基于 SVR 和 CAR 的多维时间序列分析及其在生态学中的应用

张永生,袁哲明*,熊洁仪,周铁军

(湖南农业大学生物安全科学技术学院,长沙 410128)

摘要:基于支持向量回归(SVR)并融合带受控项的自回归模型(CAR),建立了一种既反映样本集动态特征又体现环境因子影响的非线性多维时间序列分析预测方法(SVR-CAR)。用一步预测法对两个生态学样本集的预测结果表明,SVR-CAR 在所有参比模型中预测精度最高,并具结构风险最小、非线性、避免过拟合、泛化推广能力优异等诸多优点。SVR-CAR 在生态学、农业科学、经济学等多维时间序列预测领域有较广泛的应用前景。

关键词:多维时间序列;支持向量回归;非线性;预测;均方差

文章编号:1000-0933(2007)06-2419-06 **中图分类号**:Q141 **文献标识码**:A

Multidimensional time series analysis based on support vector regression and controlled autoregressive and its application in ecology

ZHANG Yong-Sheng, YUAN Zhe-Ming*, XIONG Jie-Yi, ZHOU Tie-Jun

Bio-safety Science and Technology College, Hunan Agricultural University, Changsha 410128, *China*

***Acta Ecologica Sinica*, 2007, 27(6): 2419 ~ 2424.**

Abstract: Based on support vector regression (SVR) and controlled autoregressive (CAR), we proposed a new non-linear multidimensional time series method named SVR-CAR that can show the dynamic characteristics of sample set as well as the effect of environmental factors. To evaluate the performance of SVR-CAR, we compared its predictions with those of four other commonly-used methods, using two sets of real-world data and one-step prediction. The results showed that SVR-CAR had the highest accuracy in prediction among the five methods, and had the advantages of structural risk minimization, non-linear characteristics, avoiding over-fit, and strong capacity for generalization. SVR-CAR has the potential to be widely used for predictions involving multidimensional time series data in ecology, agricultural sciences and economics.

Key Words: multidimensional time series; support vector regression; nonlinearity; forecast; mean square error

生态学积累了大量单因变量 y、多自变量 x_i 的多维时间序列数据。多维时间序列模型结合了一维时间序列分析和回归分析两类数理统计方法的优点,不仅考虑到事物发展的自身运动规律,也顾及了环境因子的作用[1,2]。经典的多维时间序列分析模型——带控制项的自回归滑动平均模型(controlled autoregressive

基金项目:国家自然科学基金资助项目(30570351);教育部新世纪优秀人才计划资助项目

收稿日期:2006-11-08;**修订日期**:2007-03-14

作者简介:张永生(980 ~),男,山西原平人,硕士生,主要从事模式识别与预测研究. E-mail:zhangysh@265.com

* **通讯作者** Corresponding author. E-mail:zhmyuan@sina.com

Foundation item: The project was financially supported by the National Natural Science Foundation of China (No. 30570351); Program for New Century Excellent Talents in University

Received date: 2006-11-08; **Accepted date**: 2007-03-14

Biography: ZHANG Yong-Sheng, Master candidate, mainly engaged in pattern recognition and forecast. E-mail: zhangysh@265.com

integrating moving average，CARMA）融合了时间序列分析和回归分析的优点但较为复杂，而任何 CARMA 模型均可以充分高阶的带受控项的自回归模型（controlled autoregressive，CAR）逼近到任意精度，因此可用 CAR 代替 CARMA 对动态系统实行统一建模[1~4]。由于生态学数据间往往更多地表现为非线性关系，CAR 模型的线性本质在实际应用中削弱了其预测能力。人工神经网络（artificial neural networks，ANN）具有很好的非线性逼近能力，基于 ANN 的多维时间序列分析已有应用[5,6]，但 ANN 本身存在基于经验最小化、模型结构难以确定、易于出现过度训练和训练不足、陷入局部最小、对连接权初值敏感、过度依赖设计技巧等诸多缺陷。基于统计学习理论的支持向量机（support vector machine，SVM）是目前发展最快的机器学习方法[7~9]，它最初用于模式识别（SVC），随 Vapnik 的 ε 不敏感损失函数的引入，SVM 已经扩展到用于非线性时间序列分析或非线性回归分析（SVR）[10~15]。SVM 基于结构风险最小，较好地解决了小样本、非线性、过拟合、维数灾和局极小等问题，泛化推广能力优异[7~9]；同时，本文使用的 LIBSVM 能自动搜索确定最佳惩罚参数、灵敏度及径向宽度等核函数参数，操作较 ANN 相对简便。

本文建立了一种基于 SVR 并融合 CAR 的非线性多维时间序列分析方法（SVR-CAR），在计算机上程序化实现并以两个生态学实例验证了 SVR-CAR 的有效性。

1 SVR-CAR 建模方法

支持向量机 LIBSVM2.8 软件包简单易用，含 4 个常用程序：svmscale 用于对原始数据规格化，svmtrain 用于训练，svmpredict 用于预测，gridregression.py 用于自动搜索核函数最优参数。各程序用法及其参数设置参见 http://www.csie.ntu.edu.tw/~cjlin/libsvm，有关 SVM 的基本原理参见文献[7~9,16]。

1.1 模型定阶

假定一多输入单输出回归模型有 N 个样本、一个因变量、$m-1$ 个自变量，由低阶到高阶递增地以 SVR 进行留一法测试（原始变量经 svmscale 规格化到 $[-1,+1]$），并依次对相邻模型采用 F 检验的方法判断模型阶次增加是否合适。对待比较的相邻两模型 SVR（n）和 SVR（$n+1$），有统计量：$F_i=\frac{Q_{\mathrm{SVR}(n)}-Q_{\mathrm{SVR}(n+1)}}{Q_{\mathrm{SVR}(n)}}\cdot\frac{N-mn-(n-1)}{m}$，$i=1,2,\cdots,n$ 服从自由度为（n，$N-mn-(n-1)$）的 F 分布；其中 $Q_{\mathrm{SVR}(n)}$ 为 SVR（n）的剩余离差平方和，$Q_{\mathrm{SVR}(n+1)}$ 为 SVR（$n+1$）的剩余离差平方和。

若 F_i 小于显著水平为 α、自由度为（n，$N-mn-(n-1)$）的 F 临界值，则 SVR（n）模型是合适的；反之，继续拓展阶数。计算 F_i 时，小均方取 $\frac{Q_{\mathrm{SVR}(n)}}{N-mn-(n-1)}$ 而不取 $\frac{Q_{\mathrm{SVR}(n+1)}}{N-m(n+1)-(n-1)}$ 对阶数拓展更为严格。因此当 $n=0$ 时，可保护性地取 $\alpha=0.1\sim1.0$；当 $n>0$ 时，取 $\alpha=0.05$。即由 SVR（0）拓展到 SVR（1）时标准放宽，进一步拓展阶数时标准从严。

1.2 变量筛选

假定多输入单输出回归模型最高阶 n 确定后有 N' 个样本、p 个输入变量，现以多轮末尾淘汰法从包含全部输入变量的 SVR 模型中以留一法（原始变量 svmscale 规格化到 $[-1,+1]$）经 F 测验逐次剔除不显著的变量。

对第一轮筛选，有统计量：$F_i=\frac{(Q_{(x_1,x_2,\cdots,x_{i-1},x_{i+1},\cdots,x_p)}-Q_{(x_1,x_2,\cdots,x_i,\cdots,x_p)})}{(Q_{(x_1,x_2,\cdots,x_i,\cdots,x_p)}/(N'-p-1))}$，$i=1,2,\cdots,p$ 服从自由度为（1，$N'-p-1$）的 F 分布。其中 $Q_{(x_1,x_2,\cdots,x_i,\cdots,x_p)}$ 为 p 个输入变量的剩余离差平方和，$Q_{(x_1,x_2,\cdots,x_{i-1},x_{i+1},\cdots,x_p)}$ 为剔除第 i 个输入变量后的剩余离差平方和。

若 $\min F_i$ 大于显著水平为 α、自由度为（1，$N'-p-1$）的 F 临界值，表明没有变量可剔除，汰选结束；反之，剔除第 i 个变量后进入下一轮筛选（注意此时 p 变为 $p-1$），直至没有变量可剔除为止。当样本太少，自由度 $N'-p-1\leqslant0$ 时，直接由（$Q_{(x_1,x_2,\cdots,x_{i-1},x_{i+1},\cdots,x_p)}-Q_{(x_1,x_2,\cdots,x_i,\cdots,x_p)}$）的正负决定变量的取舍。考虑到 SVR 对冗余变量有一定的包容能力，且过少的变量可能会影响 SVR 预测的稳健性。因此当 $p\leqslant2$ 时，可保护性地取 $\alpha=$

0.1 ~ 1.0;当 $p > 2$ 时,取 $\alpha = 0.05$。

1.3　预测评价指标

对全部 N 个样本,遍历核函数参数组合,经 svmtrain 建模并以 svmpredict 回代,能搜索到一组最优核函数参数组合使 SVR 对样本集的拟合达到极高精度,但使用该参数组合建模往往实际预测效果很差。事实上,在 LIBSVM 中 gridregression.py 并不提供针对全部 N 个样本回代的自动参数寻优,而以 n-fold 交叉验证(其极限是留一法)避免过拟合。一方面,至少对 ANN 和 SVR 而言过高的回代拟合精度并无多大实际意义;另一方面,对预测模型特别是多维时间序列模型人们真正感兴趣的是实际预测能力而非回代拟合结果。因此,本文以实际预测结果作为模型优劣的评价基准。

为避免单个样本预测的偶然性,视时间序列长短,规定至少连续选取时间序列最后 5 个以上样本作为预测样本。在预测第 i 个样本时,其后续未来样本不得参与建模训练;在预测第 $i+1$ 个样本时,第 i 个样本加入训练样本(一步预测)。预测结果优劣采用均方误差(mean squared error, MSE)和平均绝对百分误差(mean absolute percentage error, MAPE)作为评价指标:$\mathrm{MSE} = \frac{\sum (y - \hat{y})^2}{n}$, $\mathrm{MAPE}\% = \frac{\sum |y - \hat{y}| / y}{n} \times 100$,式中 y 为真值,$\hat{y}$ 为预测值,n 为预测样本数。对同一组预测样本,如 A 模型与 B 模型相比虽 MAPE 较大而 MSE 较小,则 A 模型预测更为稳健;因此,MSE 为主要评价指标。

SVR-CAR 以自编 C++程序通过调用 LIBSVM2.8 实现并经验证通过。基于一步预测法,参比模型多元线性回归(multiple linear regression, MLR)、时间序列趋势分析(包括一次滑动平均、一次指数平滑、线性回归、二次滑动平均、一次平滑、二次指数平滑、三次指数平滑等 7 种模型,默认参数设置)和 CAR 由 DPS6.55 给出[7],SVR 由 LIBSVM2.8 给出。SVR 采用与 SVR-CAR 相同核函数,原始变量 svmscale 规格化后以留一法 gridregression.py 寻优建模后预测。

2　实例分析

1970 ~ 1987 年小麦赤霉病预报要素值引自文献[8](表 1)。1983 ~ 1987 年后 5a 小麦赤霉病病穗率一步预测结果如表 2($n = 0$ 时模型定阶保护性取 $\alpha = 0.8$,$p \leqslant 2$ 时变量筛选保护性取 $\alpha = 0.2$)。从 MSE 和 MAPE% 可看出,SVR-CAR 的预测精度最高。尽管 CAR 和 SVR-CAR 均体现样本集动态特征和环境因子的综合影响,但非线性的 SVR-CAR 预测精度明显高于线性的 CAR,这表明该样本集属非线性动态系统,用线性模型进行预测是不合适的。CAR 劣于 MLR 是由 CAR 至少必然拓展一阶引起的(如果由 SVR(0)拓展到 SVR(1)时 MSE 增大,则这种必然拓展一阶的做法显然是不合适的);SVR-CAR 并不必然拓展一阶而要求一个给定的最低概率保证,并确保在 $\alpha = 1$ 时也不会出现由 SVR(0)拓展到 SVR(1)时 MSE 增大反而拓阶的情形。

杉木直径生长和气象资料引自文献[1](表 3)。树龄 19 ~ 24a 杉木年轮指数预测结果如表 4($n = 0$ 时模型定阶保护性取 $\alpha = 1$,$p \leqslant 2$ 时变量筛选保护性取 $\alpha = 0.2$;可能是由于数据集自身的原因,SVR 模型由 gridregression.py 得到最优参数进行训练时无支持向量,导致 SVR 模型每年预测值相同,故其预测值未列出)。同样地,从 MSE 和 MAPE% 看,SVR-CAR 在所有预测模型中预测精度最高。SVR-CAR 每年一步预测结果甚至均优于文献[1]中 CAR 的回代拟合结果,特别是 CAR 模型回代拟合较差的第 19、20、21 年,基于 SVR-CAR 模型第 20、21 年也得到了很好的预测,充分表明 SVR-CAR 具较强的泛化推广能力。

3　讨论

与 ANN 和 SVR 一样,由于不存在一个解析的表达式,SVR-CAR 对因子欠缺解释能力。SVR-CAR 的另一个弊端是核函数、保护性拓阶与保护性变量筛选 α 的选取仍然是经验性的;结合本文实例证实:对给定的 5 种常用核函数(线性核函数 $t = 0$,多项式核函数 $t = 1$、$d = 2$,多项式核函数 $t = 1$、$d = 3$,径向基核函数 $t = 2$,sigmoid 核函数 $t = 3$),在预测单一年份时不能依据留一法交叉测试 MSE 最小原则动态地选用核函数,即一个确定的样本集必须经验性地取一个确定的核函数。此外,尽管 LIBSVM 中的 gridregression.py 能自动搜索确定最优核函数参数,但它同时是导致 SVR-CAR 计算复杂度较高的主要原因;不使用 gridregression.py 而采用

默认核函数参数将明显降低预测精度。

表 1 1970 ~ 1987 年小麦赤霉病预报要素值

Table 1 Forecast factors of wheat scab from 1970 to 1987

年份 Year	y	x_1	x_2	x_3	x_4
1970	21.5	25.8	0.6	81.9	7
1971	5.3	25.4	0.2	82	8.4
1972	22.4	25.4	24	81.7	11.9
1973	68.5	24.3	28.9	78.2	11.6
1974	13.5	25.9	1	79.5	8.5
1975	56.7	24.5	9.8	78.9	10.7
1976	37.9	24.7	3.5	82.4	10.6
1977	50.2	25.1	66	78.8	13.1
1978	8.6	25.4	1.8	80.9	9.2
1979	16.8	25.9	2.8	81.8	9.5
1980	20.5	25.3	1.5	81.8	10.3
1981	42.5	24.9	23	70.7	12.7
1982	25.7	25.2	16.3	80.1	13.3
1983	72.3	25.4	84.3	79.3	10.2
1984	15.8	24.9	0	81.3	11.4
1985	25	25.4	24.6	80.7	8
1986	46	24.3	0	80.7	11.4
1987	56.65	24.9	11.3	79.8	14.5

y 病穗率 Diseased panic1e rate (%)；x_1 上年 7 月下旬到 8 月上旬平均最低气温 Mean minimum temperature from the last 10 days of July to the first 10 days of August last year (℃)；x_2 当年 1 月上旬雨量 Rainfall in the first 10 days of January (mm)；x_3 上年 7 ~ 9 月平均气温之和 Sum of mean temperature from July to September last year (℃)；x_4 当年 3 月中旬平均气温 Mean temperature in the middle 10 days of March (℃)

表 2 小麦赤霉病病穗率预测

Table 2 Forecasting the diseased panicle rate of wheat scab

年份 Year	True value	MLR	Time Series Analysis *	CAR	SVR **	SVR-CAR **
1983	72.3	59.0	30.4	108.7	60.5	72.8
1984	15.8	27.1	41.0	-12.6	30.6	15.0
1985	25.0	39.8	38.0	91.7	37.4	28.0
1986	46.0	42.4	36.4	27.3	32.0	55.0
1987	56.65	25.4	39.7	36.4	32.2	26.1
	MSE	302.7	588.3	1471.1	261.4	205.2
	MAPE%	42.4	64.1	114.7	46.6	18.3

* 采用 7 个时间序列模型中 MSE 最小的一次平滑模型预测值 Using the forecast values of linear smoothing model which has minimum MSE among seven time series analysis models

* * 采用多项式核函数 Using polynomial kernel ($=1$, $d=3$)

SVR-CAR 是基于 SVR 并融合时间序列分析和回归分析的非线性多维时间序列分析方法，具结构风险最小、非线性、适于小样本，能有效克服过拟合、维数灾和局极小，泛化推广能力优异、预测精度高等诸多优点。本文用两个生态学实例验证了其有效性，但其在生态学、农业科学、经济学等预测领域的广泛应用前景仍需更多研究实例的支持。

表 3　杉木直径生长和气象资料[1]

Table 3　Diameter growth of fir and climate data[1]

树龄 Age (a)	y	x_1	x_2	x_3	x_4	x_5
6	1.3125	–	–	–	–	–
7	1.1778	23.3	23.2	145	37	94.45
8	0.9286	23.3	23.2	156	43	54.55
9	1.05	22.6	22.6	176	26.5	95.55
10	1.0216	23.3	23.2	165	31	71.35
11	1.2	22.7	22.6	147	34	33.45
12	1.1818	23.3	22.2	146	30.5	41.85
13	1.0968	23.3	22.3	155	42.5	99.1
14	1.1034	21.8	21.7	138	36.5	90.65
15	0.8519	22.5	22.6	127	29.5	67.7
16	0.8077	22.5	22.6	168	36	64.15
17	0.875	22.1	22.2	164	48	130.25
18	1	22.2	22.2	180	38.5	102.6
19	0.8095	22	22.1	169	28	179.9
20	1.05	21.8	21.8	162	32.5	127.4
21	1.0526	22.6	22.7	164	30	80.45
22	1.1111	22.4	22.4	155	30	48.25
23	1	22.2	22.3	143	22.5	81.15
24	0.875	22.7	22.8	133	27.5	86.05
25	–	22.5	22.7	158	29.5	100.05

y 年轮指数 Age ring index;x_1 上年 15cm 年均地温 Annual average geotemperature under 15cm last year (℃);x_2 上年 20cm 年均地温 Annual average geotemperature under 20cm last year (℃);x_3 上年全年日均气温在 18 ~ 27℃日数 Days of mean temperature from 18 to 27 last year (d);x_4 上年 5 月和 6 月月平均最小相对湿度 Mean minimum relative humidity in May and June last year (%);x_5 上年 9 月和 10 月月平均降水量 Mean rainfall in September and October last year (mm)

表 4　杉木年轮指数预测

Table 4　Forecast/simulation values of the age ring index of fir

树龄 Age (a)	True value	MLR	Time Series Analysis*	CAR	Simulation values of CAR[1]	SVR-CAR**
19	0.8095	1.0240	0.9975	0.7709	0.9455	0.9112
20	1.0500	0.9360	0.8272	0.7933	0.9450	0.9941
21	1.0526	0.9800	1.0014	0.9066	0.9185	0.9999
22	1.1111	1.0460	1.0421	1.0783	1.0536	1.1555
23	1.0000	1.0380	1.0976	0.9139	0.9702	0.9930
24	0.875	1.0000	1.0211	0.8675	0.8894	0.8636
	MSE	0.0143	0.0205	0.0162	0.0087	0.0031
	MAPE%	11.37	13.66	9.25	8.22	4.82

* 采用 7 个时间序列模型中 MSE 最小的一次指数平滑模型预测值 Using the forecasted values of linear exponential smoothing model which has minimum MSE among seven time series analysis models

** 采用线性核函数 Using linear kernel (=0)

References:

[1] Zhou L Y, Fei H X, Zhang X X. The application of multiple dimension time series analysis method in long-term forecasting of rice leaf roller. Acta Phytophylacica Sinica, 1995, 22 (): 1 - 6.

[2] Wu C Z, Hong W. Multidimensional time series analysis on tree growth. Chinese Journal of Applied Ecology, 1999, 10 (4): 395 - 398.

[3] Deng Z L, Guo Y X. Analysis of dynamic system and its application. Shenyang: Liaoning Science and Technology Press, 1985. 31 - 76.

[4] Wu C Z, Hong W. A proposed multidimensional time series model of individual age and diameter in tsuga longibracateata. Acta Phytoecologica Sinica, 2002, 26 (4): 403 - 407.
[5] Chakraborty K, KMohan C. Forecasting the behavior of multivariate time series using neural networks. Neural Networks, 1992, 5: 961 - 970.
[6] Hu Z, Jia Y L. Neural network model of dynamic research for injection water oil fields. Journal of Southwest Petroleum Institute, 2003, 25 (6): 33 - 35.
[7] Vapnik V N. The nature of statistical learning theory. New York: Springer Verlag Press, 1995.
[8] Deng N Y, Tian Y J. Support vector machine — a new method in data mining. Beijing: Science Press, 2004. 77 - 162, 224 - 272.
[9] BURGES C J C. A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery, 1998, 2: 121 - 167.
[10] Ma X G, Hu F. Forecasting the concentration of air pollutant using support vector machine. Progress in Natural Science, 2004, 14 (6): 349 - 353.
[11] Sun D S, Wu J P, Xiao J H. The application of SVR to prediction of chaotic time series. Journal of System Simulation, 2004, 16 (6): 519 - 521.
[12] Pai P F, Hong W C. Support vector machines with simulated annealing algorithms in electricity load forecasting. Energy Conversion and Management, 2005, 46: 2669 - 2688.
[13] Thissena U, Brakela R van, Weijerb A P de, *et al*. Using support vector machines for time series prediction. Chemometrics and Intelligent Laboratory Systems, 2003, 69: 35 - 49.
[14] Pai P F, Lin C S. A hybrid ARIMA and support vector machines model in stock price forecasting. Omega, 2005, 6: 497 - 505.
[15] Chen K Y, Wang C H. A hybrid SARIMA and support vector machines in forecasting the production values of the machinery industry in Taiwan. Expert Systems with Applications, 2007, 32: 254 - 264.
[16] Fan R E, Chen P H, Lin C J. Working set selection using the second order information for training SVM. Journal of Machine Learning Research, 2005, 6: 1889 - 1918.
[17] Tang Q Y, Feng M G. DPS data processing system for practical statistics. Beijing: Science Press, 2002. 525 - 585.
[18] Zhou C H. Principal components in epidemic factors of wheat scab and a forecasting model. Acta Phytophylacica Sinica, 1990, 17 (4): 317 - 321.

参考文献：

[1] **周立阳，费惠新，张孝羲. 多维时间序列分析在稻纵卷叶螟长期预测预报上的试用. 植物保护学报，**1995，22 (): 1 ~ 6.
[2] **吴承祯，洪伟. 林木生长的多位时间序列分析. 应用生态学报，**1999，10 (4): 395 ~ 398.
[3] **邓自立，郭一新. 动态系统分析及其应用，沈阳 辽宁科学技术出版社，**1985. 31 ~ 76.
[4] **吴承祯，洪伟. 长苞铁杉种群个体年龄与胸径的多维时间序列模型研究. 植物生态学报，**2002，26 (4): 403 ~ 407.
[6] **胡 泽，贾永禄. 油田产量预报的多维时间序列神经网络模型. 西南石油学院学报，**2003，25 (6): 33 ~ 35.
[8] **邓乃扬，田英杰. 数据挖掘中的新方法——支持向量机. 北京 科学出版社，**2004. 77 ~ 162，224 ~ 272.
[0] **马晓光，胡非. 利用支撑向量机预报大气污染物浓度. 自然科学进展，**2004，14 (6): 349 ~ 353.
[1] **孙德山，吴今培，肖健华.** SVR **在混沌时间序列预测中的应用. 系统仿真学报，**2004，16 (6): 519 ~ 521.
[7] **唐启义，冯明光. 实用统计分析及其** DPS **数据处理系统. 北京 科学出版社，**2002. 525 ~ 585.
[8] **周崇和. 小麦赤霉病流行因子的主成分分析及预测模型探讨. 植物保护学报，**1990，17 (4): 317 ~ 321.