第 56 卷 第 12 期 2007 年 12 月
1000-3290/2007/56(12)/6820-08

物 理 学 报
ACTA PHYSICA SINICA

Vol.56,No.12,December,2007


# 基于选择性支持向量机集成的混沌时间序列预测*

蔡俊伟 胡寿松 陶洪峰

(南京航空航天大学自动化学院,南京 210016)

(2006 年 12 月 28 日收到;2007 年 6 月 8 日收到修改稿)

提出了一种基于聚类的选择性支持向量机集成预测模型.为提高支持向量机集成的泛化能力,采用自组织映射和 $K$ 均值聚类算法结合的聚类组合算法,从每簇中选择出精度最高的子支持向量机进行集成,可以保证子支持向量机有较高精度并提高了子支持向量机之间的差异度.该方法能以较小的代价显著提高支持向量机集成的泛化能力.采用该方法对 Mackey-Glass 混沌时间序列和 Lorenz 系统生成的混沌时间序列进行预测实验,结果表明可以对混沌时间序列进行准确预测,验证了该方法的有效性.

关键词:支持向量机,集成,混沌时间序列,聚类

PACC:0545

## 1. 引 言

混沌时间序列是非线性确定性系统产生的具有内在随机性的确定性过程,宏观上表现为无序无律的混乱运动以及对初值十分敏感的蝴蝶效应,微观上呈现无穷嵌套几何自相似性.不同于随机时间序列,混沌时间序列具有短期的可预测性和长期的不可预测性.近年来,随着混沌理论研究的不断深入及其在信号处理、自动控制和通信领域中的广泛应用,混沌时间序列的建模和预测已成为混沌信号处理领域的一个非常重要的研究方向[1—7].目前,已经有多种方法被应用于混沌时间序列预测中,如全局预测法[8]、局部预测法[9]、自适应非线性滤波预测方法[5]和基于神经网络的预测法[10,11].其中人工神经网络因其具有较强的非线性映射能力,在混沌时间序列的预测中得到了较多的应用.但因神经网络训练过程遵循经验风险最小化原则,存在过拟合、训练过程中受局部极小点的困扰、网络结构的选择过分依赖于经验等固有的缺陷,直接影响了混沌时间序列预测的精度和可靠性,从而大大限制了其进一步的应用.基于 Vapnik[12] 等提出的统计学习理论的支持向量机(SVM)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,从而提高模型的泛化能力.这一优点在小样本学习中更为突出,并且不依赖于系统的数学模型,同时具有自学习自调整模型的特点,能对各种混沌时间序列产生较好的预测性能.但是,SVM 的训练问题实质上是一个凸二次规划问题,大多数二次规划算法由于需要利用整个 Hessian 矩阵,受计算机内存容量的限制,无法处理大数据量问题,因此常用一些近似算法来降低时空消耗,这使得系统泛化性能受到了影响[13].为解决上述问题,SVM 集成方法[13,14] 利用多个子 SVM 对输出结论进行合成,可以达到更好的泛化性能.集成学习是利用多个学习机来解决同一问题,目的是更有效地提高学习系统的泛化性能.集成学习的有效性取决于子学习机的精度和子学习机之间的差异度.子学习机精度越高,子学习机之间的差异度越大,就越有利于集成泛化能力的提高[13—15].传统的 SVM 集成方法[13] 把所有的子 SVM 进行集成.本文提出的基于聚类技术的选择性 SVM 集成方法,仅选择部分精度较高差异度较大的子 SVM 进行集成,提高了组成 SVM 集成的子 SVM 的精度和子 SVM 之间的差异度,能获得优于或者接近单个最佳

* 国家自然科学基金重点项目(批准号:60234010)和航空科学基金(批准号:05E52031)资助的课题.

子 SVM 的泛化性能，且比传统 SVM 集成方法具有更好的泛化性能.

本文基于混沌时间序列固有的确定性和非线性，根据混沌动力系统的相空间延迟坐标重构理论，设计了一种基于聚类的选择性 SVM 集成的混沌时间序列的预测模型，以具有时滞特性的 Mackey-Glass 时间序列和 Lorenz 系统产生的时间序列为例验证了该模型的建模能力. 结果表明，基于聚类的选择性 SVM 集成的混沌时间序列预测模型是精确的，可得到优于或接近单个最佳子 SVM 的预测性能，比传统 SVM 集成方法具有更好的预测性能.

## 2. SVM 集成预测模型

Kearns 等[16]指出，在可能近似正确（PAC）学习模型中，若存在一个多项式级学习算法来辨别一组概念，并且辨别正确率很高，那么这组概念是强可学习的；如果学习算法辨别一组概念的正确率仅比随机猜测略好，那么这组概念是弱可学习的. 文献[16]提出了弱学习算法与强学习算法的等价性问题，即是否可将弱学习算法提升成强学习算法. 如果两者等价，那么在学习概念时，只需找到一个比随机猜测略好的弱学习算法，就可以将其提升为强学习算法，而不必直接去找通常情况下很难获得的强学习算法. 文献[16]证明了只要有足够的数据，弱学习算法就能通过集成的方式生成任意高精度的估计. 上述等价性问题可以看作是 SVM 集成思想的出发点. SVM 集成是用有限个子 SVM 对同一问题进行学习. 对于某个输入，它在集成上的输出是由集成中的各个子 SVM 在该输入上的输出共同决定的.

若给定 $N$ 组规模为 $n$ 的训练样本集

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

用各组训练样本集生成的 $N$ 个子 SVM 拟合函数形式为

$$y_i(x) = w_i\phi_i(x) + b_i \quad (i = 1,2,\cdots,N). \tag{1}$$

对 $N$ 个训练好的子 SVM 进行集成，得到输出为

$$y(x) = \sum_{i=1}^{N} \omega_i y_i = \sum_{i=1}^{N} \omega_i (w_i\phi_i(x) + b_i), \tag{2}$$

式中 $w_i$ 为子 SVM 的权向量，$b_i$ 为子 SVM 的偏差，$\phi_i(\cdot)$为子 SVM 的非线性映射函数，$\omega_i$ 为子 SVM 的权重，满足下列条件：

$$0 \leqslant \omega_i \leqslant 1, \tag{3}$$

$$\sum_{i=1}^{N} \omega_i = 1. \tag{4}$$

子 SVM 利用 $\phi_i(\cdot)$把输入空间映射到一个高维数（可能无限维）特征空间（Hilbert 空间），并在这个新空间中求取最优线性分类面实现数据的线性可分. 并且，$\phi_i(\cdot)$可通过定义在内积空间的满足 Mercer 条件的核函数 $k_i(x,x')$来代替，而核函数可以用原空间的函数来实现且无须知道 $\phi_i(\cdot)$的具体形式，计算复杂度没有增加. SVM 集成利用各个子 SVM 实现了对混沌时间序列在不同高维空间的扩展，分别提取其蕴藏的系统信息并加以综合，为混沌时间序列的状态重构提供了一个简单有效的途径.

根据统计学习理论[12]，基于 SVM 集成的混沌时间序列的拟合函数为

$$y(x) = \sum_{i=1}^{N} \omega_i \left( \sum_{k=1}^{n} \alpha_{i,k} k_i(x_{i,k}, x) + b_i \right), \tag{5}$$

式中 $\alpha_{i,k}$为子 SVM 的支持向量，$b_i$ 为子 SVM 的偏差，$k_i(x_{i,k}, x_{i,l})$为子 SVM 的核函数. 本文采用高斯核函数和多项式核函数作为子 SVM 的核函数.

高斯核函数的形式为

$$k(x_k, x_l) = \exp(- \| x_k - x_l \|^2 / 2\sigma^2), \tag{6}$$

式中 $\sigma$ 为核函数的参数，是预先选择的常数.

多项式核函数的形式为

$$k(x_k, x_l) = ((x_k \cdot x_l) + c)^d, \tag{7}$$

式中 $d \in N$，$c \geqslant 0$ 为核函数的参数，为预先选择的常数.

混沌时间序列预测的基础是状态空间的重构理论. 目前对序列动力学因素的分析广泛采用的是延迟坐标状态空间重构法. 系统的相空间维数通常很高甚至无穷，但在大多数情况下维数是未知的. 实际情况下，延迟坐标状态空间重构法是将给定的时间序列 $x_1, x_2, \cdots, x_n, \cdots$扩展到三维甚至更高维的空间，以便把时间序列中蕴藏的信息充分显露出来，并加以分类和提取.

由此看来，状态空间的重构理论和 SVM 的基本思想有相同之处：都是把输入空间的向量扩展到高维空间，提取系统蕴藏的信息和规律. 而 SVM 集成利用各个子 SVM 提取的信息加以综合，能更好地提取系统蕴藏的信息和规律.

利用 SVM 集成重构相空间的状态分量时，子 SVM 输入变量的个数应至少大于时间序列的嵌入维数和延迟时间的乘积.这样基于嵌入维 $m$，$n$ 个混沌时间序列的子 SVM 预测模型的输入向量可表示为

$$X_i(t-1)=[x_i(t-\tau),x_i(t-2\tau),\cdots,x_i(t-m\tau)]\quad(t=1,2,\cdots,n;i=1,2,\cdots,N),\tag{8}$$

式中 $x_i(\cdot)$是混沌时间序列，$\tau$ 为时延宽度，通常 $\tau$ 取不小于 1 的整数.这样，基于 SVM 集成的混沌时间序列的预测值为

$$\hat{y}=\sum_{i=1}^{N}\omega_i f(X_i(t-1)).\tag{9}$$

利用集成 SVM 根据构建的混沌时间序列输入向量和输出向量进行学习，获取混沌时间序列 SVM 集成预测模型参数 $\omega_i$，$\alpha_{i,k}$，$b_i$.预测模型的参数所蕴藏的关系就是混沌时间序列各个向量的过去和将来的关系.至此，基于 SVM 集成的预测模型已经建立.这样，就可以利用该模型预测将来的混沌时间序列的输出.

## 3. 基于聚类的选择性 SVM 集成算法

首先采用可重复取样技术(bootstrap 技术)[17]生成 $N$ 组规模为 $n$ 的训练样本集

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

用每组训练样本集训子 SVM.采用 $\varepsilon$ 不敏感损失函数，子 SVM 的训练算法可归结为下面的约束优化问题：最小化泛函

$$J(w_i,b_i,\xi_i)=\frac{1}{2}\|w_i\|^2+C_i\sum_{k=1}^{n}(\xi_{i,k}+\xi_{i,k}^{*}).\tag{10}$$

约束条件为

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

这里 $C_i$ 为调节常数，$C_i>0$.通过上述优化问题的对偶形式可以求得它的最优解，其对偶形式为约束优化问题：最大化泛函

$$W(\alpha_i,\alpha_i^{*})=-\frac{1}{2}\sum_{k=1}^{n}\sum_{g=1}^{n}(\alpha_{i,k}-\alpha_{i,k}^{*})\times(\alpha_{i,g}-\alpha_{i,g}^{*})K(x_{i,k},x_{i,g}).\tag{11}$$

约束条件为

$$-\varepsilon_i\sum_{k=1}^{n}(\alpha_{i,k}+\alpha_{i,k}^{*})+\sum_{k=1}^{n}y_{i,k}(\alpha_{i,k}-\alpha_{i,k}^{*})\times\sum_{k=1}^{n}(\alpha_{i,k}-\alpha_{i,k}^{*})=0,$$
$$0\leqslant\alpha_{i,k}\leqslant C_i,$$
$$0\leqslant\alpha_{i,k}^{*}\leqslant C_i.$$

通过求解优化问题(11)式，根据其最优解，可得

$$w_i=\sum_{x_{i,k}\in I_i}^{n}(\alpha_{i,k}-\alpha_{i,k}^{*})\phi_i(x_{i,k}),\tag{12}$$

$$b_i=\frac{1}{N_i^{\mathrm{SV}}}\Big\{\sum_{0<\alpha_{i,k}<C_i}\Big[y_{i,k}-\sum_{x_{i,g}\in I_i}(\alpha_{i,g}-\alpha_{i,g}^{*})\times K(x_{i,g},x_{i,k})-\varepsilon_i\Big]+\sum_{0<\alpha_{i,k}^{*}<C_i}\Big[y_{i,k}-\sum_{x_{i,g}\in I_i}(\alpha_{i,g}-\alpha_{i,g}^{*})\times K(x_{i,g},x_{i,k})+\varepsilon_i\Big]\Big\},\tag{13}$$

式中 $N_i^{\mathrm{SV}}$为第 $i$ 个子 SVM 的支持向量个数，$I_i$ 为第 $i$ 个子 SVM 的支持向量所组成的集合.最终可得子 SVM 回归函数的拟合函数形式

$$y_i(x)=w_i\phi_i(x)+b_i=\sum_{x_{i,k}\in I_i}(\alpha_{i,k}-\alpha_{i,k}^{*})K(x_{i,k},x)+b_i\quad(i=1,2,\cdots,N).\tag{14}$$

为了提高 SVM 集成的泛化性能，需要保证子 SVM 的精度和子 SVM 之间的差异度.可以根据第 $i$ 个子 SVM 的均方根误差(MSE)

$$e_i=\sqrt{\frac{1}{n}\sum_{h=1}^{n}(y(h)-\hat{y}_i(h))^2}\tag{15}$$

作为评价子 SVM 精度的指标.这里 $y(h)$和 $\hat{y}_i(h)$分别为第 $h$ 个时间序列的实际值和第 $h$ 个时间序列的第 $i$ 个子 SVM 的预测值，$e_i$ 为第 $i$ 个子 SVM 的 MSE.采用 bootstrap 技术训练子 SVM 可以提高子 SVM 之间的差异度.子 SVM 之间差异度可以通过它们在相同输入下的输出之间的相似性来衡量.在相同输入下，子 SVM 输出之间相似度越大差异度越小，反之则差异度越大.定义子 SVM 的聚类为对相同输入下子 SVM 输出的聚类.相同输入下输出在同一簇的子 SVM 归为一簇.同一簇中子 SVM 的差异

度较小，不同簇的子 SVM 之间差异度比同一簇中的子 SVM 间的差异度大. 由于同簇中的子 SVM 都是相似的 SVM，因此可以在每一簇中选择一个子 SVM 代表该簇中所有子 SVM. 基于上述思想，从每一簇中选择一个子 SVM，可以进一步提高集成的子 SVM 之间的差异度. 而选择的子 SVM 为该簇中精度最高的，这样可以保证集成的子 SVM 有较高精度. 最终，把所有选择出来的子 SVM 进行集成.

自组织映射(SOM)[18]神经网络是一种基于非监督和竞争学习的人工神经网络模型，可实现将数据从高维空间映射到低维空间，其中每一个神经元可代表一个模式，因此 SOM 可用于复杂的高维数据聚类分析. 但是应用 SOM 算法进行聚类时，网络收敛时间过长. $K$ 均值聚类算法[19]是实现动态聚类的一个有效方法，但是聚类结果受初始聚类中心的选择影响较大，如果初始聚类中心选取不当，将得不到好的聚类结果. 本文采用 SOM 和 $K$ 均值聚类算法结合的 SOMK 聚类算法对输出空间上的样本进行聚类，算法过程如下：(1)先执行 SOM 算法，把待聚类的数据对象输入 SOM 网络进行训练，经过网络训练出一组权值. 此阶段的训练次数可以减少，不必让 SOM 完全收敛. (2)以 SOM 的聚类结果得到的权值为初始聚类中心，对 $K$ 均值聚类算法进行初始化，执行 $K$ 均值算法进行聚类. 该聚类算法既能保持 SOM 网络自组织的特点，又能吸收 $K$ 均值聚类算法高效率的特点，同时弥补了 SOM 网络收敛时间长和 $K$ 均值聚类算法初始聚类中心选取不当造成聚类效果不佳的特点.

若在给定的样本数为 $n^*$ 的集合 $U$ 上对 $N$ 个子 SVM 进行聚类，相当于对 $N$ 个 $n^*$ 维向量进行聚类. 在 SOMK 聚类算法中，类别数 $C$ 要事先给定. 为了保证较适合的类别数 $C$，通过比较在验证集上按不同类别 $C$ 聚类所得到的 SVM 集成的精度，选取使 SVM 集成精度最高的 $C^*$ 作为最合适的类别数.

根据上述分析，下面给出基于聚类的选择性 SVM 集成混沌时间序列预测的具体学习算法步骤.

1)选择一个合适的嵌入维数 $m$，选择 $N$ 个子 SVM 的核函数(本文选径向基核函数和多项式核函数)和相应的核参数 $\sigma_i, d_i, c_i$，调节常数 $C_i$ 和 $\varepsilon_i$ ($i = 1,2,\cdots,N$)，初始化聚类类别数 $C = 1$，初始化 SOMK 网络.

2)根据选择的嵌入维数生成 SVM 集成预测所需的输入向量 $X$($M \times m$ 维)和输出向量 $Y$($M \times 1$ 维).

3)对输入样本 $X$ 和输出样本 $Y$ 采用 bootstrap 技术生成 $N$ 组规模为 $n$($n < M$)的训练样本集

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

把训练集等分成 $q$ 个验证集 $V_j$ ($j = 1,2,\cdots,q$)，其中 $n$ 能被 $q$ 整除.

4)给定样本集 $S$，求解优化问题(11)式，得到 $N$ 个子 SVM.

5)在训练集 $S$ 上，根据聚类数 $C$ 对子 SVM 进行聚类，所有子 SVM 划分成 $C$ 簇. 选择每一簇中在 $q$ 个验证集 $V_j$ 上平均精度最高的子 SVM，选择出来的 $C$ 个子 SVM 的序号记为 $j^* = 1,2,\cdots,C$，记选择出来的子 SVM 的输出为 $y_{j^*}(\cdot)$. 对选择的子 SVM 采用简单平均集成，即

$$y_{\text{ensemble}}(x) = \frac{1}{C}\sum_{j^*=1}^{C} y_{j^*}(x). \tag{16}$$

6)计算 SVM 集成在 $q$ 个验证集 $V_j$ 上的平均精度. 令 $C = C + 1$，执行上一步算法，直到 $C = N$. 选取使得 SVM 集成在 $q$ 个验证集 $V_j$ 上平均精度最高的类别数 $C^*$ 作为最适合类别数.

7)令聚类数 $C = C^*$，得最终集成输出

$$y_{\text{ensemble}}(x) = \frac{1}{C^*}\sum_{j^*=1}^{C^*} y_{j^*}(x). \tag{17}$$

## 4. 混沌时间序列的 SVM 集成预测实例

本文将 Mackey-Glass 混沌时间序列和 Lorenz 系统产生的时间序列预测作为仿真实例，以证明所提出方法的有效性. 为衡量预测模型的精确度，采用误差

$$e(n) = x(n) - \hat{x}(n) \tag{18}$$

作为评价模型的每一个时间序列的预测效果，并采用 MSE

$$e = \sqrt{\frac{1}{K}\sum_{n=1}^{K}(x(n) - \hat{x}(n))^2} \tag{19}$$

作为评价模型整体预测效果的指标. 这里 $x(n)$ 和 $\hat{x}(n)$ 分别为第 $n$ 个混沌时间序列的实际值和 SVM 集成预测模型的预测值，$e$ 为 SVM 集成预测模型的 MSE.

### 4.1. Macky-Glass 的混沌时间序列预测

Mackey-Glass 混沌时间序列产生方程为

$$\frac{\mathrm{d}x}{\mathrm{d}t} = \frac{0.2x(t-\tau)}{1+x^{10}(t-\tau)} - 0.1x(t), \qquad (20)$$

式中 $\tau$ 为时滞参数，当 $\tau \geqslant 17$ 时产生的数据是混沌的，$\tau$ 值越大混沌程度越高. 图 1 为 $\tau = 30$ 的混沌时间序列. 从图 1 不难看出，时间序列具有复杂的非线性混沌特征. 取嵌入维数 $m = 3$，抽取 2400 对输入输出数据. 1200 对数据作为训练集，训练集等分为 10 个验证集，另 1200 对数据用于验证模型的准确性. 首先通过 bootstrap 技术独立地训练出 16 个子 SVM，采用高斯核函数和多项式核函数，选择不同的核参数和调节常数 $C_i$，$\varepsilon_i$ 均取为 0.01. 按照本文所提出的方法，根据 SOMK 聚类方法选择 3 个子 SVM 进行集成. 为了进行比较，我们也列出了用全部 16 个子 SVM 进行集成的传统集成方法得到的结果. 图 2 给出了系统的实际输出和所建模型的预测值比较曲线，图 3 给出了模型各点的预测误差曲线. 整个测试集的 MSE 为 0.0016. 从图 2 和图 3 不难看出，本文所提出的 SVM 集成预测模型的预测值与系统的实际输出值符合很好，系统的预测精度比较高.

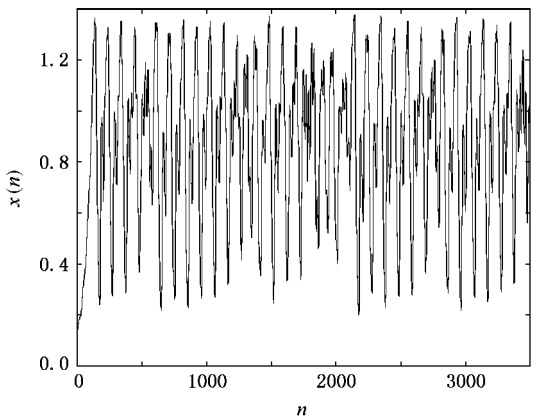


图 1 $\tau = 30$ 的 Mackey-Glass 混沌时间序列

图 4 给出了 4 次实验的结果，图中横轴为实验序号，纵轴为 MSE. 每次实验的数据中，第 1 列为最差子 SVM 在测试集上的 MSE，第 2 列为最佳子 SVM 在测试集上的 MSE，第 3 列为使用本文提出的方法的 SVM 集成在测试集上的 MSE，第 4 列为使用传统方法的 SVM 集成在测试集上的 MSE.

从图 4 可见，采用本文所提出的基于聚类的选择性集成方法都能取得比传统集成方法更好的效果. 同时还可以看出，采用本文方法能取得比单个最佳子 SVM 更好或者接近的效果. 而在实际应用 SVM 和 SVM 集成时，我们不能预先知道哪个子 SVM 泛化误差最小(或在测试集上误差最小)，因此，本文提

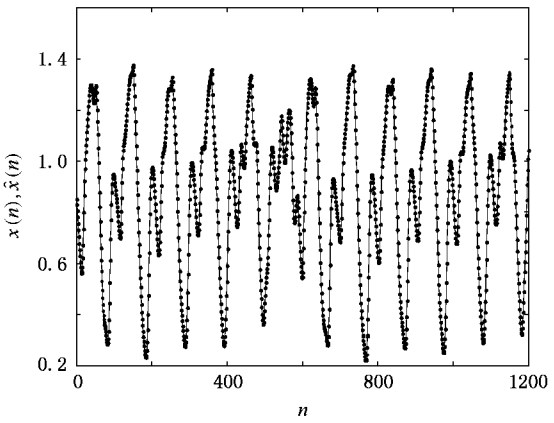


图 2 Mackey-Glass 混沌时间序列的实际输出(实线)和预测值(点线)

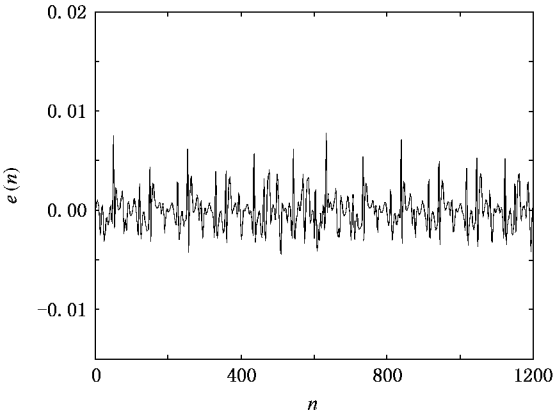


图 3 Mackey-Glass 混沌时间序列的预测误差曲线

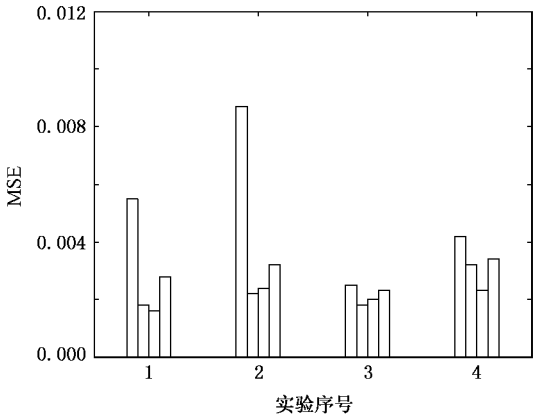


图 4 Mackey-Glass 混沌时间序列的 MSE 实验结果比较

出的方法具有实际应用价值.

### 4.2. Lorenz 混沌时间序列预测

Lorenz 混沌时间序列产生方程为

$$\begin{aligned}\dot{x} &= \sigma(y - x),\\ \dot{y} &= \rho x - y - xz,\\ \dot{z} &= -\beta z + xy,\end{aligned} \tag{21}$$

式中 $\sigma = 10$，$\rho = 28$，$\beta = 8/3$. 假定初始条件 $x_0 = 8$，$y_0 = 5$，$z_0 = 10$，利用定步长（$\Delta t = 0.02$）四阶龙格-库塔法获取变量 $x$ 的序列如图 5 所示. 从图 5 不难看出，时间序列具有复杂的非线性混沌特征. 取嵌入维数 $m = 3$，抽取 2000 对输入输出数据. 1000 对数据作为训练集，训练集等分为 10 个验证集，另 1000 对数据用于验证模型的准确性. 首先通过 bootstrap 技术独立地训练出 12 个子 SVM，采用高斯核函数和多项式核函数，选择不同的核参数和调节常数 $C_i$，$\varepsilon_i$ 均取为 0.1. 按照本文所提出的方法，根据 SOMK 聚类方法选择 3 个子 SVM 进行集成. 为了进行比较，我们也列出了用全部 12 个子 SVM 进行集成的传统集成方法得到的结果. 图 6 给出了系统的实际输出和所建模型的预测值比较曲线，图 7 给出了模型各点的预测误差曲线. 整个测试集的 MSE 为 0.0106. 从图 6 图和图 7 不难看出，本文所提出的 SVM 集成预测模型的预测值与系统的实际输出值符合得很好，系统的预测精度比较高.

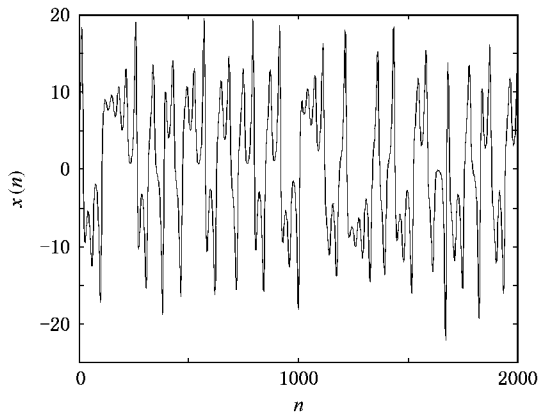


图 5　Lorenz 混沌时间序列

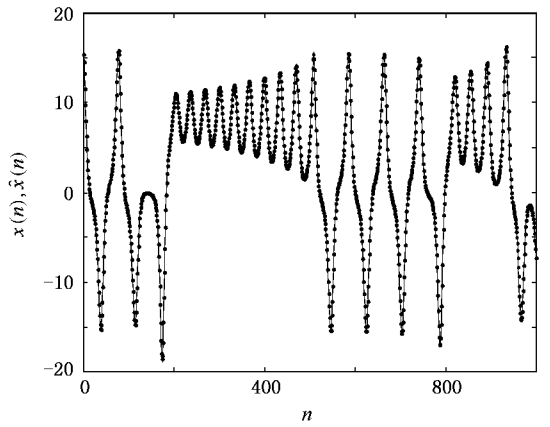


图 6　Lorenz 混沌时间序列的实际输出（实线）和预测值（点线）

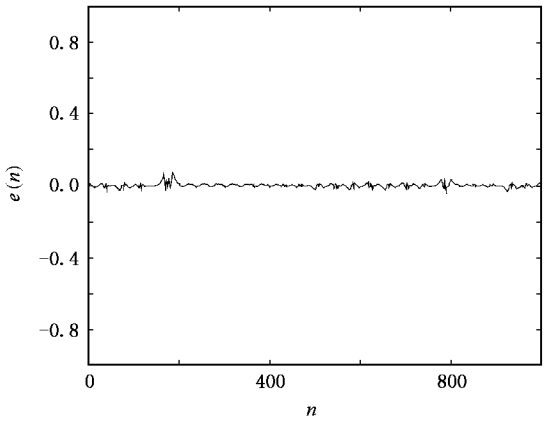


图 7　Lorenz 混沌时间序列的预测误差曲线

图 8 给出了 4 次实验的结果，图中横轴为实验序号，纵轴为 MSE. 每次实验的数据中，第 1 列为最差子 SVM 在测试集上的 MSE，第 2 列为最佳子 SVM 在测试集上的 MSE，第 3 列为使用本文提出的方法的 SVM 集成在测试集上的 MSE，第 4 列为使用传统方法的 SVM 集成在测试集上的 MSE.

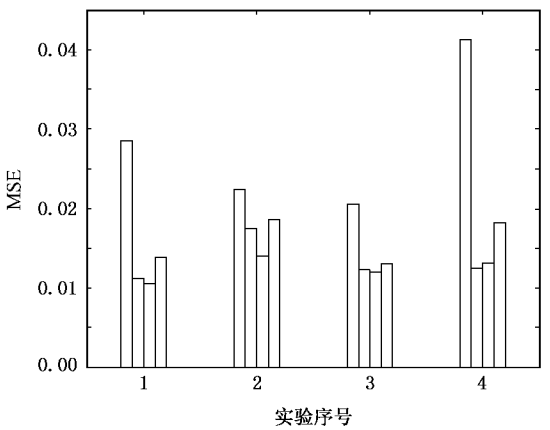


图 8　Lorenz 混沌时间序列的 MSE 比较

从以上实验结果可知，采用本文所提出的基于聚类的选择性集成方法都能取得比传统集成方法更好的效果. 从实验结果还可以看出，采用本文方法能

取得比单个最佳子 SVM 更好的或者接近的效果.而在实际应用 SVM 和 SVM 集成时,我们不能预先知道哪个子 SVM 泛化误差最小(或在测试集上误差最小),因此,本文提出的方法具有实际应用价值.

## 5. 结 论

本文所提出的基于聚类的选择性 SVM 集成建模方法,实现了混沌时间序列的建模和预测.对子 SVM 进行聚类并选择每一簇中精度最高的子 SVM 进行集成,可以保证子 SVM 有较高精度及子 SVM 之间有较大的差异度.选择的各子 SVM 实现对混沌时间序列在不同高维空间的扩展,能更好地提取其蕴藏的系统信息并加以综合,从而提高 SVM 集成的预测精度.这为混沌时间序列的预测提供了一条有效的途径.


[1] 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]

[2] Cui W Z, Zhu C C, Bao W X, Liu J H 2005 *Acta Phys. Sin.* **54** 3009 (in Chinese) [崔万照、朱长纯、保文星、刘君华 2005 物理学报 **54** 3009]

[3] Gan J C, Xiao X C 2003 *Acta Phys. Sin.* **52** 1097 (in Chinese) [甘建超、肖先赐 2003 物理学报 **52** 1097]

[4] Gan J C, Xiao X C 2003 *Acta Phys. Sin.* **52** 1102 (in Chinese) [甘建超、肖先赐 2003 物理学报 **52** 1102]

[5] 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]

[6] Ren R, Xu J, Zhu S H 2006 *Acta Phys. Sin.* **55** 555 (in Chinese) [任 韧、徐 进、朱世华 2006 物理学报 **55** 555]

[7] Zhang J S, Dang J L, Li H C 2007 *Acta Phys. Sin.* **56** 67 (in Chinese) [张家树、党建亮、李恒超 2007 物理学报 **56** 67]

[8] Cao L, Hong Y G, Fang H P, He G W 1995 *Physica* D **85** 225

[9] Farmer J D, Sidorowich J J 1987 *Phys. Rev. Lett.* **59** 845

[10] Tan W, Wang Y N, Zhou S W, Liu Z R 2003 *Acta Phys. Sin.* **52** 795 (in Chinese) [谭 文、王耀南、周少武、刘祖润 2003 物理学报 **52** 795]

[11] 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]

[12] Vapnik V N 1995 *The Nature of Statistical Learning Theory* (New York: Springer-Berlag)

[13] Kim H Y, Pang S N, Je H M, Kim D J, Bang S Y 2003 *Pattern Recognition* **36** 2757

[14] Giorgio V, Thomas G D 2004 *J. Mach. Learn. Res.* **5** 725

[15] Fu Q, Hu S X, Zhao S Y 2005 *J. Zhejiang Univ.* A **6** 387

[16] Kearns M, Li M, Valiant L 1994 *J. ACM* **41** 1298

[17] Leo B 1996 *Mach. Learn.* **21** 123

[18] Kohonen T 1990 *Proc. IEEE* **78** 1464

[19] 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

# Prediction of chaotic time series based on selective support vector machine ensemble*

Cai Jun-Wei Hu Shou-Song Tao Hong-Feng

(*College of Automation Engineering*, *Nanjing University of Aeronautics and Astronautics*, *Nanjing* 210016, *China*)

(Received 28 December 2006; revised manuscript received 8 June 2007)

Abstract

A clustering-based selective support vector machine ensemble forecasting model is presented. For improving the generalization ability of support vector machine ensemble, a hybrid clustering algorithm which combines the SOM and *K*-means algorithm is used to select the most accurate individual support vector machine from every cluster for ensembling, which ensures accuracy of individual support vector machines and improves the diversity of the individual support vector machines. This method can improve support vector machine ensemble generalization ability effectively with low cost. To illustrate the performance of the proposed forecasting model, simulations on chaotic time series prediction of the Mackey-Glass time series and the time series generated by the Lorenz systems are performed. The results show that the chaotic time series are accurately predicted, which demonstrates the effectiveness of this method.

**Keywords**: support vector machine, ensemble, chaotic time series, clustering

**PACC**: 0545

* Project supported by the Key Program of the National Natural Science Foundation of China (Grant No. 60234010) and the Aviation Science Foundation of China (Grant No. 05E52031).