湖南农业大学学报(自然科学版)　2015 年 4 月　第 41 卷 第 2 期
Journal of Hunan Agricultural University (Natural Sciences), Apr. 2015, 41(2):172–176

DOI:10.13331/j.cnki.jhau.2015.02.011
投稿网址：http://xb.hunau.edu.cn

基于混沌理论的害虫发生量非线性建模与预测

向昌盛[1,2]，万方浩[2,3,4*]

(1. 湖南工程学院计算机与通信系，湖南 湘潭 411104；2. 植物病虫害生物学国家重点实验室，北京 100193；3. 中国农业科学院植物保护研究所，北京 100193；4. 青岛农业大学农学与植物保护学院，山东 青岛 266109)

摘　要：为了提高害虫发生量预测的精度，提出一种基于混沌理论的害虫发生量非线性预测模型(PSR–LSSVM)。通过相空间重构对害虫发生量时间序列进行重构，将重构后的害虫发生量序列输入到最小二乘支持向量机进行学习，建立害虫发生量预测模型，采用云南省普洱市思茅区和浙江省仙居县的松毛虫发生面积数据对模型性能进行检验。结果表明，松毛虫发生面积预测值与实际发生值十分接近，2 个地区松毛虫发生面积预测结果的平均绝对百分误差分别为 0.90%和 2.44%，预测结果要优于 BP 神经网络、线性预测模型。

关　键　词：害虫发生量；最小二乘支持向量机；预测模型；相空间重构

中图分类号：TP181　　文献标志码：A　　文章编号：1007–1032(2015)02–0172–05

Nonlinear modelling and prediction of pest's occurrence quantity by chaotic theory

Xiang Changsheng[1,2]，Wan Fanghao[2,3,4*]

(1.Department of Computer and Communication, Hunan Institute of Engineering, Xiangtan, Hunan 411104 , China; 2.State Key Laboratory for Biology of Plant Diseases and Insect Pests, Beijing 100193, China; 3.Institute of Plant Protection, Chinese Academy of Agricultural Sciences, Beijing 100193, China; 4.College of Agriculture and Plant Protection, Qingdao Agricultural University, Qingdao, Shangdong 266109, China)

Abstract: In order to improve the prediction accuracy of pest's occurrence quantity, this paper proposed a nonlinear prediction model for pest's occurrence quantity based on chaotic theory. Time series for pest's occurrence quantity were reconstructed by phase space reconstruction, and then input into the least squares support vector machine to learn and establish prediction model for the pest's occurrence quantity, the test experiment is carried out using *Dendrolimus punctatus* occurrence area data in Simao, Puer and Xianju county, Zhejiang. The results show that the prediction values of *Dendrolimus punctatus* occurrence area were very close to the actual production, the mean absolute percent error of predicted results for *Dendrolimus punctatus* occurrence area in in Simao, Puer and Xianju county, Zhejiang were 0.90% and 2.44% respectively, the prediction results were better than that of BP neural network and linear prediction model.

Keywords: pest's occurrence quantity; least squares support vector machine; prediction model; phase space reconstruction

害虫灾害的发生是一种高度复杂的自然现象，人们对其规律进行长期探索，积累了大量的历史数据，利用这些历史数据对害虫发生趋势进行预测，能够为害虫防治提供有价值的参考意见；因此，害虫发生量建模与预测具有重要的研究意义[1]。

传统的害虫预测模型主要有多元线性回归、自回归滑动平均、趋势外推法等[2–3]。由于受到气象、作物种类、栽培管理等因素的综合影响，加上害虫发生自身独特的变化规律，使得害虫发生具有规律性和周期性及不稳定性，造成传统的模型预测精度

收稿日期：2014–12–11　　修回日期：2015–01–08
基金项目：国家“973”计划项目(2009CB119200)；公益性行业(农业)科研专项(201303024–04；201303019–02)；湖南省自然科学基金项目(2015jj2041)
作者简介：向昌盛(1971—)，男，湖南怀化人，博士，副教授，主要从事生物信息学和人工智能技术研究，cx5243879@sohu.com；*通信作者，万方浩，博士，研究员，主要从事生物入侵、生物防治及昆虫生态研究，wanfanghao@caas.cn

较低[4]。随着人工智能技术发展，出现了基于神经网络、支持向量机等机器学习算法的害虫发生量预测模型，它们可以较好地描述害虫的非线性发生趋势，提高了预测精度[5–7]。在害虫发生量建模过程中，神经网络基于结构风险最大原理，要求训练数量比较大，而害虫发生量预测是一种小样本预测问题，导致神经网络常出现“过拟合”、训练速度慢、陷入局部极小等缺陷；支持向量机虽然具有良好的泛化能力，但是训练时间相对较长，建模效率低，难以满足害虫发生预测实时性要求。最小二乘支持向量机(LSSVM)不仅克服了神经网络存在的缺陷，而且解决了支持向量机建模效率低的问题，泛化性能优异[8–10]。大量研究结果表明，害虫的发生具有弱混沌性，而当前害虫发生量预测建模过程中，害虫发生的混沌性常常被忽略，导致害虫预测精度有待提高[11]。近年来，随着混沌理论研究的不断深入，将混沌理论和机器学习方法相结合，为害虫发生预测开辟了一种新的途径[12]。

由于害虫发生量的非线性、混沌、突变性等特点，至今没有统一的预测模型。为了客观、准确描述害虫发生量的变化趋势，以提高害虫发生量的预测精度为目标，笔者提出一种基于混沌理论的害虫发生量非线性预测模型(PSR–LSSVM)。首先采用混沌理论的相空间重构(phase space reconstruction，PSR)[13]，将一维的害虫发生量时间序列转化成矩阵形式，挖掘隐藏于害虫发生量历史数据中的信息，再采用任意非线性映射和学习逼近能力的 LSSVM 对构造的时间序列进行训练，建立害虫发生量预测模型，揭示害虫复杂的发生过程，以期为害虫发生量研究提供一种新的简单、有效的方法。

1　害虫发生量非线性预测模型

1.1　相空间重构

为了研究时间序列数据的混沌性，Packard[14]提出相空间重构理论，可以从实测时间序列中的某一分量了解非线性动力系统相空间的几何特性，并在高维相空间中恢复混沌吸引子。Takens 等[15]证明了重构系统与原始系统在系统特征上具有等价关系。对于一维混沌时间序列 $x(i)$，i=1,2,⋯,n 进行相空间重构，重构后的 m 维状态向量可表示为：

$$\boldsymbol{X}(i)=\left(x(i),x(i+\boldsymbol{\tau}),\cdots,x(i+(m-1)\boldsymbol{\tau}\right)^{\mathrm{T}}\quad i=1,2,\cdots,N \tag{1}$$

式中：$N=n-(m-1)\tau$ 为相点的个数；m 为嵌入维数；τ 为延迟时间。

1.2　最小二乘支持向量机

对于训练集$\{(x_i,y_i)\}$，i=1,2,⋯,n，LSSVM 通过非线性映射函数 $\Phi(\cdot)$将其映射到高维特征空间进行线性回归，则有：

$$f(x)=\boldsymbol{\omega}^{\mathrm{T}}\Phi(x)+b \tag{2}$$

式中：$\boldsymbol{\omega}$ 为权值向量；b 为偏置量。

综合考虑预测性能和建模效率，式(2)可以变为约束条件的优化问题：

$$\begin{aligned}&\min\|\boldsymbol{\omega}\|^2+\frac{1}{2}\gamma\sum_{i=1}^{n}\xi_i^2\\&\text{s.t.}\\&y_i-\boldsymbol{\omega}^{\mathrm{T}}\varphi(x)+b=e_i(i=1,2,\cdots,n)\end{aligned} \tag{3}$$

式中：γ 为正则化参数；e_i 为预测误差。

引入拉格朗日乘子[16]，将式(3)变为对偶问题，即：

$$\begin{aligned}L(\boldsymbol{\omega},b,\zeta,\alpha)=\min\|\boldsymbol{\omega}\|^2+\frac{1}{2}\gamma\sum_{i=1}^{n}\xi_i^2+\\\sum_{i=1}^{n}\alpha_i(\boldsymbol{\omega}^{\mathrm{T}}\Phi(x)-b+e_i-y_i)\end{aligned} \tag{4}$$

式中：α_i 为拉格朗日乘子。

选择 RBF 核函数作为 LSSVM 核函数，定义为：

$$K(x_i,x_j)=\exp(-\|x_i-x_j\|^2/2\sigma^2) \tag{5}$$

最后，LSSVM 回归模型为：

$$f(x)=\sum_{i=1}^{N}\alpha_i\exp(-\|x_i-x_j\|^2/2\sigma^2)+b \tag{6}$$

从 LSSVM 建模过程可知，LSSVM 预测性能与其参数相关，因此采用粒子群优化算法对 LSSSVM 参数进行选择。

1.3　害虫发生量的非线性预测模型

设害虫发生量的历史数据时间序列为$\{x_1,x_2,\cdots,x_n\}$，x_n 是预测目标值，建立害虫发生量输入 $x=\{x_{n-1},x_{n-2},\cdots,x_{n-m}\}$与输出 $y=\{x_n\}$之间的非线性映射关系：$R^m\rightarrow R$，m 为映射维数。通过 m 对害虫发生量时间序列$\{x_1,x_2,\cdots,x_n\}$进行样本重构，得到 LSSVM 的学习样本为：

$$\boldsymbol{X}=\begin{bmatrix}x_1&x_2&\cdots&x_m\\x_2&x_3&\cdots&x_{m+1}\\\vdots&\vdots&&\vdots\\x_{n-m}&x_{n-m+1}&\cdots&x_{n-1}\end{bmatrix},\boldsymbol{Y}=\begin{bmatrix}x_{m+1}\\x_{m+2}\\\vdots\\x_n\end{bmatrix} \tag{7}$$

则害虫发生量时间序列的 LSSVM 回归函数为：

$$y_t = \sum_{i=1}^{n-m}(\alpha_i - \alpha_i^*)K(x_i, x_{n-m+1}) + b \quad (8)$$

式中：t=m+1,m+2,…,n。

基于LSSVM的害虫发生量时间序列预测模型为：

$$\hat{x}_{n+1} = \sum_{i=1}^{n-m}(\alpha_i - \alpha_i^*)K(x_i, x_{n-m+1}) + b \quad (9)$$

式中：x_{n-m+1}={x_{n-m-1},x_{n-m-2},…,x_n}。

2 应用实例

2.1 数据源

为避免单个数据集的偶然性，选取云南省普洱市思茅区 1985—2006 年松毛虫发生面积(表 1)和浙江省仙居县 1983—1989 年松毛虫发生面积(表 2)作为研究对象。浙江省仙居县松毛虫发生面积每年采集 3 次数据，收集时间分别为 4 月(出蛰)、7 月(第一代)和 9 月(第二代)。分别取 1983—2003、1983—1988 年数据作为训练集，各数据集最后 3 个数据作为测试集。

表 1 1985—2006 年普洱市思茅区松毛虫发生面积

Table 1 Occurrence area of *Dendrolimus punctatus* in Simao, Puer in 1985–2006 hm^2

年份	发生面积	年份	发生面积	年份	发生面积
1985	0.27	1993	0.41	2001	0.30
1986	0.20	1994	1.00	2002	0.17
1987	0.34	1995	0.90	2003	0.29
1988	0.71	1996	0.19	2004	0.37
1989	0.87	1997	0.29	2005	0.73
1990	0.27	1998	0.35	2006	1.00
1991	0.17	1999	0.70		
1992	0.29	2000	1.19		

表 2 1983—1989 年浙江省仙居县松毛虫发生面积

Table 2 Occurrence area of *Dendrolimus punctatus* in Xianju county, Zhejiang in 1983–1989 hm^2

时间	发生面积	时间	发生面积
1983–04	4 333.3	1986–09	3 135.6
1983–07	1 912.7	1987–04	3 205.3
1983–09	6 383.3	1987–07	2 828.3
1984–04	4 826.0	1987–09	4 772.5
1984–07	4 510.3	1988–04	4 018.0
1984–09	6 051.7	1988–07	5 341.1
1985–04	4 210.3	1988–09	4 572.1
1985–07	4 972.7	1989–04	4 309.7
1985–09	3 941.0	1989–07	4 316.0
1986–04	3 106.0	1989–09	5 309.3
1986–07	3 097.6		

2.2 归一化处理

由于 LSSVM 的核函数值通常依赖特征向量的内积，而较大的属性值影响训练效率，需在建模之前，对数据进行归一化处理，将其值缩放到[0,1]之间，具体为：

$$x_i' = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \quad (10)$$

式中：x_i 是原始数据；x_i' 为归一化后的数据；$x_{\max}$ 和 $x_{\min}$ 分别代表数据的最大值和最小值。

2.3 延迟时间和嵌入维的确定

由于害虫发生是一种弱混沌时间序列，因此松毛虫发生面积延迟时间 τ=1，采用 G–P 法[17]选择嵌入维数 m。

1) 设延迟时间的初值为 τ=1，嵌入维数的初值为 m=1。

2) 选择合适的临界距离 r，根据式(11)计算 $C_n(r)$，采用 2 个向量最大分量差作为向量距离。

$$C_n(r) = \frac{1}{M^2}\sum_{i,j=1}^{M}\theta\left[r - \|X(i) - X(j)\|\right] \quad (11)$$

式中：M 为相点的个数；r 为临界距离；θ 为 Heaviside 单位函数。

3) 用最小二乘法拟合 log$C(r)n$–log$r$ 曲线中的直线段。

4) 增加嵌入维数，返回步骤 2)。

2 种时间序列在不同嵌入维数下的关联维数，如图 1 所示。从图 1 可知，当嵌入维数 m=2 时，普洱市思茅区的松毛虫发生面积关联维数达到饱和状态，当嵌入维数 m=5 时，浙江省仙居县的松毛虫发生面积关联维数达到基本饱和状态，表明 2 种时间序列的最佳嵌入维数分别为 m=2 和 m=5。

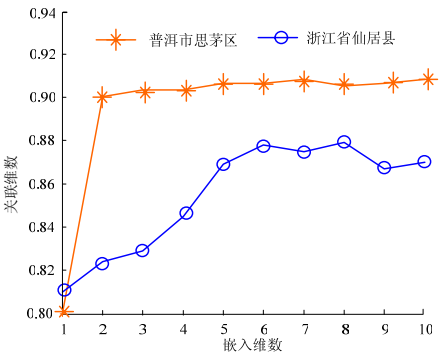

图 1 不同嵌入维数下的关联维数

Fig.1 Correlation dimension under different embedding dimensions

2.4　松毛虫发生量时间序列重构

采用上述延迟时间 τ 和嵌入维数 m 对 2 种松毛虫发生面积时间序列进行重构，并进行归一化处理，结果见表 3 和表 4。

表 3　归一化后重构的训练集(普洱市思茅区)

Table 3　Training set after normalize and reconstruction (Simao, Puer)

x_1	x_2	y
0.098	0.029	0.167
0.029	0.167	0.529
0.167	0.529	0.686
0.529	0.686	0.098
0.686	0.098	0.000
0.098	0.000	0.118
0.000	0.118	0.235
0.118	0.235	0.814
0.235	0.814	0.716
0.814	0.716	0.020
0.716	0.020	0.118
0.020	0.118	0.177
0.118	0.176	0.520
0.176	0.520	1.000
0.520	1.000	0.128
1.000	0.127	0.000
0.127	0.000	0.118

表 4　归一化后重构的训练集(浙江省仙居县)

Table 4　Training set after normalize and reconstruction (Xianju county，Zhejiang)

x_1	x_2	x_3	x_4	y
0.541	0.000	1.000	0.652	0.581
0.000	1.000	0.652	0.581	0.926
1.000	0.652	0.581	0.926	0.514
0.652	0.581	0.926	0.514	0.685
0.581	0.926	0.514	0.685	0.454
0.926	0.514	0.685	0.454	0.267
0.514	0.685	0.454	0.267	0.265
0.685	0.454	0.267	0.265	0.274
0.454	0.267	0.265	0.274	0.289
0.267	0.265	0.274	0.289	0.205
0.265	0.274	0.289	0.205	0.640
0.274	0.289	0.205	0.640	0.471
0.289	0.205	0.640	0.471	0.767
0.205	0.640	0.471	0.767	0.595
0.640	0.471	0.767	0.595	0.536
0.471	0.767	0.595	0.536	0.538
0.767	0.595	0.536	0.538	0.760

3　结果与分析

3.1　PSR–LSSVM 的预测结果

将普洱市思茅区松毛虫发生面积重构后的训练样本输入 LSSVM 学习，粒子群算法找到模型参数为：γ=100，σ=0.675。建立普洱市思茅区松毛虫发生面积预测模型，对测试集进行预测，同样地，将浙江省仙居县松毛虫发生面积重构后的训练样本输入 LSSVM 学习，粒子群算法找到模型参数为：γ=1 000，σ=1.65。建立浙江省仙居县松毛虫发生面积预测模型，对测试集进行预测，结果列于表 5。由表 5 可知，对普洱市思茅区松毛虫发生面积的预测平均相对误差为 0.9%，对浙江省仙居县松毛虫发生面积的预测平均相对误差为 2.4%。

表 5　PSR–LSSVM 的松毛虫发生面积预测结果

Table 5　Prediction results of *Dendrolimus punctatus* occurrence areas based on PSR–LSSVM

预测地点	时间	实测值	预测值	相对误差/%
普洱市思茅区	2004	0.370	0.38	2.7
	2005	0.730	0.73	0.0
	2006	1.000	1.00	0.0
浙江省仙居县	1989–04	4 177.800	4 134.40	1.0
	1989–07	2 932.300	2 980.60	1.6
	1989–09	2 041.300	2 135.80	4.6

综上所述，对混沌理论与非线性预测算法 LSSVM 进行融合建立的预测模型，可以对松毛虫发生面积进行准确预测，主要是由于一维松毛虫发生面积时间序列中包含着吸引子的结构，而重构相空间能够估计出系统演化的信息，使预测结果更符合松毛虫发生面积的实际情况，同时利用非线性预测能力强的 LSSVM 对重构后的数据进行建模，并采用粒子群优化算法优化 LSSVM 参数，进一步提高了松毛虫发生面积的精度，同时该模型要求的样本不大，克服了一般混沌预测模型要求大样本的缺陷，通用性较好，适合于小样本的害虫预测。

3.2　与其他模型预测结果对比

为了使 PSR–LSSVM 的预测结果更具说服力，采用线性预测模型(ARIMA)、PSR–BPNN 神经网络模型(PSR–BPNN)进行对比试验。模型性能评价指标为平均绝对百分误差(*MAPE*)和均方根误差(*RMSE*)。表 6 的结果表明：

1) BP 神经网络和 LSSVM 预测比线性模型

ARIMA 预测精度有大幅度提高。由于松毛虫发生具有非线性特征，BP 神经网络、LSSVM 能够更好地反映松毛虫发生量的变化趋势，有效提高了松毛虫发生量的预测精度。

2) 相对于 PSR–BP 神经网络(PSR–BPNN)，LSSVM 的预测精度更高，说明 LSSVM 能较好地克服 BP 神经网络存在的过拟合、易陷入局部最优等缺陷，对于小样本害虫时间序列预测，优势更加明显。

3) 将混沌理论和 LSSVM 用于害虫发生量预测，误差较小，结果接近真实值，说明 PSR–LSSVM 能够更加精确反映害虫发生变化趋势和变化特点，预测结果更加可靠。

表 6 各模型的预测误差对比

Table 6 Comparison for prediction error of different models

预测模型	*RMSE*		*MAPE*/%	
	普洱市思茅区	浙江省仙居县	普洱市思茅区	浙江省仙居县
ARIMA	0.073	117.23	8.33	3.97
PSR–BPNN	0.041	109.07	3.48	3.22
PSR–LSSVM	0.006	66.20	0.90	2.44

参考文献：

[1] 张孝羲．昆虫生态及预测预报[M]．3 版．北京：中国农业出版社，2002．

[2] 许晓风，马飞，丁宗泽，等．褐飞虱发生的相空间线性回归预测模型[J]．昆虫学报，2002，45(4)：548–551．

[3] 岑冠军，黄寿山，肖莉，等．ARIMA 模型在小菜蛾幼虫种群动态中的应用[J]．华南农业大学学报，2008，29(1)：109–114．

[4] 张真，李典谟，查光济．马尾松毛虫种群动态的时间序列分析及复杂性动态研究[J]．生态学报，2002，22(7)：1061–1067．

[5] 陈顺立，张华峰，张潮巨，等．神经网络在松墨天牛发生量预报中的应用[J]．福建林学院学报，2006，26(2)：6–9．

[6] 石晶晶，刘占宇，张莉丽，等．基于支持向量机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．

[7] 向昌盛，周子英，张林峰．支持向量机在害虫发生量预测中的应用[J]．生物信息学，2011，9(1)：28–31．

[8] 向昌盛，周子英，武丽娜．基于 ARIMA 和 DSVM 组合模型的松毛虫发生面积预测[J]．湖南农业大学学报：自然科学版，2010，36(4)：430–434．

[9] 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)：433–436．

[10] 向昌盛，袁哲明．最小二乘支持向量机在害虫预测中的应用[J]．湖南科技大学学报：自然科学版，2012，27(2)：71–74．

[11] 向昌盛，周子英．ARIMA 与 SVM 组合模型在害虫预测中的应用[J]．昆虫学报，2010，53(9)：1055–1060．

[12] 朱军生，翟保平，刘英智．基于小波分解的害虫发生非平稳时间序列分析和预测[J]．南京农业大学学报，2011，34(3)：61–66．

[13] 马飞，许晓风，张夕林，等．相空间重构与神经网络融合预测模型及其在害虫测报中的应用[J]．2002，22(8)：1297–1301．

[14] Packard N H，Crutchifield J P，Farmer J D，et al. Geometry from a time series[J]．Phys Rev Lett，1980，45(6)：712–716．

[15] Takens F．Determing strange attractors in turbulence [J]．Lecture Notes in Mathematics，1988，898：361–381．

[16] An S J，Liu W Q，Venkatesh S．Fast cross validation algorithms for least squares support vector machines and kernel ridge regression[J]．Pattern Recognition，2007，40(2)：2154–2162．

[17] Grassberger P，Privacies I．Characterization of strange attractors[J]．Physical Review Letters，1983，50(5)：346 – 349．

责任编辑：罗慧敏

英文编辑：罗 维