

文章编号:1001-9081(2007)05-1177-03

CFW 的 CBR 与 ART-KNN 集成智能预测

杨振刚^{1,2}, 邓飞其²

- (1. 华南农业大学 信息学院, 广东 广州 510642;
2. 华南理工大学 自动化科学与工程学院, 广东 广州 510640)
(yzg817@163.com)

摘要:结合基于案例推理(CBR)方法和 ART-KNN 网络,提出了一种黄瓜枯萎病(CFW)的集成智能预测方法。与传统的 CBR 相似案例检索任务不同的是,该方法用受训 ART-KNN 网络对新案例分类后根据提出的案例相似性测度来计算相似案例集。对 ART-KNN 网络的分类性能进行测试,确定了网络的最优相似参量 ρ , 得到最高平均分类正确率达 94.4%。对 CFW 进行预测,确定了案例相异阈值 R 的最优范围,得到病株率、病叶率的最优平均预测误差率分别达 7.4%、9.3%。综合分析结果表明,提出的 CBR 与 ART-KNN 集成预测方法可为 CFW 的防治提供较为可靠的预测数据以及辅助决策信息。

关键词:黄瓜枯萎病;智能预测;基于案例推理;ART-KNN

中图分类号: TP181 **文献标识码:** A

Intelligent forecast integrated with CBR and ART-KNN for CFW

YANG Zhen-gang^{1,2}, DENG Fei-qi²

- (1. College of Informatics, South China Agricultural University, Guangzhou Guangdong 510642, China;
2. College of Automation Science and Engineering, South China University of Technology, Guangzhou Guangdong 510640, China)

Abstract: Integrated with case-based reasoning (CBR) and Adaptive Resonance Theory-Kohonen Neural Network(ART-KNN), a Cucumber Fusarium Wilt(CFW) intelligent forecast method was proposed in this paper. Different from the traditional similar case retrieval, this method performs case classification with trained ART-KNN and then figures out a similar case set using a proposed case similarity metric. The optimal similarity parameter ρ of ART-KNN was determined by its classification performance test, and the best mean classification accuracy rate was 94.4%. The optimal interval of case dissimilarity threshold R for CBR was determined by a CFW forecast test, and the best mean prediction error rate of the ill plant rate and the ill leaf rate were 7.4% and 9.3% respectively. Comprehensive analysis results show that the proposed method can provide reliable prediction data and assistant decision information for CFW prevention and treatment.

Key words: Cucumber Fusarium Wilt(CFW); intelligent forecast; Case-based Reasoning(CBR); Adaptive Resonance Theory-Kohonen Neural Network(ART-KNN)

0 引言

据中国植保部门的病情通报,近年作物病害的发病率在 7%~12% 左右,且病害发病率有上升趋势,农作物病害防治工作依然严峻,其成败关乎整个农业的健康发展。传统的黄瓜病害预测是依靠经验进行人工预测或根据预设模型利用软件工具进行预测,其缺点是不能根据实时气候及生态条件动态地对作物病害进行智能预测。因此,本文提出集成基于案例推理(CBR)和 ART-KNN 来研究黄瓜枯萎病 CFW 智能预测方法。CBR 方法是通过适应(修订)先前相似案例来解决新面对的问题^[1],它不需要建立非常明确的作物病害知识模型,而收集病害预测历史案例及开发有效的推理算法就变得相对重要。CBR 方法要通过大量的案例来获取新知识,这是它适合于需要经验知识导向的作物病害预测领域的原因。传统的 CBR 相似案例识别任务可通过对案例库进行遍历检索,用源自 Minkowski 方法的测度^[2]、无加权相似性测度^[3]、异类

相似性测度^[4]或者异类权重敏感测度^[5]将新案例与旧案例逐一进行计算比较,确定新案例的相似案例集。其缺点是不同的案例相似性测度方法产生的案例匹配结果相异较大,对于同时含有数值型和标称型案例描述属性的案例来说,选择适宜的相似性测度显得相对较难。因此,本研究提出将 ART-KNN 神经网络集成到 CBR 过程的相似案例检索中,其思想是先用已分类好的案例对 ART-KNN 进行训练,在检索相似案例时用受训的 ART-KNN 网络进行案例分类,然后在该类案例中用提出的案例相似性测度来计算确定相似案例集,以提高案例检索的性能和效率。在基于 CBR 和 ART-KNN 的集成预测方法下,CFW 可能发生的情况和处理方法可以有效预知,从而增加其防治胜算并可辅助生产决策的动态修订。

1 CBR 与 ART-KNN 的集成预测原理

1.1 CBR

CBR 执行过程如图 1 所示,具体包括四大主要任务:检

收稿日期:2006-11-03;修订日期:2007-01-15

基金项目:国家自然科学基金资助项目(60374023);华南农业大学校长基金资助项目(2006K040)

作者简介:杨振刚(1976-),男,广西北海人,讲师,博士研究生,主要研究方向:信息系统工程;邓飞其(1961-),男,湖南益阳人,教授,博士,主要研究方向:控制理论与系统工程。

索相似案例、重用相似案例并推断新案例解决方案、修订解决方案、保存新案例以备后用^[6]。本质上 CBR 是一种依据原有经验来模拟人脑推理的方法学,它先将大量的问题及其解决方案以案例形式进行表达和存储,当新问题(案例)出现时,系统根据案例描述用一定的相似性测度匹配案例库中的旧案例,检索出最相似案例并据此重用、修订其解决方案以产生新问题的解决方案。当新案例具有典型性时将其保留到案例库,实现 CBR 一次学习过程。

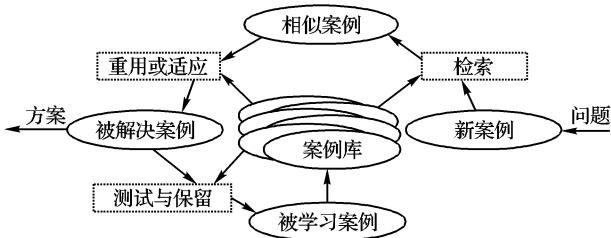


图1 CBR 执行过程

1.2 集成系统框架

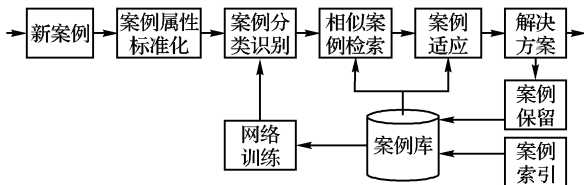


图2 集成预测系统框架

集成预测系统的框架如图 2 所示。依据旧案例中的解决方案对案例库进行分类后对每类案例建立索引,然后用各类

表 1 案例库

案例 ID	问题描述部分 (ATR)			方案部分 1 (PDT)			方案部分 2 (STM)		
	温度	土壤 PH 值	...	病株率	病叶率	...	化学防治	...	
1	25.2℃	6.8	...	9.5%	18.7%	...	50%多菌灵可湿性粉剂 500 倍液	...	
2	25.6℃	6.7	...	10.1%	20.3%	...	50%多菌灵可湿性粉剂 500 倍液	...	
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	
107	27.3℃	6.9	...	23.6%	37.0%	...	40%多菌灵胶悬剂 400 倍液	...	
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	
432	26.2℃	6.8	...	25.6%	36.8%	...	瑞代合剂 (1:2) 140 倍液	...	

1.4 ART-KNN 网络及其训练

1.4.1 ART-KNN 网络

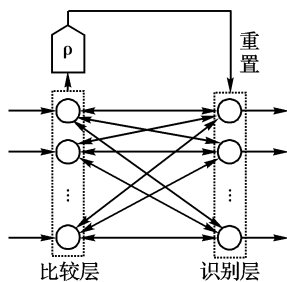


图3 ART-KNN 神经网络结构

ART-KNN 网络结合了自适应谐振理论网络和 Kohonen 学习策略,结构如图 3 所示。与 ART1 网络相似,ART-KNN 网络由比较层(输入层)和识别层(输出层)组成,通过反复地将输入学习模式由输入层向输出层自左而右地识别、由输出层向输入层自右而左地比较来实现分类学习的。当自左而右的识别和自右而左的比较达到共振,即输入向量可以正确反映输入学习模式的分类而且网络原有记忆没有受到不良影响时,一个输入学习模式的记忆和分类就可完成。当输入模式与样板模式在比较层不能匹配时网络向识别层发送一个重置

信号^[7]。换言之,当输入模式被确认为新模式时,识别层阻止原先组织好的神经元去学习该模式,同时适应网络结构并立即以另外的权重存储新模式。当一个输入模式根据相似标准得到匹配时,网络将改变获胜神经元的权重。

1.3 案例表达

CBR 方法需要正确描述表征案例的问题属性及方案属性。设系统案例库 $CB = \{c_a | a = 1, 2, \dots, n\}$, 案例 c_a 问题描述部分属性集合 $ATR_a = \{atr_{a1}, \dots, atr_{al}\}$, 其中 atr_{al} 为第 l 个问题描述属性。案例 c_a 的方案部分 $SLT_a = \{PDT_a, STM_a\} = \{\{pdt_{a1}, \dots, pdt_{ah}\}, \{stm_{a1}, \dots, stm_{ak}\}\}$, 其中 PDT_a 为预测属性集合, STM_a 为病害处理措施及其他描述内容集合。因此案例可以表示为 $c_a = \{ATR_a, PDT_a, STM_a\}$ 。

CFW 是世界性病害,在中国各地都有发生。由于连年耕作,黄瓜产地土壤中的枯萎病菌量会逐年积累增多,发病率增大。CFW 病菌以菌丝体、厚垣孢子和菌核的形式,在土壤里、未腐熟的农家肥里或附着在种子和病株残体上越冬,成为翌年的初侵染源。CFW 的发生同时受自然环境、生态条件等因素影响,如气温暖和、土壤偏酸、土质粘重等都会助长病菌的生长。因此 CFW 案例的问题部分为影响病害发生的自然条件及生态条件,如温度、土壤 PH 值、土壤湿度、平均菌量、作物生长期、土壤线虫密度、品种特性和栽培方式 ($l = 8$)。方案部分的预测属性集合 PDT 包括病株率、病叶率 ($h = 2$), 而处理措施及其他描述内容集合 STM 包括化学防治方法、生态防治方法及其他防治方法等 ($k = 13$)。案例库如表 1 所示。

信号^[7]。换言之,当输入模式被确认为新模式时,识别层阻止原先组织好的神经元去学习该模式,同时适应网络结构并立即以另外的权重存储新模式。当一个输入模式根据相似标准得到匹配时,网络将改变获胜神经元的权重。

在网络的 Kohonen 学习策略^[8]中,输入向量 X 与识别层中神经元间的匹配程度用欧氏距离进行评价,产生获胜神经元后输入向量 X 反馈至比较层。相似系数 S_c 定义为:

$$S_c = 1 - \|W_j - X\| / \|W_j\| \quad (1)$$

W_j 为获胜神经元权重,它与 X 越相似,系数 S_c 越趋近 1,引入相似参量 ρ 作为相似评价标准时,当 $S_c > \rho$ 表明输入模式 X 与第 J 类模式充分相似。为使代表相应类的权重更准确,第 J 类的权重通过下式进行改进:

$$W_j = \frac{X + nW_j^0}{n + 1} \quad (2)$$

其中 W_j 为改进后权重, W_j^0 为原始权重, n 为迭代次数。当 $S_c < \rho$ 时,输入模式 X 不属于第 J 类,网络模型中没有与 X 相匹配的类,需要调整网络,通过识别层重置方式增加新的神经元记忆新的模式。新神经元权重设置为 $W_{n+1} = X$ 。

1.4.2 ART-KNN 网络训练

设 f 类共 F 个案例被输入网络进行训练。ART-KNN 网络

的训练程序可描述如下:

(1) 训练前 ART-KNN 网络为空,第一个案例输入时没有比较神经元,故增加神经元记忆该案例;

(2) 计算输入案例 X 与每个已记忆神经元 N 的欧氏距离,确定

$$\|W_j - X\| < \|W_j - X\|, j, J = 1, 2, \dots, f; j \neq J \quad (3)$$

即神经元 N_j 被视作获胜神经元。

(3) N_j 与输入案例 X 的相似性用(1)式进行评价。判断条件 $S_c > \rho$, 为真时确认案例 X 属于 N_j 对应的第 j 类案例且 N_j 的权重用(2)式进行改进,为假时增加新的神经元记忆案例 X 。

(4) 重复第 2,3 步直至所有 F 个案例训练完毕。

1.5 相似案例检索

向系统输入新案例 X 并标准化属性后,受训 ART-KNN 网络对其进行所属案例类识别。设识别结果为 $X \in C_j$ 类 ($j \in \{1, 2, \dots, f\}$), 且 C_j 类有 m_j 个案例,则依次计算 C_j 中每个案例与 X 的相异距离,即:

$$dt(c_a, X) = \sqrt{\sum_{s=1}^l [w_s (atr_{as} - atr_x)]^2}, a = 1, \dots, m_j \quad (4)$$

其中 w_s 为 ATR 中第 s 个属性的重要性权值,它是影响案例解决方案的案例问题属性的重要度,总体上温度、土壤 PH 值、平均菌量、土壤湿度等属性权值相对较大。新案例 X 与 C_j 类案例中 c_a 之间的相异距离小于等于预定义的相异阈值 R 时,系统认定 c_a 为 X 的相似案例集元素。即最终相似案例集为 $SCS_X = \{c_a \mid dt(c_a, X) \leq R, a = 1, 2, \dots, m_j\}$ 。

1.6 案例适应

案例方案部分中的 PDT 属性为 CFW 病害影响预测属性,其中病株率、病叶率根据加权平均法修正。设相似案例集 SCS_X 中的案例数为 g , 即 $SCS_X = \{c_s \mid s = 1, \dots, g\}$ 。设 c_a 为 SCS_X 中与 X 相异距离最小的案例,当 $g = 0$ 时,重用 c_a 的 PDT 属性,即 $PDT_X = PDT_{c_a}$; 当 $g = 1$ 时,重用 c_1 的 PDT 属性,即 $PDT_X = PDT_{c_1}$; 当 $g > 1$ 时,设 SCS_X 中 g 个案例的权重为 $w_s^X (s = 1, \dots, g)$, 定义:

$$w_s^X = \frac{[dt(c_{\#}, X) / dt(c_s, X)]}{\sum_{c_a \in SCS_X} [dt(c_{\#}, X) / dt(c_a, X)]}, s = 1, \dots, g \quad (5)$$

则预测属性修正为

$$pdt_x = \sum_{s=1}^g [w_s^X \times pdt_{si}], i = 1, \dots, h \quad (6)$$

案例方案部分中的 STM 属性为 CFW 案例处理措施,其中描述性属性如药剂、药含量、倍液数、使用方式、其他防治措施等可直接重用,而药剂用量、使用天次等属性通过经验公式或 CFW 防治知识推理得出。

2 实验与结果分析

实验案例库包含 8 类共 432 个 CFW 防治案例,其中四分之三共 324 个案例用于训练 ART-KNN 网络,其余四分之一共 108 个案例用于测试系统性能。在相似参量 ρ 的不同取值下用所有测试案例对 ART-KNN 网络的分类性能进行测试,所得案例平均分类正确率与 ρ 的关系如图 4 所示。实验结果表明,相似参量 ρ 越大,平均分类正确率 ($MCAR$) 越高, $\rho = 0.9$ 和 $\rho = 1.0$ 时分类正确率均达 94.4%。由于 ρ 越大,ART-KNN

网络的计算量越大,故取最优 ρ 为 0.9。

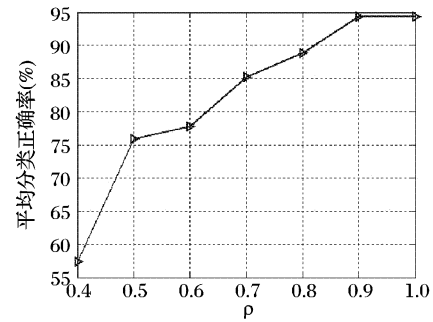


图 4 平均分类正确率

相似案例检索的结果会受相异阈值 R 的影响,进而影响到 PDT 中属性的预测值。为确定最优相异阈值 R , 基于 $\rho = 0.9$ 的条件,分别在不同 R 值下用所有测试案例对 CFW 进行预测,结果用平均预测误差率表示,定义为:

$$MPER_j = \left[\frac{1}{n_T} \sum_{i=1}^{n_T} \left(\frac{|pdt_i^i - pdt_i^{j*}|}{pdt_i^i} \right) \right] \times 100\%, j = 1, 2 \quad (7)$$

其中 $MPER1$ 为病株率平均预测误差率, $MPER2$ 为病叶率平均预测误差率, n_T 为测试案例数, pdt_i^i 为第 i 测试案例的第 j 个 PDT 属性的实际值, pdt_i^{j*} 为第 i 测试案例的第 j 个 PDT 属性的预测值。在不同 R 下用 108 个测试案例分别对 CFW 进行预测,统计得到的平均预测误差率如图 5 所示。结果表明,随着相异阈值 R 增大, $MPER1$ 和 $MPER2$ 均出现先减小后增大的趋势,原因是 R 很小时检索得到的相似案例较少(甚至 SCS_X 可能为空),因而预测误差会相对大些,而 R 值较大时,许多准相似案例被归到相似案例集中,使预测结果可靠性降低。故最优相异阈值 $R \in [0.2, 0.3]$, 当 $R = 0.2$ 时, $MPER1 = 10.2\%$, $MPER2 = 9.3\%$; 当 $R = 0.3$ 时, $MPER1 = 7.4\%$, $MPER2 = 12.0\%$ 。

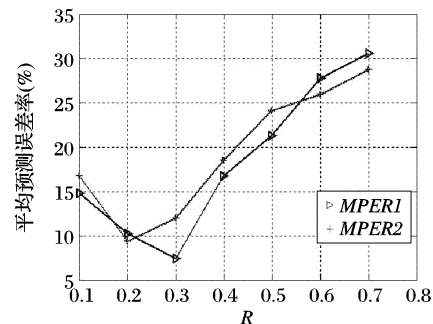


图 5 平均预测误差率

3 结语

结合 CBR 和 ART-KNN 网络,本文提出了一种 CFW 的集成智能预测方法。与传统的 CBR 相似案例检索任务不同的是,这种方法用受训 ART-KNN 网络对新案例分类后用提出的案例相似性测度来计算相似案例集。案例适应任务中对 PDT 属性采用加权平均法修正,而 STM 属性依据知识来推理。在相似参量 ρ 的不同取值下对 ART-KNN 网络的分类性能进行测试后发现,相似参量 ρ 越大,平均分类正确率 ($MCAR$) 越高,最优 ρ 为 0.9,最高平均分类正确率达 94.4%。在 $\rho = 0.9$ 的条件下,基于不同相异阈值 R 对 CFW 进行预测,结果表明随着相异阈值 R 增大,CFW 的病株率平均预测误差率

(下转第 1203 页)

