# 基于大数据的玉米田四代棉铃虫发生量的预测模型

**赵雷，杨波，刘勇，牟少敏，温孚江**
山东农业大学农业大数据研究中心，山东 泰安 271018

**摘要**
提出了一种基于支持向量机的预测模型。根据山东省1999 - 2013年玉米田第四代棉铃虫发生程度采集的数据，采用支持向量回归（SVR）算法，构建了玉米田第四代棉铃虫发生程度与其关联因子间的非线性关系模型，并对该方法进行了测试与分析。结果表明，由SVR预测模型得到的预测发生量与实际发生量基本一致，预测的平均绝对百分比误差为4.36%，预测值与实际值的相关系数为0.960 6，为玉米田第四代棉铃虫的有效防控提供了科学指导。

**关键词**
农业大数据；棉铃虫；支持向量回归；监测预警；玉米

中图分类号：S431.9 文献标识码：A
doi: 10.11959/j.issn.2096-0271.2016008

# *Forecasting model for the fourth generation of cotton bollworm in corn fields based on big data*

**ZHAO Lei, YANG Bo, LIU Yong, MU Shaomin, WEN Fujiang**
Agricultural Big Data Research Center, Shandong Agricultural University, Taian 271018, China

***Abstract***
The monitoring and forecasting model was put forward based on support vector machine program. According to the data collection of the fourth generation occurrence degree of the corn bollworm in Shandong province from 1999 to 2013, the support vector regression (SVR) method was adopted to build the nonlinear correlation model between the occurrence degree of the fourth generation bollworm and the associated factors. The method and the model were tested and analyzed. The results showed that the SVR forecasting model for prediction was almost in accord with the actual insect occurrence situation. The mean absolute percentage error was 4.36%, and the actual and estimated value of the correlation coefficient was 0.960 6. It could provide effective and accurate guidance to the cotton bollworm control in corn fields.

***Key words***
agricultural big data, cotton bollworm, support vector regression, monitoring and forecasting, corn

# 1 引言

随着云计算、物联网等技术的迅猛发展，数据正以前所未有的速度不断地增长和累积，大数据时代已经来临。依靠数据分析获得洞察力，做出更好的业务决策是数据分析挖掘的主要目的[1,2]。

预测是大数据的核心。跳出传统的因果关系的思维局限，通过对大量数据的搜集、挖掘和分析，发现数据间隐藏的相关关系，从而揭示事物发生和发展的内在规律，能做出更快、更符合实际的预测[3,4]。如英国和美国科学家在《Science》发文指出，全球变暖会导致非洲和南美洲高原地带疟疾的流行[5]；基于多年的数据挖掘和分析，荷兰科学家指明，新烟碱类杀虫剂吡虫啉的应用是本地食虫鸟类种群数量减少的主要原因[6]。因此，借助于大数据的研究手段和方法，能够使已有的农业数据“活起来”，认识其内在的关联性，预测发展趋势，使其在指导生产活动中产生价值[7-10]。

棉铃虫属鳞翅目、夜蛾科，是一种重要的农业害虫，寄主范围广。近年来，随Bt棉的大范围种植，第四代棉铃虫对玉米叶片和果穗，特别是果穗的危害逐渐加重，造成玉米产量下降，品质降低[11]。截至目前，国内外对棉铃虫的监测预警研究主要是依靠有限的气象因素，如温度、降雨和光周期等[12-14]，采用线性回归分析，建立相关模型，开展对棉花田棉铃虫的预测研究。而对玉米棉铃虫的监测预警未见报道。本文基于大数据理念，依据已有的数据积累，采用支持向量回归（SVR）算法，建立了预测玉米田第四代棉铃虫发生量的支持向量机模型，为指导玉米田第四代棉铃虫的发生预测及科学防治打下了基础。

# 2 资料与方法

## 2.1 数据来源

本文涉及的变量主要有1999-2013年山东省滨州地区玉米田第四代棉铃虫的发生量及气象数据，主要包括7月中旬到8月上旬的平均气温、降水量、最高气温（≥35℃）的日数、降水（≥10 mm）的日数、极大风速、平均本站气压、平均风速、平均水汽压、平均相对湿度、日最低本站气压、日最低气温、日最高本站气压、日最高气温、最大风速、最大风速的风向和最小相对湿度，分别计算出每年月气象因子的平均数。1999-2013年第四代棉铃虫的发生量资料来自山东省滨州地区植保部门；该时期的逐日气象观测资料来自国家气象信息中心。

## 2.2 支持向量机

### 2.2.1 支持向量机的基本思想

支持向量机（support vector machines，SVM）是Vapnik等人根据统计学习理论提出的机器学习方法[15]。基本思想是通过一个非线性映射把样本空间映射到一个高维特征空间中，将寻找最优线性回归超平面的算法归结为求解一个凸约束特性下的凸规划问题，并得到全局最优解。同时支持向量机通过定义核函数（kernel function），将高维空间中的内积运算转化为原空间中的核函数运算（如图1所示）。由于棉铃虫的发生具有非线性、不稳定、多变量的特点，对于其虫害发生系统，很难用确切的公式和解析方法将棉铃虫发生的规律表达出来[16]。而这种信息处理方式正是支持向量机所具备的，因

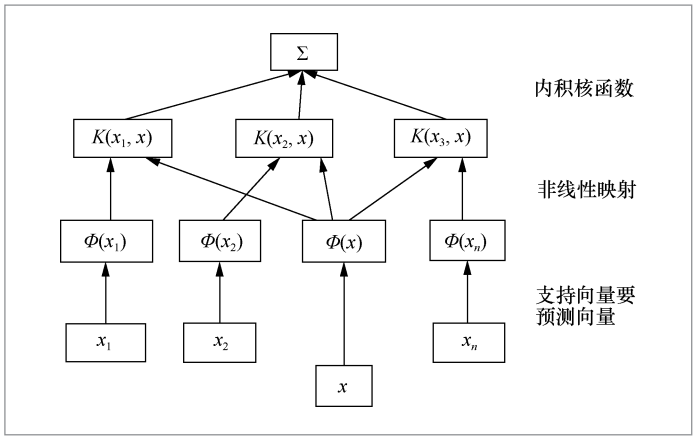



图1 支持向量机结构示意

此，将支持向量回归用于棉铃虫发生量的建模和预测分析。

### 2.2.2 支持向量回归预测方法

支持向量机的回归[17]函数拟合分为线性拟合和非线性拟合，首先考虑线性拟合函数。假设有一个样本集为：$(y_1,x_1)$，$(y_2,x_2)$，…，$(y_i,x_i)$，$y\in R$，回归函数线性方程表示如下：

$$f(x)=W^{\mathrm{T}}x+b \tag{1}$$

跟基础定义的线性可分原理一样，通过函数的最小值找到最佳的回归函数，得出：

$$\min\frac{1}{2}w^{\mathrm{T}}w+c\sum_{i=1}^{N}(\xi_i+\xi_i^*) \tag{2}$$

其中，$w$代表$W$的维数，$\xi$和$\xi^*$为松弛变量，$\xi$为上限，$\xi^*$为下限。Vapnik定义不敏感耗损函数为：

$$L(y)=\begin{cases}0, & |f(x)-y\leqslant\varepsilon|\\ |f(x)-y|-\varepsilon, & |f(x)-y>\varepsilon|\end{cases} \tag{3}$$

得到对偶优化方程为：

$$\begin{aligned}&\max_{a,a^*}w(a,a^*)\\&=\max_{a,a^*}\left\{-\frac{1}{2}\sum_{i=1}^{i}\sum_{j=1}^{i}(a_i-a_i^*)(a_j-a_j^*)(X_i-X_j)+\right.\\&\left.\sum_{i=1}^{i}a_i(y_i-\varepsilon)-a_i(y_i-\varepsilon)-a_i^*(y_i+\varepsilon)\right\}\end{aligned} \tag{4}$$

它的约束条件是：$0\leqslant a\leqslant c,i=1,\cdots,N$；$0\leqslant a^*\leqslant c,i=1,\cdots,N$；$\sum_{i=1}^{K}(ai-a_i^w)=0$。

于是得到支持向量机的回归函数：

$$f(x)=\sum_{i=1}^{k}(a_i-a_i^*)(x\cdot x_i) \tag{5}$$

其中，$a_i$、$a_i^*$将只有小部分不为0，它们对应的样本就是支持向量。

对于非线性的支持向量机的回归，通过一个非线性映射把数据$x$映射到高维特征空间，然后可以在这个空间进行线性回归，也就是类似于分类问题。跟支持向量机定义的线性不可分的原理类似，在它的基础上进行回归，需用到一个非敏感性损耗函数，且目标函数为：

$$\begin{aligned}&\max_{a,a^*}w(a,a^*)\\&=\max_{a,a^*}\left\{\sum_{i=1}^{i}a_i^*(y_i-\varepsilon)-a_i(y_i+\varepsilon)-\right.\\&\left.\frac{1}{2}\sum_{i=1}^{i}\sum_{j=1}^{j}(a_i^*-a_i)(a_j^*-a_j)K(X_i-X_j)\right\}\end{aligned} \tag{6}$$

其约束条件是跟式(4)的约束条件一样。可通过求得的$a_i$和$a_i^*$，得到回归函数：

$$f(x)=\sum_{i=a}^{i}(a_i-a_i^*)k(x_i,x)+b \tag{7}$$

### 2.2.3 核函数

核函数在支持向量回归中起着重要的作用，它不仅可以解决非线性问题，克服维数灾难问题，而且还可以代替高维特征空间中的内积运算，避免高维度运算的复杂性。

支持向量机核函数的回归预测在诸如疾病预测、天气预测、市场预测、股价预测等很多方面都得到了广泛应用。核函数主要有4种。本文将径向基核函数（RBF）应用于支持向量回归模型中，支持向量回归对玉米田第四代棉铃虫的发生量预测从线性到非线性转换，是通过径向基核函数。径向基核函数能够针对棉铃虫的发生具有非线性、不稳定性和多变量的特点进行处理。

径向基核函数为：

$$K(x_i,x_j)=\exp\left\{-\frac{|x_i-x_j|^2}{\sigma^2}\right\} \quad (8)$$

# 3 棉铃虫发生量模型实例研究

## 3.1 数据的预处理

本文是以滨州地区玉米地1999-2010年玉米田第四代棉铃虫的实际发生量作为训练样本（见表1），基于支持向量机理论，建立支持向量回归第四代棉铃虫发生量的预测模型，以2011-2013年第四代棉铃虫发生量数据进行预测（见表2）。为了提高数据之间的可比性和收敛速度、缩短训练时间，本文在对数据处理时，先对原始数据进行归一化处理。

$$x_i'=\frac{x_i-x_{\min}}{x_{\max}-x_{\min}} \quad (9)$$

其中，$x_i$是原始数据；$x_i'$为归一化后的数据；$x_{\min}$和$x_{\max}$分别代表数据的最小值和最大值。将数据压缩到[0,1]。

## 3.2 模型参数的选取

本研究中，采用LIBSVM-3.20软件包来建立SVR模型。通过对各种核函数进行测试，最终确定预测模型的核函数为径向基核函数。模型的参数按照相应的标准来选取，本文选定的滑动窗口宽度（嵌入维数）为16，用gridregression.py自动搜索最佳惩罚参数、灵敏度及径向宽度等核函数参数。

## 3.3 基于SVR的第四代棉铃虫发生量预测模型

根据式（10）得到实际值与拟合值之间的相关系数[18]，得到的数据表明实际值和拟合值之间有高度的相关性，相关系数接近1，拟合效果好。数据表明实际值与预测值之间的相关系数为0.96，有高度的相

表1 支持向量回归拟合结果与实际值对比

| 年份 | 真实值 | 拟合值 | 绝对误差 | 相对误差 |
|---|---|---|---|---|
| 1999年 | 15.5 | 15.599 8 | - 0.099 8 | 0.64% |
| 2000年 | 49 | 48.899 8 | 0.100 2 | 0.20% |
| 2001年 | 14 | 14.099 8 | - 0.099 8 | 0.71% |
| 2002年 | 24.5 | 24.399 9 | 0.100 1 | 0.41% |
| 2003年 | 5 | 5.100 1 | - 0.100 1 | 2.00% |
| 2004年 | 19 | 18.574 9 | 0.425 1 | 2.24% |
| 2005年 | 16 | 16.100 3 | - 0.100 3 | 0.63% |
| 2006年 | 39.3 | 37.699 3 | 1.600 7 | 4.07% |
| 2007年 | 42.5 | 42.399 8 | 0.100 2 | 0.24% |
| 2008年 | 47 | 47.100 3 | 0.100 3 | 0.21% |
| 2009年 | 34.5 | 34.600 3 | - 0.100 3 | 0.29% |
| 2010年 | 38 | 38.099 9 | - 0.1 | 0.26% |

表2 支持向量回归预测结果与实际值对比

| 年份 | 真实值 | 预测值 | 绝对误差 | 相对误差 |
|---|---|---|---|---|
| 2011年 | 52 | 52.992 3 | - 0.992 3 | 1.90% |
| 2012年 | 42.6 | 41.243 2 | 1.356 8 | 3.17% |
| 2013年 | 32.5 | 35.101 7 | - 2.601 8 | 8.01% |

关性，预测结果与实际值相符合。

通过回归模型得到的训练集样本的拟合值与实际值相符合（如图2所示），而测试集样本的预测值与实际值相匹配（如图3所示）。

$$R=\frac{n\sum_{i=1}^{n}y_i\cdot\hat{y}_i-\left(\sum_{i=1}^{n}y_i\right)\left(\sum_{i=1}^{n}\hat{y}_i\right)}{\sqrt{\sum_{i=1}^{n}y_i^2-\left(\sum y_i\right)^2}\sqrt{\sum_{i=1}^{n}\hat{y}_i^2-\left(\sum_{i=1}^{n}\hat{y}_i\right)^2}} \quad (10)$$

其中，$y_i$为样本的实际值，$\hat{y}_i$为样本的预测值，$n$为预测样本数。

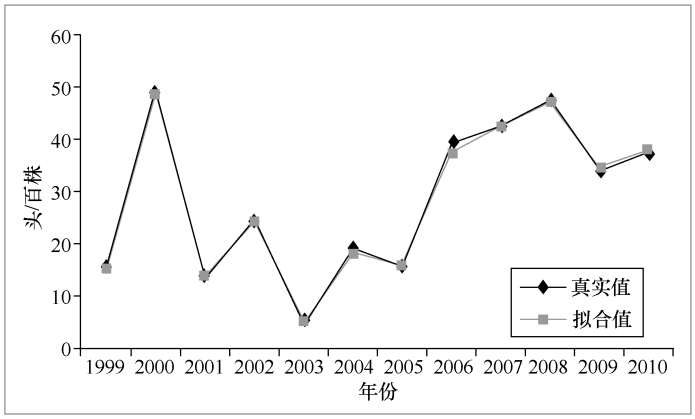



图 2 训练集拟合结果

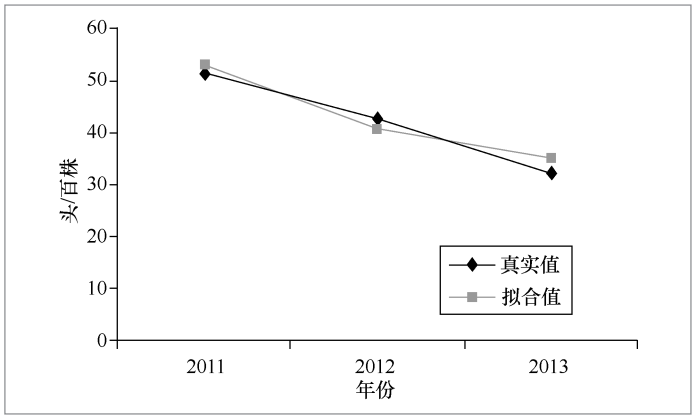



图 3 样本的预测结果

表 3 预测样本的 MAE、MRE、RMSE 值

| | MAE | MRE | RMSE |
|---|---|---|---|
| 训练值 | 0.24 | 0.99% | 1.68 |
| 预测值 | 1.65 | 4.36% | 3.1 |

支持向量机的预测准确率用偏差来表示，主要包括平均绝对误差（MAE）、平均相对误差（MRE）、均方根误差（RMSE），表达式如下：

$$MAE=\frac{1}{2}\sum_{i=1}^{n}|y_i-\hat{y}_i| \quad (11)$$

$$MRE=\frac{1}{n}\sum_{i=1}^{n}\frac{|y_i-\hat{y}_i|}{y_i} \quad (12)$$

$$RMSE=\sqrt{\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n}} \quad (13)$$

其中，$y_i$为样本的实际值，$\hat{y}_i$为样本的预测值，$n$为预测样本数。

用式（11）~式（13）来计算偏差，计算结果见表3。结果表明，拟合误差水平相当低，预测误差水平高于拟合误差。总体而言，误差水平总体较低，*MAE*低于2，*MRE*低于5%，*RMSE*低于4。而由表2可知，除了2013年的相对误差达到8.01%之外，其他样本的相对误差均较小。由此可知，基于SVR的玉米田第四代棉铃虫发生量预测模型具有可行性。

## 4 应用分析与讨论

玉米田棉铃虫的发生具有复杂的非线性变化规律，要建立精确的数学模型相当困难，因此，根据大数据的理念，将支持向量机引入第四代棉铃虫发生量的预测中。支持向量机不需要对数据分布性质做任何假设，通用性较强，实验结果表明，将SVR用于玉米田第四代棉铃虫发生量建模与预测，能较好地揭示玉米田第四代棉铃虫的发生规律。传统的模型主要集中在对棉田棉铃虫的发生预测上。传统的方法主要有多元线性回归分析[12]、二级分辨率模型、Fisher模型、期距法等数十种。近年来，随着物联网、机器学习等技术的迅速发展，

出现了更为方便、准确率更高的预测方法。如黄健等人[14]利用人工神经网络（ANN）模型对新疆地区1990-2007年第二代棉铃虫的发生等级进行预测的结果表明，BP神经网络预报模型的拟合精度和预报精度高于逐步回归模型。朱军生等人[19]利用1966-1995年山东惠民县棉铃虫的监测数据建立了基于径向基小波神经网络的第二代棉铃虫卵峰日期预测模型，结果表明，在5年的预测数据中，4年的预测数据偏差在3天以内，另外一年的预测数据偏差为4天。这对于卵峰日的预测偏差较大。以前对棉铃虫的预测，大都是对棉田第二、三代棉铃虫发生期、发生等级的预测，而未见对玉米田主害代棉铃虫进行预测。虽然基于人工神经网络的模型比传统的预测模型效果好，但是人工神经网络容易出现过拟合、维数灾难等问题；径向基小波神经网络模型结构复杂。而支持向量机模型可以很好地解决小样本、非线性、过拟合、维数灾难和局部最优的问题，结构简单，便于应用。因此，首次将SVR应用到玉米田第四代棉铃虫发生量的预测上。

从理论上讲，SVR算法得到全局最优，可解决其他神经网络算法无法避免的局部最优问题；SVR预测是把线性回归转为非线性，需要将内机核函数转换为高维空间的非线性映射，计算的复杂性取决于支持向量机的数目，不是样本空间的维数，因此，从一定层面上避免了维数灾难的问题。

但是，SVR算法是半监督式学习算法，其模型具有一定的局限性，由表1可知，2011-2013年的预测值的相对误差逐渐增大，与实际值相差逐渐偏大，说明预测的年份距建立模型的年份越远，预测结果偏差越大。这需要连续跟踪现实数据的采集，不断优化预测模型。

气象因子是影响玉米第四代棉铃虫种群发生和发展的重要因子。本文通过对气象因子进行分析，建立基于SVR的玉米第四代棉铃虫的预测模型，得到较为准确的结果。但在2006年，预测相对误差达到4.07%，推测原因主要有以下两点。

- 历史数据的限制。过去十几年采集的数据量小、数据范围比较窄。2014年以后，扩展了数据采集面，将农田中诸如生物因子（包括自然天敌、周边生物环境等）、非生物因子（包括土壤性质等以及地块类型、施肥、浇水等）都列于采集范围之内。随着数据采集的不断完善以及数据分析方法的逐步改进，将会使预测结果与实际情况更加接近。
- SVR本身的不足，如参数优化等。

## 5 结束语

本研究首次将SVR应用到玉米田第四代棉铃虫发生量的预测上，根据1999-2010年第四代棉铃虫采集的数据构建了玉米田第四代棉铃虫发生量的SVR模型，并对2011-2013年进行了测试，得到的预测发生量与实际发生量基本一致，呈现高度相关性。这一模型的应用，能及时和准确地发布第四代棉铃虫监测预警信息，有效地指导玉米田棉铃虫的科学防控。

本文首次将SVR用于第四代棉铃虫发生量建模与预测中，回归精度与泛化能力都较高。以上研究表明，SVR应用于玉米田棉铃虫发生量预测是可行的。

## 参考文献:

[1] GANDOMI A, HAIDER M. Beyond the hype: big data concepts, methods, and analytics[J]. International Journal of Information Management, 2015, 35(2): 137-144.

[2] KAMBATLA K, KOLLIAS G, KUMAR V, et al. Trends in big data analytics[J]. Journal of

Parallel and Distributed Computing, 2014, 74(7): 2561-2573.

[3] LAZER D, KENNEDY R, KING G, et al. The parable of Google flu: traps in big data analysis[J]. Science, 2014, 343(6176): 1203-1205.

[4] SUN Z F, DU K M, ZHENG F X. Perspectives of research and application of big data on smart agriculture[J]. Journal of Agricultural Science and Technology, 2013, 15(6): 63-71.

[5] SIRAJ A S, SANTOS-VEGA M, BOUMA M J, et al. Altitudinal changes in malaria incidence in highlands of ethiopia and colombia[J]. Science, 2014, 343(6175): 1154-1158.

[6] HALLMANN C A, FOPPEN R P B, TURNHOUT C A M V, et al. Declines in insectivorous birds are associated with high neonicotinoid concentrations[J]. Nature, 2014, 511(7509): 341-343.

[7] 宋长青, 牟少敏, 周虎, 等. 农业大数据研究中心的建设、研究与实践[J]. 中国现代教育装备, 2015(3): 8-11.
SONG C Q, MU S M, ZHOU H, et al. Construction, research and practice of agricultural big data research center[J]. China Modern Educational Equipment, 2015(3): 8-11.

[8] 辛妍. 大数据在农业中的应用[J]. 新经济导刊, 2015(4): 46-51.
XIN Y. Big data applications in agriculture[J]. New Economy Weekly, 2015(4): 46-51.

[9] 温孚江. 农业大数据与发展新机遇[J]. 中国农村科技, 2013(3): 4-7.
WEN F J. Agricultural big data and development of new opportunities[J]. Agricultural Sciences in China, 2013(3): 4-7.

[10] 杨波, 刘勇, 牟少敏, 等. 大数据背景下山东省二代玉米螟发生程度预测模型的构建[J]. 计算机研究与发展, 2014(S2): 160-165.
YANG B, LIU Y, MU S M, et al. Based on big data: the establishment of meteorological forecast model for the occurrence degree of the second generation of corn borer in Shandong[J]. Journal of Computer Research and Development, 2014(S2): 160-165.

[11] 陈广泉. 河西走廊玉米田棉铃虫发生规律与药剂防治技术研究[D]. 咸阳: 西北农林科技大学, 2004: 12-16.
CHEN G Q. The occurrence regularity and cottrol technology research of cotton bollworm in the hexi corridor cornfield[D]. Xianyang: Northwest A&F University, 2004: 12-16.

[12] CHANGNON D, SANDSTROM M, ASTOLFI J, et al. Using climatology to predict the first majorsummer corn earworm (lepidoptera: noctuidae) catch in north central Illinois[J]. Meteorological Applications, 2010, 17(3): 321-328.

[13] FENG H, GOULD F, HUANG Y, et al. Modeling the population dynamics of cotton bollworm Helicoverpa armigera (Hübner) (lepidoptera: noctuidae) over a wide area in northern China[J]. Ecological Modelling, 2010, 221(15): 1819-1830.

[14] RAJ K R, KARDAM A, ARORA J K, et al. Artificial neural network (ANN) design for Hg-Se interactions and their effect on reduction of Hg uptake by radish plant[J]. Journal of Radioanalytical and Nuclear Chemistry, 2010, 283(3): 797-801.

[15] VAPNIK V N. The nature of statistical learning theory[J]. IEEE Transactions on Neural Networks, 1995, 10(5): 988-999.

[16] 赵仲华, 沈佐锐. 昆虫种群动态非线性建模理论与应用[J]. 生物数学学报, 2001, 16(4): 439-447.
ZHAO Z H, SHEN Z R. Insect population dynamics of nonlinear modeling theory and application[J]. Journal of Biomathematics, 2001, 16(4): 439-447.

[17] 李永娜. 基于支持向量机的回归预测综述[J]. 信息通信, 2014(11):32-33.
LI Y N. Regression forecast review based on support vector machine[J]. Message Communication, 2014(11): 32-33.

[18] ERDAL H I. Two-level and hybrid ensembles of decision trees for high performance concrete compressive strength prediction[J]. Engineering Applications of Artificial Intelligence, 2013, 26(7): 1689-1697.

[19] 朱军生, 翟保平, 董保信. 基于径向基小波网络的二代棉铃虫卵峰日预测模型[J]. 昆虫学报, 2010, 53(12): 1429-1435.
ZHU J S, ZHAI B P, DONG B X. Forecasting model for the oviposition peak day in the

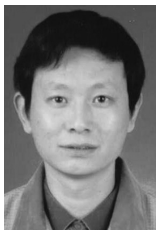
second generation of helicoverpa armigera (lepidopeter: noctuidae) based on radial basis wavelet network[J]. Acta Entomologica Sinica, 2010, 53(12): 1429-1435.

作者简介

**赵雷**(1990-),女,山东农业大学硕士生,主要研究方向为农业科研与大数据。

**杨波**(1988-),女,山东农业大学硕士生,主要研究方向为农业大数据。

**刘勇**(1968-),男,山东农业大学教授、博士生导师,主要研究方向为害虫绿色防控和农业大数据。



**牟少敏**(1964-),男,博士,山东农业大学教授,主要研究方向为大数据、机器学习和模式识别。

**温孚江**(1955-),男,现任山东农业大学校长、教授,农业大数据创新战略联盟理事长,全国人民代表大会常务委员会委员。早年留学美国,并获博士学位。主要从事植物保护研究和宏观农业研究工作。发表论文210余篇,专著5部。最近一部专著《大数据农业》由中国农业出版社于2015年9月出版。目前主要从事农业大数据应用研究工作,是我国农业大数据研究主要发起人之一。

**收稿日期**:2015-10-28

**通信作者**:刘勇,liuyong@sdau.edu.cn;温孚江,fjw@sdau.edu.cn

**基金项目**:山东省农业重大应用技术创新课题基金资助项目

**Foundation Item:** Major Innovation of Applied Technology in Agriculture of Shandong Province