基于最大似然法集成的黄曲条跳甲预警模型

李亭[1], 杨敬锋[2], 彭晓琴[3], 陈志民[4*]　(1 .中山火炬职业技术学院,广东中山 528436 ;2 .华南农业大学工程学院,广东广州 510640 ;3 .西南财经大学天府学院,四川绵阳 651000 ;4 .华南农业大学公共基础课实验教学中心,广东广州 510640)

摘要　采用最大似然法模型,建立蔬菜黄曲条跳甲的预警模型,并且针对最大似然法一般需要比较多的训练样本才能准确预测的缺点,提出能够显著地提高学习系统的泛化能力的集成算法,即最大似然集成算法以减少对训练样本数量的要求。通过对广东省蔬菜黄曲条跳甲数据验证表明,最大似然集成算法的预警准确率比最近邻算法、k-mean 聚类和支持向量机预警准确率都要高。

关键词　预警;黄曲条跳甲;最大似然法;集成算法
中图分类号　TN18　　**文献标识码**　A　　**文章编号**　0517 - 6611(2008)25 - 10963 - 02

Study on the Early Warning Model of *Phyllotreta striolata* **Based on the Maximum Likelihood Integration**

LI Ting et al　(Zhongshan Torch Vocational College , Zhongshan , Guangdong 528436)

Abstract　The early warning model of *Phyllotreta striolata* in vegetables was set up by using the maximum likelihood model . Aiming at the disadvantage of the maximum likelihood that more training samples were needed for accurate prediction generally , its integration algorithm that could enhance the generalization ability of the learning system significantly was put forward . The maximum likelihood integration algorithm reduced the quantity demands of the training samples . The data of the test on *P . striolata* in vegetables in Guangdong Province , the accuracy rate of early warning by the maximum likelihood integration algorithm was higher than that by nearest-neighbor algorithm, k-mean clustering and support vector machine .

Key words　Early warning ; *Phyllotreta striolata* ; Maximum likelihood ; Integration algorithm

近年来,农作物病虫害发生面积不断增大,发生程度也愈发严重,给农民造成巨大的经济损失。化学农药的不合理使用,造成农业成本提高、品质下降、环境污染等一系列难以解决及回避的环境与社会问题[1]。对农作物病虫害进行有效的预测预报,从而为科学防治提供依据迫在眉睫,为此,已建立了多种相关农业病虫害的预测系统和模型[2-7]。以往的各种预测系统和算法虽然已经取得一定的成效,但是害虫生长周期和发病特征不尽相同,采用单个预测模型的预测值往往受制于固定的训练样本,从而限制了预测准确率。笔者首先建立最大似然法预测模型,并针对最大似然法一般需要比较多的训练样本才能准确预测的缺点,提出最大似然集成算法以减少对训练样本数量的要求。

1　最大似然法

最大似然法(Maximum Likelihood) 是一种统计方法,它用来求一个样本集的相关概率密度函数的参数[8]。给定一个概率分布样本集,假定其概率密度函数(连续) 或者概率聚集函数(离散分布) 以及分布参数。从这个概率分布中抽出一个具有 n 个值的采样,通过利用概率密度函数或者概率聚集函数可以计算出其概率。最大似然法就是寻找关于分布参数的最可能的值,即在所有可能的分布参数取值中,寻找一个分布参数值使这个采样的可能性最大化[9]:

假设由 n 个样本组成的集合 $D=\{x_1, x_2, \cdots, x_n\}$,这些样本都是未标记的,并且是独立地从一个混合密度采样获得的,其混合密度为 $p(x|\theta)=\sum_{j=1}^{c} p(x|\omega_j, \theta_j) P(\omega_j)$;其中,参数向量 θ 具有确定但未知的值。定义样本集的似然函数具有以下的联合概率密度形式为 $p(D|\theta)\equiv\prod_{k=1}^{n} p(x_k|\theta)$;使得该密度达到最大的参数值 θ 就是 θ 的最大似然法 $p(D|\theta)$。

2　最大似然法集成算法

集成学习(Ensemble Learning) 可以显著地提高学习系统的泛化能力[10]。采用 Bagging(Bootstrap Aggregation) 算法作为集成算法[11]。Bagging 算法的目标是提高任何给定的学习算法的准确率,其主要思想是对每一个成员分类器赋予权值,通过测试得到加权组合作为算法的输出。首先给每一个训练样本赋予相同的权重,然后训练第1 个基本分类器并用它来对训练集进行测试,对于那些分类错误的测试样本提高其权重(实际算法中是降低分类正确的样本的权重) ,用调整后的带权训练集训练第2 个基本分类器,重复这个过程直到最后得到一个足够好的学习器[12-13]。

建立 Bagging 分类器集成由原始训练样本集采用重复取样技术(Bootstrap Samples) 生成 T 个训练样本集(S_1, S_2, …, S_T) ,然后由 $S_T(t-1,2,\cdots,T)$ 训练生成相应的 N_t,分类器集成的结果由 N_1, N_2, …, N_T 成员分类器投票表决。Bagging 集成算法为:

输入:训练集 S,学习机 L,重复采样 T 次

输出:集成 N^*

For t = 1 to T

{ S_T = 由原始训练样本集 S 重复取样

$N_t = L(S_T)$ }

End for

$$N^*(x)=\arg\max_{y\in Y}\sum_{t}^{T} N_t(x)=y$$

3　实证

根据广东省的气候特点和作物生长特征,采用地区、蔬菜种类、生长阶段、温度、湿度和天气状况6 个指标作为建模的主要涉及因素。这6 个因素的属性数据从广东省蔬菜病虫害黄曲条跳甲数据库中提取,包括2004 年1 月～2008 年4 月的数据,共有2 649 条记录。其中,2004 ～2006 年的数据,共有1 724 条纪录,2007 年1 月1 日～2008 年4 月16 日共有925 条纪录。根据《蔬菜主要病虫发生程度分级标准》,黄曲条跳甲的预警等级划分标准1、2、3、4、5 级的百株虫量分别为0 ～100、100 ～200、200 ～500、500 ～1 000、> 1 000 。

采用最大似然法集成算法,选取2004 ～2006 年1 724 条记录作为训练样本,2007 年1 月1 日～2008 年4 月30 日925

基金项目　华南农业大学校长基金项目(2007K017) 。
作者简介　李亭(1979 -) ,女,河南周口人,硕士,助教,从事数据挖掘与智能计算方面的研究。* 通讯作者。
收稿日期　2008-06-10

条记录作为预测样本。表1 列出了混淆矩阵结果。表1 中第2 列表示,在最大似然法预警结果中,等级1 被正确预警的数量为122 个,实际预警结果为等级2 的有28 个被最大似然法错误预警为等级1,实际预警结果为等级3 的有41 个被最大似然法错误预警为等级1,如此类推。

从表1 可以看出,最大似然法预警出现"跳级"错误预警情况比较严重,并且无法对等级5 的样本准确预警;而最大似然法集成预警则仅仅出现少量"跳级"错误预警,对角线以下的错误预警比对角线以上的多,说明算法的预警比实际的预警的等级更高。表2 是最大似然法集成与其他算法结果的比较。

表1 2007 年1 月～2008 年4 月黄曲条跳甲混淆矩阵结果

Table 1 The confusion matrix results of Phyllotreta striolata **from January of 2007 to April of 2008**

算法 Algorithm	测试值 Test value	实际值 Practical value					准确率//% Accuracy rate
		等级1 Level 1	等级2 Level 2	等级3 Level 3	等级4 Level 4	等级5 Level 5	
最大似然法预警	等级1 Level 1	122	28	41	0	0	62.27
Early warning	等级2 Level 2	66	107	14	7	3	
by the maximum	等级3 Level 3	26	34	280	31	17	
likelihood	等级4 Level 4	17	24	20	67	21	
	等级5 Level 5	0	0	0	0	0	
最大似然法集成预警	等级1 Level 1	142	18	0	0	0	74.38
Early warning by	等级2 Level 2	78	139	18	1	0	
the maximumlike-	等级3 Level 3	11	22	295	23	2	
lihood integration	等级4 Level 4	0	14	42	81	8	
	等级5 Level 5	0	0	0	0	31	

从表2 可以看出,最近邻算法是所列方法中准确率最低的;k-mean 聚类和支持向量机的预警准确率比最近邻算法都有所提高;而采用最大似然法集成预警算法,预警准确率达到了74.38%,比最近邻算法、k-mean 聚类和支持向量机预警准确率分别提高了25.19%、23.36%、8.43%,达到令人较为满意的预警准确率。

表2 黄曲条跳甲预警结果

Table 2 The early warning results of Phyllotreta striolata

算法 Algorithm	准确预警//次 Accurate early warning	预警准确率//% Accurate rate of early warning
最大似然法预警	576	62.27
Early warning by the maximumlike-lihood		
最大似然法集成预警	688	74.38
Early warning by the maximumlike-lihood integration		
k-mean 聚类 k-mean clustering	472	51.03
最近邻算法	455	49.19
Nearest-neighbor algorithm		
支持向量机	610	65.95
Support vector machine		

4 结论

(1) 最大似然法一般要求数量比较大的训练样本,而最大似然法集成算法可减少其对训练样本的要求。

(2) 采用最大似然法集成算法对黄曲条跳甲进行预警,预警结果表明,最大似然法集成算法比最近邻算法、k-mean 聚类和支持向量机预警准确率都要高。

(3) 从混淆矩阵结果可以看出,最大似然法集成算法仅仅出现少量严重的"跳级"错误预警,算法的预警比实际的预警的等级更高。

参考文献

[1] 孙虎. 小麦全蚀病的生物防治研究及品种抗性鉴定[D]. 郑州: 河南农业大学,2004.

[2] 李祚泳,彭荔红. 基于人工神经网络的农业病虫害预测模型及其效果检验[J]. 生态学报,1999,19(5):759-761.

[3] 张建兵,诸叶平. 基于模糊规则的病虫害预防研究[J]. 农业系统科学与综合研究,2000,16(4):283-285.

[4] 彭晓琴,杨敬锋,胡月明,等. 基于半监督学习的黄曲条跳甲预警方法[J]. 农机化研究,2008,30(3):150-153.

[5] 胡小平,梁承华,杨之为,等. 植物病虫害BP 神经网络预测系统的研制与应用[J]. 西北农林科技大学学报,2001,29(2):73-76.

[6] 熊雪梅,姬长英. 基于参数化遗传神经网络的植物病害预测方法[J]. 农业机械学报,2004,35(6):110-114.

[7] 任春风,李淼. 病虫害预测预报中适应性函数的研究[J]. 计算机工程与应用,2007,43(6):197-243.

[8] ALDRICH JOHN R A. Fisher and the making of maximumlikelihood 1912-1922 [J]. Statistical Science,1997,12(3):162-176.

[9] DUDA R I O, HART P E, STORK D G. Pattern classification[M]. Second Edition. John Wiley & Sons Inc,2001:549-556.

[10] DIETTERICH T G. Machine learning research: Four current directions[J]. AI-Magazine,1997,18(4):97-136.

[11] BREIMAN L. Bagging predictors[J]. Machine Learning,1996,6(2):123-140.

[12] FREUND YOAV, SCHAPIRE ROBERT E. A decision-theoretic generalization of on-line learning and an application to boosting[J]. Journal of Computer and System Sciences,1997,55(1):119-139.

[13] SCHAPIRE ROBERT E. The boosting approach to machine learning: An overview[M]. Proc MSRI Workshop Nonlinear Estimation and Classification, 2002.