DOI:10.3724/SP.J.1238.2010.00430

基于 ARIMA 和 DSVM 组合模型的松毛虫发生面积预测

向昌盛[a]，周子英[b]，武丽娜[a]

(湖南农业大学 a.东方科技学院；b.资源环境学院，湖南 长沙 410128)

摘　要：提出一种基于 ARIMA 和动态 ε 支持向量机(ε-DSVM)的组合预测模型(ARIMA-ε-DSVM)，预测松毛虫发生面积．先采用 ARIMA 模型进行时间序列线性趋势建模，为非线性部分确定输入阶数，根据确定的输入阶数进行时间序列样本重构，再采用 ε-DSVM 模型进行时间序列非线性特征建模，将这两模型预测值相加得到组合模型预测值．对辽宁省朝阳市松毛虫时间序列进行仿真试验，结果表明，ARIMA-ε-DSVM 模型预测精确度比单一模型 ARIMA 和 SVM 及简单组合模型 ARIMA-SVM 要高，ARIMA-ε-DSVM 模型大幅度改善预测效果，显著地减少预测误差，泛化能力强.

关　键　词：支持向量机；松毛虫；时间序列；差分自回归移动平均

中图分类号：S431　　文献标志码：A　　文章编号：1007-1032(2010)04-0430-04

Dendrolimus punctatus forecasting based on hybrid ARIMA and dynamic SVM model

XIANG Chang-sheng[a], ZHOU Zi-ying[b], WU Li-na[b]

(a.College of Orient Science and Technology; b. College of Resources and Environment, HNAU, Changsha 410128, China)

Abstract: A novel forecasting model combinating autoregressive integrating moving average(ARIMA) with dynamic ε-insensitive cost function support vector machine(ε-DSVM)was brought forth, which showed the complicated and dynamic characteristics of *Dendrolimus punctatus* occurrence. ARIMA model was used to capture the linear feature of the time series and ε-DSVM model to fit the nonlinear component of the time series to obtain the ensemble forecasting result by adding ARIMA to ε-DSVM. The prediction performances of the method was tested by *Dendrolimus punctatus* occurrence, and the results showed that the hybrid model, which took advantage of the unique strength of the two models in linear and nonlinear modeling, had better accuracy than the single model and simple ensemble forecasting model incorporating ARIMA and SVM. As a novel model combinated ARIMA with ε-DSVM , the combinatining model had the advantages of structural risk minimization and non-linear characteristics, which was suitable for small samples, being able to avoid the over-fit. It is a new, powerful tool in pests forecasting work.

Key words: support vector machines; *Dendrolimus punctatus*; time series; auto regression integrated moving average(ARIMA)

近年来，国内外学者运用一些经验模型或基于统计回归方法和时间序列预测模型等，对松毛虫发生面积的预测展开了广泛的研究，但对于发生既有全域性又有区域性，在时间上表现出无序的不稳定性，有序的规律性和周期性的松毛虫发生系统，用统计回归法很难用解析方法或确切的公式表达[1-2].

收稿日期：2010-01-17

基金项目：国家自然科学基金项目(30570351)

作者简介：向昌盛(1971—)，男，湖南怀化人，博士，副教授，从事生物信息学及农业昆虫与害虫防治研究，cx5243879@sohu.com

时间分析方法中最具代表性的是基于线性的差分自回归移动平均(autoregressive integrating moving average，ARIMA)，由于ARIMA无法捕捉松毛虫发生过程中的非线性数据的特征，导致其预测准确率不高[3-4]．20 世纪 80 年代以来，非线性的神经网络算法迅速发展，为松毛虫的预测研究开拓了新的空间[5-6]，但神经网络是基于经验风险最小化原则，要求数据样本大，而实际松毛虫发生面积的历史数据属于小样本数据，往往不能满足大样本这一要求，所以在预测过程中容易出现结果过拟合、泛化能力差等现象[7]．基于结构风险最小化的支持向量机(support vector machines，SVM)，是一种新的机器学习方法，较好地解决了小样本、非线性、过拟合、维数灾和局极最优等问题，且泛化推广能力优异，在害虫预测领域里得到了广泛应用[8-9]．

基于著名的M-竞争理论[10]，一些学者利用ARIMA和SVM的组合模型(ARIMA-SVM)来进行时间序列预测研究，用ARIMA捕捉线性特征，再用SVM揭示非线性规律，结果表明，ARIMA-SVM能够较大限度地利用各种预测样本信息，比单个预测模型考虑问题更系统，更全面[11-12]．Tay等[13]提出折扣最小平方法(discounted least squares，DLS)，把输入输出间关系的逐渐变动列入考虑，对较近期资料的误差，给予较大的惩罚，但ARIMA-SVM并没有把时间因素考虑进去，会造成预测准确率有时不佳[14]．

笔者拟使用具有更强的动态行为和计算能力，能处理时间间动态变化的动态ε支持向量机(dynamic ε-insensitive cost function support vector machines，ε-DSVM)并结合ARIMA模型，产生一种新的时间序列预测模型，即ARIMA-ε-DSVM，对松毛虫发生面积进行预测，旨在探讨有效的松毛虫灾害预测技术，提高松毛虫的综合治理水平.

1　材料与方法

1.1　数据来源及处理

数据来源于辽宁省朝阳市 1986—2006 年松毛虫发生面积(表 1)[15]．由于对预测模型的评价主要需考察其预测能力而非回代拟合结果，因此，对数据分别进行拟合和独立预测．将 1986—2001 年数据作为训练样本进行拟合和建模．为避免单个样本预测的偶然性，以 2002—2006 年连续 5 年数据作为独立测试集，采用一步预测，检验模型的泛化能力.

表 1　辽宁朝阳市历年松毛虫发生面积

Table 1　*Dendrolimus punctatus* occurrence of Chaoyang in Liaoning province　　$\times 10^4$ hm^2

年份	发生面积	年份	发生面积
1986	9.0	1997	11.4
1987	13.8	1998	27.0
1988	17.4	1999	20.4
1989	16.2	2000	20.4
1990	10.8	2001	29.4
1991	9.6	2002	18.0
1992	16.2	2003	26.4
1993	20.4	2004	19.2
1994	21.0	2005	22.8
1995	15.0	2006	25.9
1996	21.6		

1.2　研究方法

1.2.1　线性部分预测模型

线性部分预测模型采用ARIMA[16]，其建模过程为：1) 样本平稳化处理；2) 模型定阶；3) 模型检验；4) 预测.

1.2.2　非线性部分预测模型

为了提高SVM的预测准确度，采用体现时间结构特征的ε-DSVM模型对非线性部分进行预测，令ε值[17]进行动态变化，近期数据ε值相对较小，而远期数据ε值相对较大，这样ε-DSVM就能够同时捕捉时间序列的非线性和动态特征.

1.2.3　ARIMA-ε-DSVM 模型

由于在实际应用中，很难完全知道数据的特点，因此结合线性和非线性方法去估计较为合理．首先通过 ARIMA 建立线性预测模型，得到数据的线性预测值，再用 ε-DSVM 模型对数据的非线性数据残差序列进行预测，并调整模型参数，得到最佳参数时的非线性部分预测值，最后 2 种模型值相加得到组合模型的预测值.

1.2.4 参比模型及评价指标

为了考察 ARIMA-ε-DSVM 的预测效果，选择 SVM、ARIMA、ARIMA-SVM 作为参比模型，并以均方误差(*MSE*)和平均绝对误差百分比(*MAPE*)作为模型性能评价指标 .ARIMA 由 DPS6.55 给出，SVM 使用 LIBSVM 2.86 实现，ARIMA-ε-DSVM 程序在 MATLAB 7.01 平台和 LIBSVM 工具箱编程实现.

2 结果与分析

2.1 朝阳市松毛虫发生面积的ARIMA模型

利用 DPS6.55 构建 ARIMA 模型，首先利用差分对数据进行平稳化，发现 3 阶差分后，数据已经基本平稳化，所以确定 ARIMA 模型参数 d=3．采用从低阶到高阶逐步试探法来识别模型的类型和阶数 ,经过比较分析 ,发现选择 ARIMA(5,4,3)模型，拟合效果(表 2)较好 ,进行松毛虫发生面积预测 ,预测样本的 *MSE* 为 5.33 ,说明 ARIMA 对线性部分进行了较好的拟合.

2.2 朝阳市松毛虫发生面积的ε-DSVM模型

模型定阶采用袁哲明[18]提出的基于SVM非线性定阶方法，模型的阶数确定为3．从模型阶数可知，当年松毛虫发生面积的残差受到前2年的发生面积残差的影响较大，这就意味着需将前2年的松毛虫发生面积的残差作为ε-DSVM的输入进行样本重构 ,组成新的样本集 .由于针对ε参数随时间赋予不同的权重 ,故需先决定采用传统的SVM进行参数寻优，得到最优参数(C,δ,ε)=(400,2,20)，先暂定(C,δ)=(400,2)，并将ε的初值设为40，使得当i=7时，ε收敛于最佳值20，将a的范围设为[- 14,0]，用ε-DSVM进行建模预测 ,得到 :当a为 - 5时 ,ε-DSVM模型*MSE*最小，最后以(C,δ,a)=(400,2, - 5)为参数，利用ε-DSVM对2002—2006年松毛虫发生面积残差序列进行预测，得到非线性部分的预测结果.

2.3 ARIMA-ε-DSVM预测结果及性能分析

将 ARIMA 模型得到的线性预测结果和ε-DSVM 模型的非线性预测结果进行简单相加，得到 ARIMA-ε-DSVM 模型的预测结果，各模型拟合值和预测结果的 *MSE* 和 *MAPE* 分别见表 2 和表 3 .

表 2 各种模型对 1986—2001 年朝阳市松毛虫发生面积的拟合值

Table 2 Fitted values of models for occurred *Dendrolimus punctatus* of Chaoyang city in 1986—2001 $\times10^4$ hm^2

年份	面积拟合值			
	ARIMA	SVM	ARIMA-SVM	ARIMA-ε-DSVM
1986	9.7	10.1	9.3	9.2
1987	11.9	12.8	13.0	13.4
1988	14.9	17.4	17.6	17.7
1989	16.7	17.3	16.5	16.5
1990	10.4	12.1	11.7	11.3
1991	12.6	11.9	10.6	10.4
1992	16.8	16.2	16.5	16.5
1993	20.3	19.4	20.4	20.6
1994	19.0	19.9	20.7	21.1
1995	14.9	16.5	15.3	15.3
1996	23.8	22.6	21.9	21.9
1997	12.8	13.8	12.1	11.8
1998	28.6	25.1	28.1	27.7
1999	20.8	20.7	20.6	20.7
2000	22.6	22.5	22.7	21.8
2001	25.9	27.6	28.1	29.1

表 3 各种模型的预测误差

Table 3 The forecasting errors of various models

模型	*MSE*	*MAPE*
ARIMA	5.33	10.50
SVM	5.45	8.80
ARIMA-SVM	1.57	5.50
ARIMA-ε-DSVM	1.17	3.87

(1) 从表3中可知 ,非线性的SVM模型和线性的ARIMA模型的预测准确度不高 ,主要是因为它们都不能同时捕捉到线性和非线性特征，故对松毛虫发生面积这种复杂时间序列数据进行预测时，预测效果不好.

(2) ARIMA-SVM模型与ARIMA、SVM相比，预测准确度有了较大的提高，说明组合模型可以提升预测准确度，这是因为2个模型结合时可以互补，相对单一模型具有一定的优势．而且ARIMA-SVM是先用ARIMA模型求得线性部分的预测值 ,因此在神经网络中经常发生的过学习现象不会发生.

(3) ARIMA-ε-DSVM的*MSE*和*MAPE*值远低于其他3种模型，其泛化能力最强，ARIMA-ε-DSVM为松毛虫发生面积的最佳预测模型，证明ε-DSVM把数据的时间相关性考虑进去，令ε随时间变化，越近期的数据对预测结果影响越大，越远期的数据对预测结果影响则越小，改变了传统支持向量机固定ε的作法，这样的处理较为合理.

3　讨　论

将 ARIMA 和 ε-DSVM 模型相结合，构建了一种松毛虫发生面积时间序列组合预测方法 ARIMA-ε-DSVM 模型，并以辽宁朝阳市 1981—2006 年松毛虫发生面积预测为例，对预测准确度进行评估，预测结果与实际情况吻合．由于考虑了数据时间相关性，ARIMA-ε-DSVM 对松毛虫发生面积作出了比较准确有效的预测，比 ARIMA 和 SVM 及 ARIMA-SVM 模型有更高的预测准确率.

由于数据获得困难，仅对松毛虫发生面积的一维时间序列进行预测，而松毛虫的发生面积不仅与环境因素有关，也受到天敌、寄主营养、林木抗性等的影响，是一个多维时间序列问题，因此，随着预测时间的延长，ARIMA-ε-DSVM 的预测准确性会受到影响，尚需要考虑多因素影响，并需及时补充新的资料来修正预测模型，以提高预测的准确度.

参考文献：

[1] 陈绘画，崔相富，朱寿燕．马尾松毛虫发生量灰色系统模型的建立及其预报[J]．东北林业大学学报，2004，32(4)：19-21．

[2] 柴守权，许国莲，卢南．云南松毛虫发生期与危害程度预测预报研究[J]．西北林学院学报，2002，17(2)：54-57．

[3] 秦华光，李家才，穆丹，等．时间序列自回归模型预测茶园小绿叶蝉种群动态的探讨[J]．安徽农业大学学报，2008，35(4)：564-570．

[4] 贾春生．ARIMA 模型在马尾松毛虫发生面积预测中的应用[J]．安徽农业科学，2007，35(19)：5672-5673．

[5] 陈绘画，朱寿燕，崔相富．基于人工神经网络的马尾松毛虫发生量预测模型的研究[J]．林业科学研究，2003，16(2)：159-165．

[6] 朱寿燕，陈绘画，崔相富．应用人工神经网络预测马尾松毛虫的有虫面积[J]．中国农业气象，2004，25(1)：51-53．

[7] 张爱兵，陈建，王正军，等．BP 网络模型和 LOGIT 模型在森林害虫测报上的应用初报——以安徽省潜山县马尾松毛虫为例[J]．生态学报，2002，21(12)：2159-2166．

[8] 谭泗桥，林雪梅，陈渊，等．基于地统计学的多维时间序列模型及其在生态学中的应用[J]．湖南农业大学学报：自然科学版，2009，35(4)：433-437．

[9] 石晶晶，刘占宇，张莉丽，等．基于支持向量机(SVM)的稻纵卷叶螟危害水稻高光谱遥感识别[J]．中国水稻科学，2009，23(3)：331-334．

[10] Bate J M，Granger C W．The combination of forecasts [J]．Operational Research Quarterly，1969，20(1)：451-468．

[11] Pai P F，Lin C．S．A hybrid ARIMA and support vector machines model in stock price forecasting[J]．Omega，2005，33(6)：497-505．

[12] Huseyin Ince，Theodore B Trafalis．A hybrid model for exchange rate prediction[J]．Decision Support Systems，2005，9(1)：1-9．

[13] Tay F E，Cao L J．Modified support vector machines in financial time series forecasting[J]．Neurocomputing，2002，48(4)：847-861．

[14] Huang W，Nakamori Y，Wang S Y．Forecasting stock market movement direction with support vector machine [J]．Computers & Operations Research，2005，32(10)：2513-2522．

[15] 刘青松．时间序列分析方法在预测松毛虫发生面积中的应用[J]．河北农业科技，2008(12)：58-59．

[16] Zhang G P．Time series forecasting using a hybrid ARIMA and neural network model[J]．Neurocomputering，2003，50(1)：159-175．

[17] Huang Y S，Wang B L．A hybid arima and adaptive SVM model in forecasting stock market index [J]．Journal of E-Business，2008，10(4)：1041-1066．

[18] 袁哲明，张永生，熊洁仪．基于SVR的多维时间序列分析及其在农业科学中的应用[J]．中国农业科学，2008，41(8)：2485-2492．

责任编辑：罗慧敏
英文编辑：胡东平