DOI: 10.13718/j.cnki.xdzk.2020.01.019

基于机器学习和图像处理技术的病虫害预测

杭　立,　车　进,　宋培源,　王晨宇,　田　斌

宁夏大学 物理电子电气工程学院,银川 750021

摘要: 针对传统病虫害预测过于繁琐、准确度低的现状,提出一种基于图像处理与 SVM(支持向量机)结合的病虫害预测算法,并对未来几年病虫害可能发生的面积进行了有效预测. 首先通过图像滤波、特征提取等图像处理技术得到昆虫数学形态学特征,然后结合往年病虫害数据对特征进行标签设定和科学分类,继而对未来病虫害的发生进行合理预测. 通过构建动态预测模型进行有效的、科学的病虫害预测预报. 最后,通过与实际值进行对比,预测精度达到了 90%. 实验结果表明,该方法具备较好的预测精度,是一种合理科学的预测方法.

关　键　词: 病虫害预测; 图像处理; 形态学特征; 机器学习; SVM

中图分类号: TP181　　**文献标志码**: A　　**文章编号**: 1673-9868(2020)01-0134-08

作为人口大国,农业是关乎我国社会稳定的重要基础. 近年来,我国在预防农业病虫害领域投入巨大,但是农业灾害,尤其是作物病虫害依旧屡见不鲜. 频繁的虫害不但造成了难以挽回的经济损失[1],而且还导致了生态效益与社会效益下降. 如果能够有效地对害虫发生趋势进行预测,就可以提前掌握虫害的发生动态,预防和采取相应的防治措施,更有效地控制灾害,进而减少虫害造成的损失. 因此,能够准确地对虫害发生趋势进行预测预报,具有非常重要的意义.

现阶段,对害虫发生趋势预测的传统方法主要有期距预测、有效积温预测、多元线性回归预测及有效基数预测等[2]. 王淑芬等[3]研制了马尾松毛虫防治专家系统; 王霓虹等人[4]基于 Web GIS 技术,研发了一款森林病虫害预测的专家系统. 这类预测系统本质上都基于传统的专家系统. 但是,我国幅员辽阔,气候地貌复杂,虫害的发生概率大小受到气候、天敌等众多因素的协同影响,不仅仅是简单的线性关系,而是一种复杂的非线性关系[5]. 同时,预测率较高的专家系统需要依赖于农业专家的人工预测,由于人工成本较高,普及率较低.

因此,寻求一种合理且更为准确的预测方法就显得尤为重要. 本文提出一种以图像处理技术结合 SVM 机器学习的方法,通过这种方法,在不影响准确度的条件下,可以有效减少预测时的自变量数量,对虫害进行合理预测.

1　高斯滤波

作为一种线性平滑去噪算法,采用高斯滤波算法,有算法成型快、容易实现的优点. 由于绝大多数噪

收稿日期: 2018-04-22
基金项目: 国家自然科学基金项目(61861037); 宁夏回族自治区重点研发计划项目(NZ1512).
作者简介: 杭　立(1994—),男,硕士研究生,主要从事图像处理与模式识别研究.
通信作者: 车　进,博士,教授.

声可以近似为高斯分布的白噪声[6]，因此本文采用低通高斯滤波器，过滤掉采集图像中的高斯白噪声，进而实现图像平滑.

根据选择的固定窗口大小及窗口内任意像素与中心像素点的距离，利用高斯函数实现系数权值的分配，即高斯滤波可以表示为

$$I_{Rf}(x, y)=\frac{\sum_{(i, j)\in W_{x,y}}\omega_d(i, j)I(i, j)}{\sum_{(i, j)\in W_{x,y}}\omega_d(i, j)} \tag{1}$$

其中：$W_{x,y}$ 表示中心像素(x, y)的 $M\times M$(M 为奇数)大小的领域；ω_d 为空间距离相似度权重因子，即图像中任意 2 个像素之间的空间距离比例，距离越小，权重越大. 根据图 1，各个昆虫距离明显分布不均匀，若单纯为了计算权重，则误差较大. 为了方便程序实现和减小误差，进行单位化，区间为[0，1]，图 1 对应的 ω_d 为 0.78.

由公式(1)可知，在低频区间，高斯滤波算法有良好的处理效果，但其仅仅考虑了像素间的边界关系，而忽略了整体图像区域可能存在拟合特征[7]，由于采用固定的掩模窗口，对于该图像区域进行求和取平均值实现归一化，可能导致丢失平滑区域中的细节信息，所以本文通过高斯滤波，对图像区域内的不同领域像素设置了不同权值，在保证图像可以平滑的同时，能够更多地保留图像总体灰度分布特征. 同时，本文采用的是提取数学形态特征，对于纹理、触角等传统特征没有进行提取，细节信息丢失对于特征的提取影响不大.

图 1 左边部分是采集的原始图像，右边部分是进行高斯滤波后的原始图像(图片采集：诱虫灯大田拍照)，去除了原始图像中的高斯白噪声.

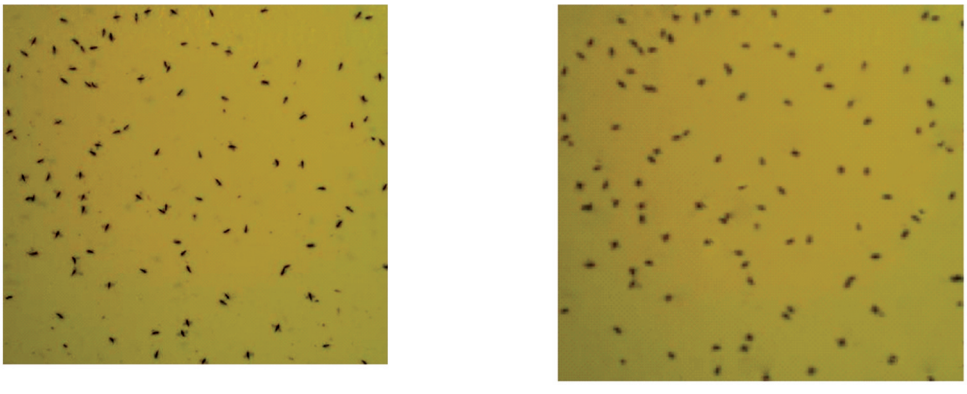


图 1　高斯滤波前后对比图

2　特征处理

2.1　特征提取

将图像中任何一点(x, y)处的积分图像表示为它左上角的所有像素的总和：

$$ii(x, y)=\sum_{x'\leqslant x,\ y'\leqslant y}i(x', y') \tag{2}$$

其中，$ii(x, y)$表示积分图像也称 Harr 特征；$i(x, y)$表示原始图像；如图 2 所示，点(x, y)处的积分图像为灰色区域的像素和[8].

事实上，大多数昆虫在农田里都是运动状态，或可观察的视角角度并非水平或垂直，拍摄过程中难以保证昆虫以固定角度出现，所以本文特征的提取来源，主要是水平或侧面角度随机进行拍摄的图片. 以蚜

虫为例见图3，不同角度拍摄的Harr特征.

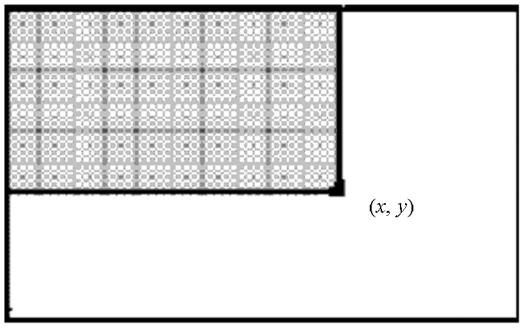
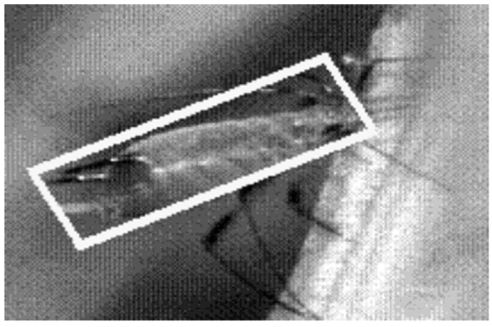


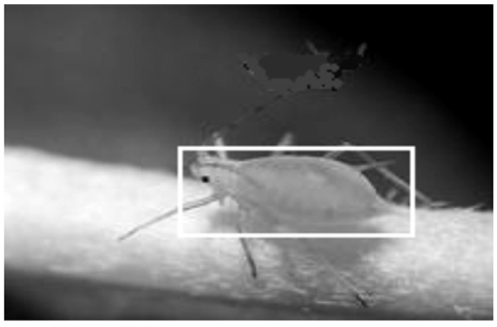
图2 (x, y)的积分表示

2.2 特征分类

昆虫的特征很多，比如颜色、斑纹、大小体长、体宽等[9]. 识别精度与选取的特征值有着密切的关系，但是并非所有特征值都对识别精度有促进作用. 事实上，存在部分特征值对于识别精度有降低的效果，因此我们需要淘汰一些不适用的特征，进行特征值筛选过滤. 尝试在特征向量中去除一个特征值$x_{i,j}$，将剩下的特征向量构成样本采集合用于Jackknife检测[10]，得到A_j和B_j. 通过检测每个Harr特征的互补特征向量，得到一组对于整体特征向量集合的影响特征向量A_0和B_0.



(a) 侧面积分图



(b) 背面积分图

图3 不同角度的积分图表示

图4中有部分特征值未识别，原因是该类型昆虫体积过小，也符合实际的工程情况.

SVM的特点是适合两元分类，对于小部分比较浅或者不规则的成虫数据遗失影响不大，很适合做小样本监督学习.

其中，图4中图片总像素$S=6\ 250$，通过各个昆虫大小累加去和值$K=\sum_{i=1}^{n}k_i$，得到图4中昆虫图像部分像素的累计面积.

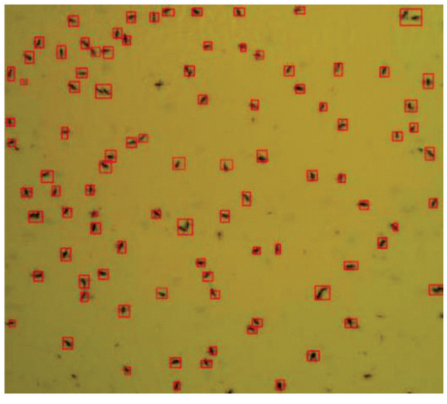


图4 对于诱虫灯图的特征提取

虫口密度δ为

$$\delta=\varphi\frac{K}{S}=\varphi\frac{\sum_{i=1}^{n}k_i}{S} \tag{3}$$

其中，S为图4图像部分的像素总值，K为图4图像中的昆虫累计面积，φ为比例系数，比例系数根据季节随之改变. 这里实验采用的是0.25.

3 基于支持向量机方法的建模与预测

3.1 支持向量机简介

支持向量机(Support Vector Machine，SVM)由训练和核函数组成，基于有限的样本信息，对样本数

据进行分类，在线性可分二元分类中，所有的数据点都在二维平面上，所以此时分割超平面就只是一条直线. 但是，如果给定的数据集是三维的，此时用来分割数据的就是一个平面，该平面为最优超平面.

支持向量机就是离分割平面最近的那些点，是支持向量积的基本原理[11-14]. 其中，最大化支持向量到分割面的距离，就是支持向量机的目标. 最优超平面可以提高模型的预测能力和减少错误分类. 图 5 展示了什么是最优超平面，用“红色”代表的样本类型 1，“蓝色”代表样本类型－1.

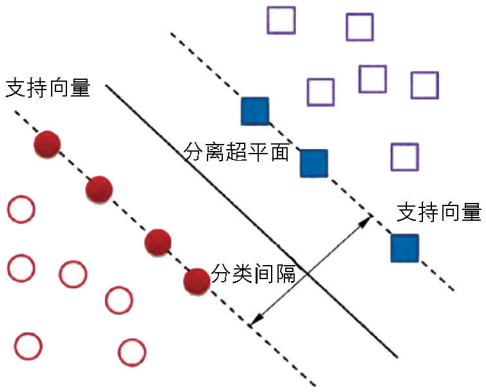


图 5　支持向量确定最优超平面的位置

根据图 4 提取的特征，得到了图 4 昆虫的图像特征值，见表 1.

表 1　特征提取的数值

编号	昆虫大小	昆虫坐标	球状性	单一虫口密度
1	21.4	－574，662	0.42	0.003 42
2	25	－633，666	0.411	0.004 00
3	18.5	－483，653	0.51	0.002 96
4	18.5	－175，657	0.541	0.002 96
5	19	－40，651	0.612	0.003 04
6	15	－462，646	0.632	0.002 40
7	8.5	－29，629	0.23	0.001 36
8	17.5	－564，544	0.547	0.002 80
9	16.5	－287，539	0.491	0.002 64
10	23.6	－128，512	0.788	0.003 78
11	6.5	－320，480	0.891	0.001 04
12	17	－60，456	0.91	0.002 72
13	14.3	－23，469	0.405	0.002 29
14	17.7	－376，456	0.803	0.002 84
15	15.3	674，429	0.709	0.002 45

SVM 可以很好地应用于函数拟合问题，本文采用支持向量机，根据表 1 特征值的类型，设置回归函数的参数，可求得回归式

$$Y=f(x)=\sum_{i=1}^{n}(T_i-T_i^*)(x_i,x)+b \tag{4}$$

其中，T_i 和 T_i^* 为拉格朗日乘子；x_i 为待预测因子向量；x 为支持向量的样本子向量；b 为偏置量.

对非线性问题，要用核函数方法将原始数据映照到高维特征空间，使其转化为线性问题求解，可求得回归式

$$Y=f(x)=\sum_{i=1}^{n}(T_i-T_i^*)K(x_i,x)+b \tag{5}$$

其中：$K(x_i,x)$为支持向量的核函数.

近年来，SVM 这类机器学习算法在实际应用中越来越普及. 余秀丽等[15]利用 SVM 模型对小麦叶部病

害进行识别，结果表明SVM模型的预测识别准确率较高，并且模型算法实现较容易，误差低，对于小麦叶部的病虫害识别成功率有提升效果. 向昌盛等[16]、夏永泉等[17]利用SVM模型对黏虫的病害发生量进行预测，认为该模型可以有效提高病虫害发生的预测精度，对于病虫害发生概率事件，这类规模较小且非线性的样本预测比较适合. 基于前人的研究成果，本文也选取了SVM模型进行建模预测.

3.2 建模方法

一个模型的数据一般分为训练集和测试集，训练集用于模型的自我学习，进行规律的总结；测试集用于验收模型经过训练后的成型效果. 一般来说，训练集占总体数据集比例为75%～85%，测试集为15%～25%，本研究中随机选取了1992,1998,2000,2006,2011,2013的数据作为测试集(表2)，1980－2016年中剩余年份可以作为训练集(由中国气象局国家气候中心和人地系统主题数据库提供，可根据年限地区筛选自由选择).

表2 模型系数表

模型	标准系数			非标准系数
	g	c	d	b
x_1	10.71	0.56	4	0.05
x_2	10.49	0.61	4	0.01
x_3	9.89	0.60	4	0.06
x_4	9.98	0.59	4	0

根据公式(5)建立模型. 将核函数$K(x_i, x)$的参数定义为4个，x_1为当年7－8月平均最低气温；x_2为当年1月降水量；x_3为当年7－9月份平均气温之和；x_4为对应年份特征提取后的虫口密度. 其中，$x_4=\delta$，计算过程将图像处理中的单一虫口密度逐条累加，再根据公式(3)计算.

$$f(x)=\sum_{i=1}^{n=4}(T_i-T_i^*)K(x_1, x_2, x_3, x_4)+b \tag{6}$$

根据归一法，b趋近于0，由于变量多，本文SVM采用的是多项式内积核函数

$$K(x, y)=[g(x*y)+c]^d \tag{7}$$

其中，g,c,d为参数，d为变量个数，即核函数的变量个数为4，可根据数据集的有效数据类型数量进行改变；g为时期参数，对于昆虫而言，成虫、幼虫时期在同一采集面积下的虫口密度肯定不一样，本文由于设备和条件有限，采集病害昆虫都为成虫期. 本文筛选的病虫成虫期的体积大概是该类虫害幼虫期体积的10～14倍，为了方便统计，将虫害的成虫和幼虫图像大小体积显示比设置为10∶1，这里g选择为10；c为区间(0, 1)的一个常数，当该值减小时，对应的直线斜率会减小. 建立模型，模型相关系数α如表2所示.

为保证预测结果直观且有连续性，本文将模型的参考时间序列的长短设置为至少5个连续样本. 在预测第i个样本时，后续样本不会参与模型的训练；在预测第$i+1$个样本时，再将第i个样本进行模型的训练预测. 采用Python建立SVM模型，并以均方根误差(Root mean square error，RMSE)作为指标来衡量多元线性回归，进行SVM性能比对. RMSE定义为

$$RMSE=\sqrt{\frac{\sum_{i=1}^{n}(X_{obs,i}-X_{model,i})^2}{n}} \tag{8}$$

其中，$X_{obs,i}$为实际值；$X_{model,i}$为预测值；n为预测样本数. 表3为部分实验数据展示.

本文样本数据来源：1980－2016年全国各地区小麦病穗率数据，实际值可以从中国气象局国家气候中心进行查询对比.

表 3　部分实验数据集合

年份	$f(x)$	x_1	x_2	x_3	x_4
1992	0.767	25.4	24.0	81.7	11.9
1998	0.770	26.4	1.8	80.9	10.6
2000	0.691	28.9	2.3	81.8	9.2
2006	1.101	25.9	1.6	76.5	10.1
2011	0.792	24.7	24.8	80.7	8.4
2013	0.971	24.5	11.2	78.2	14.5

注：$f(x)$为当年病穗率；x_1 为当年 7－8 月平均最低气温；x_2 为当年 1 月降水量；x_3 为当年 7－9 月份平均气温之和；x_4 为对应年份特征提取后的虫口密度(计算公式为公式(3))．

3.3　SVM 预测结果

通过 Pycharm 软件，通过选取样本建立 SVM 模型，再利用 6 组测试样本对建立的 SVM 模型进行测试，预测结果和实际结果对比见表 4.

表 4　SVM 预测结果

年份	实际值/hm^2	预测值/hm^2
1992	800	767
1998	667	770
2000	733	691
2006	1 000	1 101
2011	760	792
2013	867	971
RMSE		72

根据表 4 可知：SVM 的 *RMSE* 为 72，SVM 模型对 6 组测试样本都达到了准确预测，通过公式(8)计算平均误差为 10%，即预测准确率为 90%．该模型的预测结果可用图 6 中的点来表示，线性直线表示实际的虫害面积，图 6 表明 SVM 预测结果与 6 组测试样本中的实际虫害面积均较为接近．根据图 7、图 8 可以看出，SVM 所描绘的点均围绕残差等于零的直线上下随机散布，说明 SVM 模型对虫害面积预测有较好的效果．

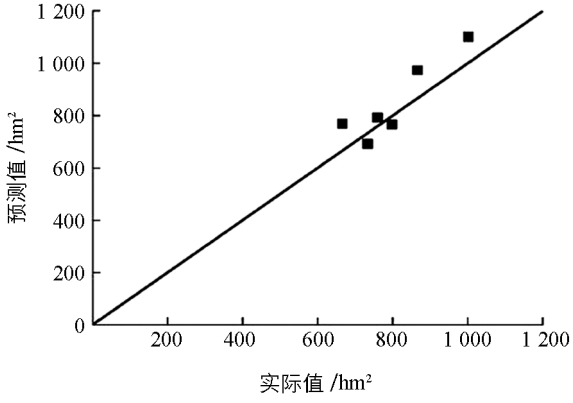


图 6　SVM 模型预测病害面积结果

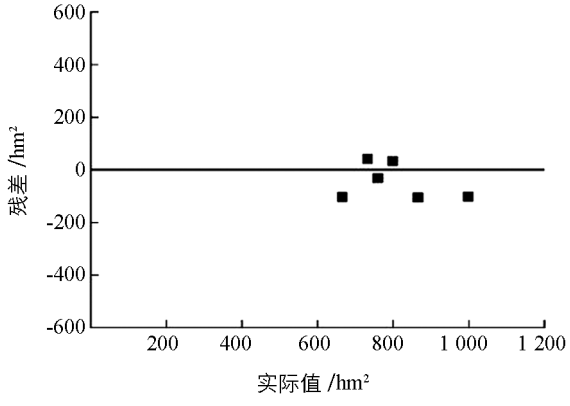


图 7　病害发生面积实际值

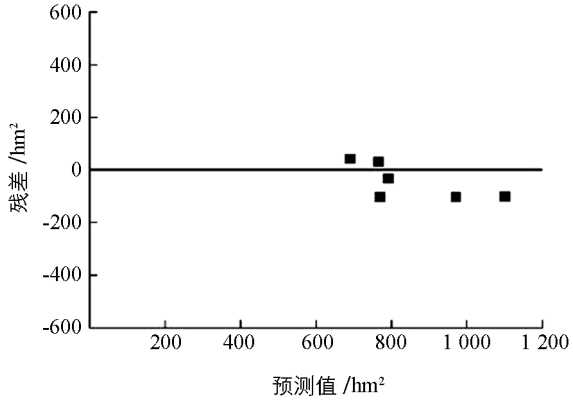


图 8　病害发生面积预测值及偏差

4 结 语

通过实验不难发现，图像处理可以得到害虫的部分特征，而对这部分特征合理地筛选、计算可以得到相应的数据. SVM模型适合处理小样本问题，分类预测的精度较高.

相比于传统的病虫害发生面积预测方法，本研究利用图像处理和SVM减少了传统预测方法中复杂的人工预测部分. 通过利用有限的自变量进行病虫害预测，节约预测时间，节约人力成本，比较符合实际工程上的需求. 预测结果显示预测准确性高，科学有效，证明了本文方法的优越性.

参考文献：

[1] 白庆红. 新形势下森林病虫害的发生动态及防治对策[J]. 现代园艺，2017(15)：150-150.

[2] 张文一，景天忠，严善春. 基于机器学习的落叶松毛虫发生面积预测模型[J]. 北京林业大学学报，2017，39(1)：85-93.

[3] 王淑芬，张 真，陈 亮. 马尾松毛虫防治决策专家系统[J]. 林业科学，1992，28(1)：31-38.

[4] 王霓虹，缪天宇，王阿川. 基于WebGIS的森林病虫害预测预报专家系统的设计与应用[J]. 东北林业大学学报，2008，36(1)：79-80.

[5] 戚 莹. 基于模糊神经网络的森林虫害预测预报的应用研究[D]. 哈尔滨：东北林业大学，2011.

[6] 龚声蓉. 数字图像处理与分析[M]. 北京：清华大学出版社，2014.

[7] 姒绍辉，胡伏原，顾亚军，等. 一种基于不规则区域的高斯滤波去噪算法[J]. 计算机科学，2014，41(11)：313-316.

[8] 张继红. 基于计算机视觉的蚜虫图像识别方法研究[D]. 杨凌：西北农林科技大学，2009.

[9] 袁 锋. 昆虫分类学[M]. 北京：中国农业出版社，2006.

[10] ZHANG Z，LIU T，ZHANG B. Jackknife Empirical Likelihood Inferences for the Population Mean with Ranked Set Samples[J]. Statistics & Probability Letters，2016，108：16-22.

[11] SCHOELKOPF B，SUNG K，BURGES C，et al. Comparing Support Vector Machines with Gaussian Kernels to Radial Basis Function Classifiers[J]. IEEE Transactions on Signal Processing，1996，45(11)：2758-2765.

[12] JENICKA S，SURULIANDI A. Fuzzy Texture Model and Support Vector Machine Hybridization for Land Cover Classification of Remotely Sensed Images[J]. Journal of Applied Remote Sensing，2014，8(1)：083540.

[13] SHAO Y H，CHEN W J，ZHANG J J，et al. An Efficient Weighted Lagrangian Twin Support Vector Machine for Imbalanced Data Classification[J]. Pattern Recognition，2014，47(9)：3158-3167.

[14] 林 卓，吴承祯，洪 伟，等. 基于BP神经网络和支持向量机的杉木人工林收获模型研究[J]. 北京林业大学学报，2015，37(1)：42-47.

[15] 余秀丽，徐 超，王丹丹，等. 基于SVM的小麦叶部病害识别方法研究[J]. 农机化研究，2014(11)：151-155.

[16] 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.

[17] 夏永泉，李耀斌，黄海鹏. 基于平均影响值和支持向量机的小麦病害识别[J]. 电子技术应用，2015，41(6)：136-138，142.

Studies on Pest Prediction Based on Machine Learning and Image Processing Technologies

HANG Li, CHE Jin, SONG Pei-yuan,
WANG Chen-yu, TIAN Bin

School of Physics and Electronic-Electrical Engineering, Ningxia University, Yinchuan 750021, China

Abstract: Traditional pest prediction methods are too complicated and their accuracy is too low. Taking this problem into consideration, the authors put forward in this paper a pest prediction algorithm based on image processing technology and SVM (support vector machine) separator, and forecast the probable incidence of pests and diseases in the coming few years. First, the morphological characteristics of the insects are obtained through image processing techniques such as image filtering and feature extraction. Then, based on the data of pests and diseases in the previous years these characteristics are tagged and scientifically classified to predict the future occurrence of pests and diseases. By constructing a dynamic prediction model, forecasting of pests and diseases is made more efficient, effective and scientific. Finally, the predicted value obtained with the model is compared with the actual value, and the prediction accuracy is as high as 90%. Experimental results show that this method has satisfactory prediction accuracy.

Key words: pest forecast; image processing; morphological characteristics; machine learning; SVM

责任编辑　夏　娟