

数据挖掘技术在药物配方中的研究及应用

Research and Application of Data Mining in medicine

吴 艳 (浙江工业大学之江学院信息分院 浙江杭州 310024)

摘要: 在新型的杀虫剂配方研制过程中,了解影响新药发挥功效的主要因素以及药物的副作用对整个研发过程来说是至关重要的。通过数据挖掘技术对相关数据进行分析和处理,提出了一种多维数据模型及计算推导方案,旨在揭示药物在杀虫过程中各个影响因素的变化规律,从而为合理利用杀虫剂进行最优化杀虫提供了理论依据。

关键词: 数据挖掘 数据立方体 多元线性回归方程 数据仓库

1 引言

数据挖掘,作为目前计算机领域三根强大的技术支柱之一,在越来越多的学科领域得到运用,被广泛应用于生物医学和 DNA 数据分析、金融数据分析、零售业顾客行为预测和市场趋势预测、电信业数据通信服务等^[1,2]。在药物制剂应用中,通常要对药物配方进行大量的实验,从大量的实验数据提取药物的知识性规律,获取药物的药理和使用方法。基于影响杀虫剂效果的因素的繁多,在此采用多维数据模型及推导方法,并结合数理统计分析工具,挖掘出杀虫剂的最佳配药方案。

2 研究内容

在配制杀虫剂的过程中,必须对药物的浓度、浸泡时间和温度对杀虫剂的功效有关。下面通过生物碱灭杀血吸虫^[3]为例,将各种考虑因素分成若干个参考实验值进行配制实验,实验数据如表 1 所示。

表 1 不同浓度、温度和时间杀虫结果

浓度 /mg/L	杀虫率%											
	24h			48m			72m			96m		
	20℃	25℃	30℃	20℃	25℃	30℃	20℃	25℃	30℃	20℃	50℃	30℃
10.00	12	32	46	38	90	96	60	100	100	82	100	100
5.00	10	18	38	26	80	94	32	100	100	64	100	100
2.50	8	12	16	14	48	86	16	100	100	40	100	100
1.25	4	12	16	10	32	76	14	26	100	40	68	100
0.625	2	8	10	8	30	48	12	16	92	16	36	98

从表 1 可以看出,当药物浓度越高,温度越高,且浸泡时间越长时效果越好。但在实际应用时,会出现一些问题:(1) 药物浓度越高,会对其他动植物产生不良影响,并且成本就越高;(2) 温度要求越高,对自然条件要求就越苛刻,且也会影响其他动植物的生长;(3) 延长药物浸泡时间可能会消耗人力和物力。因此,研究杀虫过程中各个因素的内部规律,探索合理的解决方案是必要的。回归分析不仅能给出整个方案的统计结果,而且创建的模型能匹配预测属性中的值,即能进行数据预测。

3 多因素函数模型的建立与分析

要研究上述实验各个因素的内部规律,最直观清晰的方法是描绘出效果与各个因素之间的函数轨迹。比较合理的解决途径是首先用客观兴趣度作为第一级过滤器,选出潜在的感兴趣的模式,然后再用主观兴趣度来对数据进行第二级筛选和特定方法的处理,得到真正感兴趣的知识。

3.1 数据的预处理

可以表 1 所列出的实验数据看作一个数据立方体。这个数据立方体包括三维:药物浓度(X1),药物浸泡温度(X2)和药物浸泡时间(X3),如图 1 所示。

为了获取这个立方体中潜在的知识,将这个立方体依据某种条件进行切片(slice)。例如,将药物浓度的每个实验值为条件对数据立方体进行切片操作,可以得到 5 个浓度立方体数据片,即:

浓度(X1) = 10.000, 杀虫效果与温度(X2)和浸泡

时间(X3)的子集 1

浓度(X1) = 5.000, 杀虫效果与温度(X2)和浸泡

时间(X3)的子集 2

浓度(X1) = 2.500, 杀虫效果与温度(X2)和浸泡

时间(X3)的子集 3

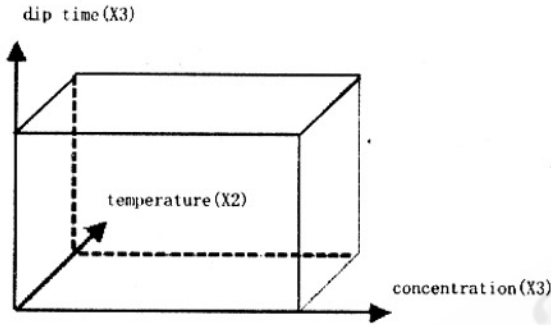


图 1 数据立方体

浓度(X1) = 1.250, 杀虫效果与温度(X2)和浸泡时间(X3)的子集 4

浓度(X1) = 0.625, 杀虫效果与温度(X2)和浸泡时间(X3)的子集 5

同时,将浸泡温度实验值为条件对数据立方体进行切片操作,可以得到 3 个温度立方体数据片:

温度(X2) = 20, 杀虫效果与浓度(X1)和浸泡时间(X3)的子集 6

温度(X2) = 25, 杀虫效果与浓度(X1)和浸泡时间(X3)的子集 7

温度(X2) = 30, 杀虫效果与浓度(X1)和浸泡时间(X3)的子集 8

最后,将浸泡时间的对数据立方体进行切片操作,可以得到 4 个时间立方体数据片:

浸泡时间(X3) = 24, 杀虫效果与浓度(X1)和温度(X2)的子集 9

浸泡时间(X3) = 48, 杀虫效果与浓度(X1)和温度(X2)的子集 10

浸泡时间(X3) = 72, 杀虫效果与浓度(X1)和温度(X2)的子集 11

浸泡时间(X3) = 96, 杀虫效果与浓度(X1)和温度(X2)的子集 12

3.2 数据挖掘过程和分析

数据挖掘的过程就是对不同的应用制定出挖掘知识的具体方法来获取知识的过程。由于数据立方体中的属性类型均为数型,可以将每个数据子集分别进行逐步线性回归处理,全部的工作由 matlab7.0 完成^[4]。

(1) 首先考虑对浓度立方体中每个数据子集的处理。可以得到 5 个有效的多元线性回归方程:

浓度(X1) = 10.000, $y = 3.750X_2 + 0.850X_3$ --
- 模型 1 $p \leq 0.001$

浓度(X1) = 5.000, $y = 5.000X_2 + 0.859X_3$ --
- 模型 2 $p \leq 0.001$

浓度(X1) = 2.500, $y = 5.600X_2 + 0.944X_3$ --
- 模型 3 $p \leq 0.001$

浓度(X1) = 1.250, $y = 5.600X_2 + 0.764X_3$ --
- 模型 4 $p \leq 0.001$

浓度(X1) = 0.625, $y = 5.250X_2 + 0.589X_3$ --
- 模型 5 $p \leq 0.001$

表 2 显示了上列 5 个方程的系数以及生成回归方程条件(浓度实验值)。

表 2 温度-时间回归系数对照表

Regression Model	concentration (X1)	temperature modulus ($\partial y / \partial X_2$)	time modulus ($\partial y / \partial X_3$)
1	10.000	3.750	0.850
2	5.000	5.000	0.859
3	2.500	5.600	0.944
4	1.250	5.600	0.764
5	0.625	5.250	0.589

从表 2 中可以看到,所有的回归系数都是正的,也就是说,只要回归方程中任意个自变量增大时,都能起到杀虫作用。当药物浓度由低到高变化时,回归方程中药物浸泡时间系数也有所递增,当浓度为 2.5mg/L 时,达到最高值 0.944,随后又逐渐变小,变化规律类似抛物线形状(如图 2)所示。同样,药物浸泡温度回归系数的变化也呈现出这样的规律,当药物浓度达到 2.5mg/L 和 1.5mg/L 时,最高值为 5.6。

图 2 画出了药物浓度(X1)与药物浸泡时间对杀虫效用($\partial y / \partial X_3$)之间大致的函数关系。根据表 2 中浸泡温度系数的变化规律,也可以得出与图 2 类似的药物浓度(X1)与药物浸泡温度对杀虫效用($\partial y / \partial X_2$)之间大致的函数关系。

(2) 考虑对温度立方体中每个数据子集的处理。

可以得到 3 个有效的多元线性回归方程:

温度 (X_2) = 20, $y = 4.473X_1 + 0.407X_3$ --- 模型 6 $p \leq 0.001$

温度 (X_2) = 25, $y = 7.378X_1 + 0.652X_3$ --- 模型 7 $p \leq 0.001$

温度 (X_2) = 30, $y = 5.113X_1 + 0.933X_3$ --- 模型 8 $p \leq 0.001$

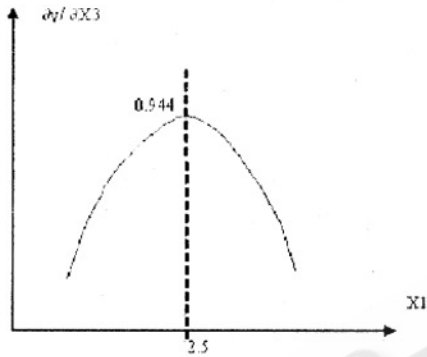


图 2 浓度和时间对杀虫效用的函数关系

根据这些回归方程,可以得到表 3 所示的温度 - 时间回归系数对照表:

表 3 温度 - 时间回归系数对照表

Regression Model	temperature (X_2)	deepness modulus (ay/aX_1)	time modulus (ay/aX_3)
6	20	4.473	0.407
7	25	7.378	0.652
8	30	5.113	0.933

从表 3 可以看出,当温度升高时,浓度回归系数呈现抛物线的变化规律。当温度达到 25°C 时,浓度系数达到最大。温度和浓度对杀虫率 (ay/aX_1) 的函数关系类似图 2 所示的变化规律。而温度和浸泡时间对杀虫率 (ay/aX_3) 则随着温度的升高单调递增。

(3) 考虑对时间立方体中每个数据子集的处理。可以得到 3 个有效的多元线性回归方程:

浸泡时间 (X_3) = 24, $y = 2.762X_1 + 1.343X_2$ --- 模型 9 $p \leq 0.001$

浸泡时间 (X_3) = 48, $y = 6.534X_1 + 4.629X_2$ --- 模型 10 $p \leq 0.001$

浸泡时间 (X_3) = 72, $y = 6.996X_1 + 5.686X_2$ --- 模型 11 $p \leq 0.001$

浸泡时间 (X_3) = 96, $y = 6.328X_1 + 5.200X_2$ ---

- 模型 12 $p \leq 0.001$

根据这些回归方程,又可以得到表 4 所示的浓度 - 温度回归系数对照表:

表 4 浓度 - 温度回归系数对照表

Regression Model	time (X_3)	concentration modulus (ay/aX_1)	temperature modulus (ay/aX_2)
9	24	2.762	1.343
10	48	6.534	4.629
11	72	6.996	5.686
12	96	6.328	5.200

表 4 则说明,当时间延长时,浓度回归系数和温度回归系数都呈现出抛物线的变化规律。当时间达到 72h 时,浓度系数和温度系数达到最大。时间和浓度对杀虫率 (ay/aX_1) 的函数关系和时间对浸泡温度的函数关系均类似图 2 所示的变化规律。

3.3 分析结果

根据上述实验数据的统计结果来看,所有的回归系数都是整数,即揭示了这样一个规律:当影响杀虫剂杀虫效果中的三个因素中的任意一个因素增大时,都会增大杀虫效果。但必须考虑杀虫剂的副作用。由于本例中杀虫剂是在常温下使用,药物撒入水中不可回收(时间可以延长),因此,只有药物浓度可以控制。

不难看出,表 2、表 3 和表 4 中的数据、图 2 中 ay/aX_3 和 X_1 的轨迹,类似的 ay/aX_2 和 X_1 、 ay/aX_1 和 X_2 、 ay/aX_3 和 X_2 、 ay/aX_1 和 X_3 、 ay/aX_2 和 X_3 的轨迹可以证明:当浓度达到 25mg/L,浸泡温度对杀虫的效果达到最高,浸泡时间对杀虫的效果也达到最高;当温度在 25°C 时,浓度对杀虫的效果达到最高,而杀虫效果也随着浸泡时间的延长而增加;当浸泡时间为 72h 时,浓度对杀虫的效果最好,浸泡温度对杀虫的效果也达到最大。并且,此条件下的药物副作用实验证明,满足环境保护要求。

4 软件的实现

(1) 数据预处理:将表 1 中实验数据进行整理并存入关系数据库中,利用 SQL 查询语句 SELECT...WHERE 将数据生成数据立方体的切片子集;(2) 数据统计处理:利用数学软件 MATLAB7.0 的数据分析工具生成对应于数据立方体中 12 个切片子集的多元回归线性方程,并用拟合工具绘制回归系数的变化规律;(3) 结果

分析:制定辨别规则,用 MATLAB7.0 进行编程方式实现智能分析,程序运行结果证明与人工分析结果一致。

5 结束语

在杀虫剂配方研制领域中,传统的做法是将数据信息存在数据库中,然后通过数据统计工具进行简单的分组对照和统计分析。但是这种做法存在着几个问题:(1)不能通过有效的数据模型组织和存储大量关联数据,因此,不能充分挖掘数据中隐藏的有用知识。(2)传统的数据库技术不能有效地建立和实现多维视图,因此,对数据的分析和处理不尽人意。(3)简单的统计分析最终的结论往往是人工完成,也不能根据已知数据进行预测。解决的方法是利用数据仓库和数据挖掘技术合理地组织和存取大量实验数据,并从中获取有用的、容易被忽视的信息或知识^[5,6]。

在研制新型杀虫剂过程中,一方面需要获取大量的实验数据,通过观察和数据建模将数据有效地组织起来,建立合理的多维视图实现对数据的分析和统计。另一方面,还需要注意杀虫剂所产生的副作用和对生态环境的破坏因素。因此,本研究有助于实现类似应

用的一般过程和方法,充分利用数据挖掘工具,为实践应用提供理论依据。

参考文献

- 1 Jiawei Han, Micheline Kamber. Data Mining: Concepts and Techniques. 1th ed., New York: Morgan Kaufmann Publishes, Inc. 2001. 70 - 100.
- 2 Luo Ke, Cai Biye. Research of Data Mining and Its Development. Computer Engineering and Application. 2002, (14): 182 - 185.
- 3 黄琼瑶、刘年猛、彭飞、杨华中、卢岳华、胡琦、冯芳、孙慧、刘铭,血本草生物碱杀灭钉螺及日本血吸虫尾灭蚴的实验研究,热带病与寄生虫学,2003, 1 (4): 216 - 219.
- 4 王正林、刘明,精通 MATLAB7,北京:电子工业出版社. 2007. 184 - 203.
- 5 E. F. Codd, S. Codd, and CT. Salley. Beyond decision support. Computer World, 1993, 27 (30): 223 - 227.
- 6 杨泽民、陈莉、范全润,加权关联规则的并行挖掘算