为验证集, 后 700 个作为预测对比, 使用选择性 SVM 集成进行单步预测, 经计算目标距离单元 RMSE 为 0.0177, 海杂波距离单元 RMSE 为 0.0264, 比较图 4(a) 与 (b), 在有目标信号的距离单元中, 预测误差中存在明显尖峰, 证明海杂波基于选择性 SVM 集成的一步预测误差对目标的存在具有较强的敏感性, 比较目标距离单元与海杂波距离单元预测误差的差别, 可以初步判定是否含有目标

实验结果表明, 选择性集成的预测方法结合了选择性集成学习与支持向量机的特性, 能够有效的预测含有微弱信号的混沌时间序列, 预测精度高于传统 RBF 神经网络和 SVM 回归方法, 先对海杂波数据进行可重复抽样处理, 扩充训练样本集, 再分别训练各子支持向量机, 选择回归精度高差异度大的子 SVM 进行集成, 克服了传统集成学习方法计算量大的缺点, 较传统混沌背景下微弱信号检测方法, 有很低的检测门限和预测误差

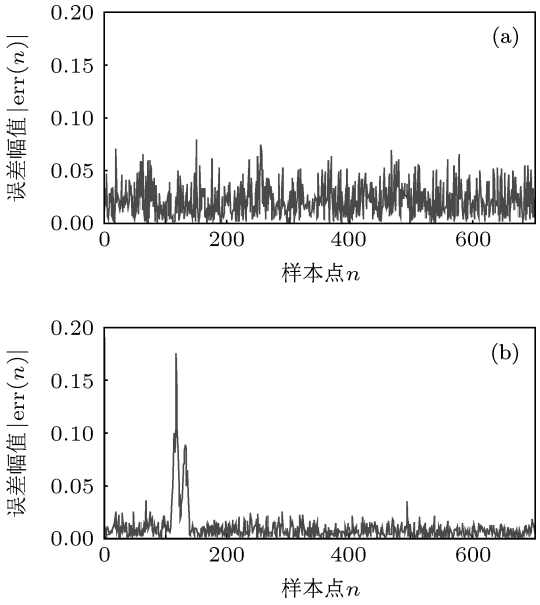

图 4 54# 数据预测结果 (a) 海杂波距离单元预测误差绝对值; (b) 目标距离单元预测误差绝对值

5 结 论

结合混沌时间序列相空间重构以及局部可预测的特点, 本文提出基于 SVM 集成的预测模型, 并对子 SVM 进行聚类, 选择每一簇中精度最高的子 SVM 进行集成, 保证子 SVM 有较高的精度和差异度, 将混沌背景下微弱信号检测问题转化成从预测误差中提取有用信号; 实验表明这种检测方法有很低的检测门限和预测误差, 与传统的 RBF 神经网络和 SVM 回归方法相比, 以很小的运算代价显著提高了学习系统的精度、稳健性、可行性、泛化能力. 我们将继续研究该算法与在线学习相结合, 以达到更高的预测速度和更准确的多步预测结果.

[1] Haykin S, Li X B 1995 *Proceedings IEEE* **83** 95

[2] Zhang J S, Xiao X C 2000 *Acta Phys. Sin.***49** 403 (in Chinese) [张家树, 肖先赐 2000 物理学报 **49** 403]

[3] Xing H Y, Xu W 2007 *Acta Phys. Sin.* **56** 3771 (in Chinese) [行鸿彦, 徐伟 2007 物理学报 **56** 3771]

[4] Zhang J F, Hu S S 2007 *Acta Phys. Sin.* **56** 713 (in Chinese) [张军峰, 胡寿松 2007 物理学报 **56** 713]

[5] Cui W Z, Zhu C C, Bao W X, Liu J H 2004 *Acta Phys. Sin.* **53** 3303 (in Chinese) [崔万照, 朱长纯, 保文星, 刘君华 2004 物理学报 **53** 3303]

[6] Zhou Z H, Wu J X 2002 *Artif. Intell* **137** 239

[7] Grassberger P, Procaccia I 1983 *Phys. Rev. Lett.* **50** 346

[8] Cao L Y 1997 *Physica* D**110** 43
[9] Kaplan D T, Glass L 1992 *Phys. Rev. Lett.* **68** 427
[10] Aguirre L A 1995 *Phys. Lett.* A **203** 88
[11] Kim H S, Eykholt R, Salas J D 1999 *Physica* D **127** 48
[12] Takens F 1981 *Lecture Notes in Mathematics* **898** 366
[13] Leo B 1996 *Mach. Learn.* **21** 123
[14] Zhang L, Zhou W, Jiao L 2004 *IEEE Trans. Syst. Man Cyb.* B **34** 34
[15] Wu J X, Zhou Z H, Shen X H, Chen Z Q 2000 *J. Comput. Res. Dev.* **37** 2000 (in Chinese) [吴建鑫, 周志华, 沈学华, 陈兆乾 2000 计算机研究与发展 **37** 2000]
[16] Kanungo T, Mount D M, Netanyahu N S, Piatko C D, Silverman R, Wu A Y 2002 *IEEE Trans. Pattern Anal. Mach. Intell.* **24** 881
[17] Xing H Y, Jin T L 2010 *Acta Phys. Sin.* **59** 140 (in Chinese) [行鸿彦, 金天力 2010 物理学报 **59** 140]
[18] Du J Y, Hou Y B 2007 *J. Sci. Instru.* **28** 555 (in Chinese) [杜京义, 侯媛彬 2007 仪器仪表学报 **28** 555]
[19] Wang F Y, Yuan G N, Xie Y J, Qiao X W 2009 *Radar. Sci. Technol.* **7** 53 (in Chinese) [王福友, 袁赣南, 谢燕军, 乔相伟 2009 雷达科学与技术 **7** 53]

Weak signal estimation in chaotic clutter using selective support vector machine ensemble*

Xing HongYan† Qi ZhengDong Xu Wei

(*Jiangsu Key Laboratory of Meteorological Observation and Information Processing, Nanjing University of Information Science and Technology, Nanjing* 210044*, China*)

(*College of Electronic and Information Engineering, Nanjing University of Information Science and Technology, Nanjing* 210044 *China*)

(Received 23 May 2012; revised manuscript received 23 June 2012)

Abstract

A method of detecting weak signals embedded in chaotic noise by selective support vector machine ensemble based on the theory of phase space reconstruction of the complicated nonlinear system is presented. For improving the generalization ability of support vector machine ensemble, K-means algorithm is used to select the most accurate individual support vector machine from every cluster for ensembling It is established a one-step predictive model that detects the weak signal, including transient signal and period is signals, from the predictive error in the chaotic sequences. It is illustrated in the experiment which is conducted to detect weak signals from Lorenz chaotic background and IPIX Sea Clutter, that the proposed method is highly effective to detect weak signal from a chaotic background and to minimize the influence of noise on weak signals, Compared wich the RBF neural network and SVM model, the new method presents great value in predicting accuracy and detection threshold.

Keywords: support vector machine, ensemble, clutter, weak signal estimation

PACS: 05.45.Pq

* Project supported by the National Natural Science Foundation of China (Grant No. 61072133), and the Jiangsu Sensor Network and Modern Meteorological Equipment Preponderant Discipline Platform.

† E-mail: xinghy@nuist.edu.cn