融入监督信息的 *k-mean* 聚类瓜蓟马预警模型

陈志民[1],李 亭[2*],杨敬锋[1,3],彭晓琴[4]　(1. 华南农业大学公共基础课实验教学中心,广东广州 510642;2. 中山火炬职业技术学院,广东中山 528436;3. 广东瑞图万方科技有限公司,广东顺德 528305;4. 西南财经大学天府学院,四川绵阳 651000)

摘要　[目的]为提高瓜蓟马病虫害的预警效果。[方法]采用 *k-mean* 聚类建立了瓜蓟马预警模型,并针对瓜蓟马数据中在 *k-mean* 聚类算法下难以判断的情况,引入了监督信息,即模糊关联规则进行进一步划分。[结果]引入监督信息的 *k-mean* 聚类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。[结论] *k-mean* 聚类过程中引入模糊关联规则能较有效地提高预警准确率。

关键词　预警;*k-mean* 聚类;模糊关联规则;瓜蓟马
中图分类号　TP18　　**文献标识码**　A　　**文章编号**　0517 - 6611(2009)30 - 14738 - 02

The Warning Model of Melon Thrips Based on *k-mean* Clustering Combining Fuzzy Association Rules Algorithm

CHEN Zhi-min et al　(Center of Experimental Teaching for Common Basic Courses, South China Agricultural University, Guangzhou, Guangdong 510642)

Abstract　[Objective] The aim of the study was to improve warning effect of melon thrips diseases and insect pests. [Method] The warning model of melon thrips by *k-mean* Clustering was firstly established. Considering the challenge of quantity of samples that were difficult to classify in the process of *k-mean* Clustering, an iterative algorithm combining fuzzy association rules was discussed. [Result] The study showed that the warning accuracy of *k-mean* Clustering combining fuzzy association rules provided a higher accuracy rate than that of nearest Neighbor Clustering, *k-means* Clustering and Support Vector Machine in the same condition. [Conclusion] *k-mean* Clustering combining fuzzy association rules algorithm could effectively improve warning accuracy rate.

Key words　Warning ; *k-mean* Clustering; Fuzzy association rules; Melon thrips

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、环境污染等一系列问题[1]。因此,对农作物病虫害进行有效的预测预报,为科学防治提供依据迫在眉睫。为此,许多研究者进行了大量研究[2-9]。已往的各种预测系统和算法虽然已取得一定的成效,但是病虫害生长周期、影响环境以及发病特征不尽相同,采用单个预测模型的预测值往往受制于固定的训练样本,对其他病虫害数据类型和结构往往不能推广,从而限制了预测准确率。笔者针对蔬菜病虫害瓜蓟马数据在使用 *k-mean* 聚类过程中出现数量较多难以通过距离计算判断测试样本属于哪个聚类中心的特点,提出引入模糊关联规则,对这些样本进行进一步判别的改进算法,以达到完全划分的目的。

1　研究方法

1.1　*k-mean* 聚类算法[10]　*k-mean* 算法把对象集合 D 划分成一组聚类 $\{C_1, C_2, \cdots, C_k\}$,其中,$\bigcup_{i=1}^{k} C_i = D$,$k$ 是聚类的个数。聚类的结果可以用一个隶属矩阵 $W = \{w_{ij}, 1 \leqslant i \leqslant n, 1 \leqslant j \leqslant k\}$ 来表示,而 $M_{hk} = \{W \mid w_{ij} \in \{0,1\}, \forall i,j; \sum_{i=1}^{k} w_{ij} = 1, \forall i \in [1,n]; \sum_{i=1}^{n} w_{ij} > 0, \forall j \in [1,k]\}$ 则为 D 的 k 组聚类空间。

k-mean 聚类的目标函数为 $\sum_{j=1}^{k}\sum_{i=1}^{n} w_{ij} d(x_i, z_j)$,其中,$x_i$ 是第 i 个对象;z_j 是第 j 个聚类的中心。*k-mean* 算法的步骤为:首先随机选取 k 个初始聚类中心,把每个对象分配给离其最近的据点,从而得到一组聚类。然后计算当前每个聚类的中心作为新的聚点,把每个对象重新分配到最近的聚类中心。如果满足终止条件则算法结束,否则用新聚类代替原聚类。该研究采用最常用的欧式距离作为黄曲条跳甲数据库中样本相似性度量的标准,模式样本向量 x 与中心 z 的距离为:$D = d(x_i, z_j) = \sqrt{\sum_{k=1}^{m} |x_{ik} - z_{jk}|^2}$。

1.2　引入监督信息的 *k-mean* 聚类　*k-mean* 聚类算法中,测试样本与分别两个聚类中心距离相同难以进行划分的情况虽然很容易出现,但是一般不普遍,对测试结果的影响也不大,对于数量比较少的这一类测试样本通常很少专门设置算法进行处理。然而,所采集的瓜蓟马数据中出现了数量较大的类似情况,与不同聚类中心具有相同距离难以进行划分的测试样本直接影响了测试准确率。改进的方法为:引入模糊关联规则,对这些样本进行进一步判别,以达到完全划分的目的。

设属性集合 $I = \{A_1, A_2, \cdots, A_m\}$ 和数据库 $D = \{d_1, d_2, \cdots, d_n\}$,模糊关联规则 $A \Rightarrow B$ 的最小支持度和最小置信度的计算表达式分别为[11-12]:

$$\text{Support}(A \Rightarrow B) = \frac{\sum_{d \in D}(\mu_{A \wedge B}(d) \mid \mu_{A \wedge B}(d) \geqslant \varepsilon)}{|D|}$$

$$\text{Confidence}(A \Rightarrow B) = \frac{\sum_{d \in D}(\mu_{A \wedge B}(d) \mid \mu_{A \wedge B}(d) \geqslant \varepsilon)}{\sum_{d \in D}(\mu_{A}(d) \mid \mu_{A}(d) \geqslant \varepsilon)}$$

式中,ε 为指定的阈值;D 为事务数据库总集;$|D|$ 为事务总数;$\mu_A(d)$ 为样本 d 对模糊集 A 的模糊隶属度;$\mu_{A \wedge B}(d)$ 为样本 d 对模糊集 $A \wedge B$ 的模糊隶属度;$\sum_{d \in D}(\mu_{A \wedge B}(d) \mid \mu_{A \wedge B}(d) \geqslant \varepsilon)$ 为项集 $A \wedge B$ 中 $\mu_{A \wedge B}(d)$ 大于指定阈值 ε 的和;$\sum_{d \in D}(\mu_A(d) \mid \mu_A(d) \geqslant \varepsilon)$ 为项集 A 中 $\mu_A(d)$ 大于指定阈值 ε 的和。

结合 *k-mean* 聚类算法,模糊关联规则的引进的步骤是:①随机抽取训练样本,并进行 *k-mean* 聚类,得到聚类中心;②从随机抽取的训练样本挖掘相应的模糊关联规则,并以模糊关联规则中的预警等级作为该模糊关联规则的标签;③寻找聚类中心与所挖掘出来的模糊关联规则以最大隶属度为匹配标准的最匹配模式,并把该模糊关联规则对应的标签作为聚类中心的预警等级。所采用的匹配算法为高斯曲线的隶属函数,论域中的元素 x 对模糊子集 F 的隶属度为:$\mu_F(x)$

基金项目　华南农业大学校长基金项目(2007K017)。
作者简介　陈志民(1981 -),男,广东湛江人,实验师,从事数据挖掘与智能计算方面的研究。* 通讯作者。
收稿日期　2009-06-15

$=\exp[-(\frac{x-c}{\sigma})^2]$，其中 σ 为方差，c 为中心值；④测试样本中能直接被带标签的聚类中心判别等级则通过 *k-mean* 算法进行聚类，并以该聚类中心的标签作为输出；若测试样本与分别两个聚类中心距离相同，则用模糊关联规则作为聚类中心通过计算测试样本与模糊关联规则的距离，以距离最短的模糊关联规则的标签作为该测试样本的输出。

2　资料来源

主要数据来源于广东省蔬菜病虫害瓜蓟马数据资料。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。该 6 个因素的属性数据从广东省蔬菜病虫害瓜蓟马数据库中提取，包括 2004 年 1 月 ~2008 年 9 月的数据，共有 1 807 条记录，其中 2004 ~2006 年的数据共有1 049 条记录，2007 年 1 月 1 日 ~2008 年 9 月 18 日共有 758 条记录。根据《蔬菜主要病虫发生程度分级标准(2006 年 11 月，第二版)》，瓜蓟马的预警等级划分标准如表 1。百株虫量越多，预警等级越高。

表 1　瓜蓟马的预警等级划分标准

Table 1　The forecast level dividing standard for melon thrips

预警等级 Forecast level	百株虫量 Larvae amount per 100 plants
1	0 ~100
2	100 ~200
3	200 ~500
4	500 ~1 000
5	>1 000

3　结果与分析

选取 2004 ~2006 年 1 049 条记录作为训练样本，2007 年 1 月 1 日 ~2008 年 9 月 18 日 758 条记录作为预测样本。混淆矩阵结果见表 2，由表 2 可知，在 *k-mean* 聚类算法预警结果中，等级 1 被正确预警的数量为 58 个，实际预警结果为等级 2 的有 109 个被 *k-mean* 聚类算法错误预警为等级 1，实际预警结果为等级 3 的有 8 个被 *k-mean* 聚类算法错误预警为等级 1，如此类推。

表 2　2007.1 ~2008.9 瓜蓟马混淆矩阵结果

Table 2　Fusion matrix result of melon thrips from Jan. 2007 to Sep. 2008

算法 Algorithm	测试值 Test value	实际值 Actual value					准确率//% Accuracy rate
		等级 1	等级 2	等级 3	等级 4	等级 5	
k-mean 聚类	等级 1	58	109	8	0	0	50.13
	等级 2	31	161	27	34	13	
	等级 3	11	28	97	35	13	
	等级 4	4	3	35	64	27	
	等级 5	0	0	0	0	0	
引入监督信息的 *k-mean* 聚类	等级 1	75	43	0	0	0	65.17
	等级 2	17	214	24	7	0	
	等级 3	12	31	111	35	15	
	等级 4	0	13	32	71	15	
	等级 5	0	0	0	20	23	

从混淆矩阵可以看出，*k-mean* 聚类算法预警出现“跳级”错误预警情况较严重，且无法对等级 5 的样本准确预警；而引入监督信息的 *k-mean* 聚类算法预警则仅出现少量“跳级”错误预警，对角线以下的错误预警比对角线以上的多，说明引入监督信息的 *k-mean* 聚类算法的预警比实际的预警等级更高。

引入监督信息的 *k-mean* 聚类算法与其他算法结果的比较见表3。由表3 可知，最近邻算法的准确率最低；*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；而采用引入监督信息的 *k-mean* 聚类预警算法预警准确率达 65.17%，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 20.58%、15.04% 和 15.54%，达到较满意的预预警准确率。

表 3　瓜蓟马预警结果

Table 3　Forecast result of melon thrips

算法 Algorithm	准确预警数量 Correct forecast amount	预警准确率//% Forecast accuracy rate
k-mean 聚类	380	50.13
引入监督信息的 *k-mean* 聚类	494	65.17
最近邻算法	338	44.59
支持向量机	452	59.63

4　结论

(1)该研究提出的在 *k-mean* 聚类过程中引入模糊关联规则能较有效地提高预警准确率，对难以划分的测试样本进行进一步比较准确地划分。

(2)与最近邻算法、支持向量机算法相比，引入监督信息的 *k-mean* 聚类取得更高的准确率。

(3)从混淆矩阵的结果来看，引入监督信息的 *k-mean* 聚类的预警等级比实际的预警等级更高。

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] 吴小芳，包世泰，胡月明，等. 多因子空间插值模型在农作物病虫害监测预警系统中的构建及应用[J]. 农业工程学报，2007，23(10)：162 -166.
[7] 张谷丰，朱叶芹，翟保平. 基于 WebGIS 的农作物病虫害预警系统[J]. 农业工程学报，2007，23(12)：176 -181.

(下转第 14754 页)

的用量比例做了更进一步的分析。由表 4 可知,乳化剂对 602#、CJ2 及 CJ1 的用量比例为 9∶3∶5 时合格。

2.3 水的选择 以不同水质配制浓度 10% 除尽微乳剂,进行稳定性试验。由表 5 可知,水质硬度对制剂的稳定性影响不是很大。因此,从保证产品质量和成本最低化等方面考虑,在生产过程中使用自来水就可以生产浓度 10% 除尽微乳剂。

表 4 乳化剂 602#、CJ2 和 CJ1 最佳配比筛选结果

Table 4 Screening results of optimum proportion of emulsifiers 602#、CJ2 and CJ1

序号 Code	比例 Percentage	起始外观 Initial appearance	乳液稳定性 Emulsion stability	冷贮外观 Cold storage appearance	热贮外观 Hot storage appearance	稀释稳定性 Dilution-stable	是否合格 Qualified or not
1	5∶2∶4	浑浊	–	–	–	–	否
2	5∶3∶5	浑浊	–	–	–	–	否
3	5∶4∶5	浑浊	–	–	–	–	否
4	7∶3∶4	浑浊	–	–	–	–	否
5	7∶3∶5	澄清透明	浑浊	–	–	–	否
6	8∶3∶5	澄清透明	浑浊	–	–	–	否
7	8∶3∶6	澄清透明	浑浊	–	–	–	否
8	9∶3∶5	澄清透明	烟状扩散	澄清透明	澄清透明	澄清透明	是
9	10∶3∶5	澄清透明	烟状扩散	澄清透名	澄清透明	沉淀	否
10	10∶4∶6	澄清透明	油丝状扩散	澄清透明	澄清透明	沉淀	否

表 5 不同水质对制剂稳定性的影响

Table 5 Effects of different water qualities on stability of preparation

水质 Water quality	乳液稳定性 Emulsion stability	起始外观 Initial appearance	冷贮外观 Cold storage appearance	热贮外观 Hot storage appearance	稀释稳定性 Dilution-stable	结果 Results
去离子水 Deionized water	烟状扩散	澄清透明	澄清透明	澄清透明	澄清透明	合格
自来水 Tap water	烟状扩散	澄清透明	澄清透明	澄清透明	澄清透明	合格
标准硬水(342 mg/L)Standard hard water	烟状扩散	澄清透明	澄清透明	澄清透明	澄清透明	合格
硬水(1 140 mg/L)Hard water	烟状扩散	澄清透明	澄清透明	澄清透明	澄清透明	合格

3 结论

通过对助溶剂、乳化剂进行筛选,确定该制剂最佳助溶剂为环己酮,最佳乳化剂组合 602#、CJ1 及 CJ2。10% 除尽微乳剂的最佳配方组成为:除尽 95% 原药 10% 、环己酮 15% 、苯甲醇 5% 、二甲苯 10% 、CJ1 5% 、CJ2 3% 、602# 9% ,水余量。

该配方制剂的起始外观、冷贮外观、热贮外观、稀释稳定性均达到要求。浓度 10% 除尽微乳剂以水为分散介质,不燃不爆,贮运安全,环境效益好,符合当今农药剂型发展方向,同时该制剂低毒低残留。因此,随着人们环境意识的增强、相关政策的出台、贸易壁垒、有机溶剂价格的持续飞速上扬,农药生产企业研制开发微乳剂农药的前景仍是非常乐观的。

参考文献

[1] 王军,许培援. 绿色农药剂型——微乳剂[J]. 精细与专用化学品,2004,21(12):12 – 13.

[2] 张春华,王忠伟. 微乳剂农药的发展概况及其优越性[J]. 农药研究,2003,18(1):19 – 20.

[3] 谢毅,吴学民. 现代农药剂型新进展[J]. 精细与专用化学品,2006,21(14):14 – 15.

[4] 孙家隆,范本荣. 微乳剂农药的发展概况及其优越性[J]. 山东农业科学,2000,21(14):7 – 8.

[5] 欧晓明, 黄明智, 王晓光,等. 昆虫抗性靶标部位及其在杀虫剂创制中的作用[J]. 现代农药, 2003,2(5):11 – 15.

[6] 徐尚成,蒋木庚. 虫螨腈的研究与开发进展[J]. 农药,2003,42(2):5 – 8.

[7] 印家厚. 新型杀虫剂除尽[J]. 农药市场信息,2001,6(6):7 – 8.

[8] 沈晋良. 农药加工与管理[M]. 北京:中国农业出版社,2002:6.

[9] 吴秀华. 农药微乳液物理稳定性的探讨[J]. 农药,1999,11(3):36 – 38.

[10] 郭武棣. 液体制剂[M]. 3 版. 北京:化学工业出版社,2003:11.

[11] 刘步林. 农药剂型加工技术[M]. 2 版. 北京:化学工业出版社,1998:3.

[12] SHABRIAR S. Fonulation of fine emulision by emulsification at high viscoaity or 1ow intialtension[J]. Colloid and Surfaa,2007,299:7 – 38.

[13] TMBE D E,SHANNA M M. Factors conlling the stability of colloid-stabed emulsions III. Measuremnent of the rheological Properties of colloid-laden interfaa[J]. Colloid Interfaee Sci,1995,171(2):456 – 462.

(上接第 14739 页)

[8] 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.

[9] 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.

[10] 陈安,陈宁,周龙骧,等. 数据挖掘技术及应用[M]. 北京:科学出版社,2006:3.

[11] 杨敬锋,薛月菊,胡月明,等. 基于关联规则和模糊判据的土地评价方法[J]. 农业工程学报,2008,24(5):74 – 78.

[12] 杨敬锋,薛月菊,胡月明,等. 基于精简模糊分类关联规则的分组模糊判决方法[J]. 系统工程理论与实践,2008,28(5):139 – 143.