Apriori 算法在农业病虫害分析中的应用

黄光明 (柳州职业技术学院, 广西柳州 545006)

摘要　介绍了Apriori 算法进行规则挖掘的方法,并以某地区越冬代二化螟规则分析为实例介绍Apriori 算法的应用。结果表明,Apriori 算法在农业病虫害数据分析中具有良好的应用前景。

关键词　Apriori 算法; 关联规则; 病虫害; 数据挖掘
中图分类号　S11 +9　　文献标识码　A　　文章编号　0517 - 6611(2009)13 - 06028 - 02

Application of Apriori Algorithm in Analysis of Agricultural Disease and Insect Pest

HUANG Guang- ming　(Liuzhou Vocational and Technical College , Liuzhou , Guangxi 545006)

Abstract　The method of using Apriori algorithm mine association rules was introduced in this essay , and the application of Apriori algorithm was explained by taking rule analysis of overwintering Chilo suppressalis (Walker) as an example .The results showed that Apriori algorithm had the good application prospect in data analysis of agricultural disease and insect pest .

Key words　Apriori algorithm; Association rule ; Agricultural disease and insect pest ; Data mining

病虫害直接危及农作物的产量和质量,我国每年因病虫害造成的经济损失达15 %～25 %[1] 。因此,分析及发现农业病虫害规则对防治病虫害具有重大的意义。关联规则挖掘是近年来数据挖掘领域的一个研究热点,它用于发现数据中项集之间隐含的联系,通过关联规则形式表示。Apriori 算法是最有影响的挖掘关联规则的算法之一,该算法目前已应用于商业、林业、电信和教育等方面[2 - 5] ,但未见在农业上的应用。因此,笔者将Apriori 算法用于农作物病虫害规则分析,以期为农作物病虫害分析开辟一条新途径。

1　Apriori 算法

1.1　相关概念　设$I = \{i_1, i_2, \cdots, i_m\}$是项集合,$T = \{t_1, t_2, \cdots, t_n\}$是事务集合,其中$P\ t_i \subseteq I (1 \leq i \leq n)$。$A \Rightarrow B$称为T中的关联规则,其中$A < I$、$B < I$、$A \cap B = \varnothing$。

在事务集合T中,包含$A \cup B$的事务占全部事务的百分比称为T中关联规则$A \Rightarrow B$的支持度,记为$support(A \Rightarrow B) = P(A \cup B)$。

在事务集合T中,包含$A \cup B$的事务占包含A的事务的百分比称为T中关联规则$A \Rightarrow B$的置信度,记为$confidence(A \Rightarrow B) = P(B|A)$。

设min_sup是最小支持度阈值,min_conf是最小置信度阈值。如果事务集合T中的关联规则$A \Rightarrow B$同时满足

$support(A \Rightarrow B) \geq min\_sup$

$confidence(A \Rightarrow B) \geq min\_conf$

则$A \Rightarrow B$称为T中的强关联规则。

包含k个项的集合称为k-项集。如果项集满足最小支持度,则称它为频繁项集(Frequent itemset)。频繁k-项集的集合通常记作L_k。

1.2　Apriori 算法描述　Apriori 算法是R.Agrawal 和R.Srikant 在1994 年提出的关联规则挖掘算法。为了减少候选数据项集的数目,Apriori 算法使用逐层搜索的迭代方法:k-项集用于搜索$(k+1)$-项集。首先扫描事务集合,找出频繁1-项集集合L_1,基于L_1找出频繁2-项集集合L_2,而基于L_2找出L_3,以此类推,直至不能找到频繁项集为止。每一次搜索都要扫描1 次事务集合,为提高频繁项集逐层产生的效率,Apriori 算法利用一个重要性质:频繁项集的所有非空子集都必须是频繁的,这个性质被称为Apriori 性质。

Apriori 算法由以下两部分组成:

(1) 使用候选项集找出频繁项集。利用频繁k-项集L_k找出频繁$(k+1)$-项集L_{k+1},是Apriori 算法的核心,具体过程可以描述为:第一步基于频繁k-项集L_k,采用自连接方法产生所有可能频繁的$(k+1)$-项集,即候选$(k+1)$-项集C_{k+1}。第二步扫描1 次事务集合,统计C_{k+1}中每个候选的支持计数,与最小支持计数相比,形成频繁$(k+1)$-项集L_{k+1}。

(2) 由频繁项集产生关联规则。具体过程可以描述为:第一步对于每个频繁项集l,产生l 的所有非空子集。第二步对于l 的每个非空真子集l_u,如果l 的支持计数除以l_u的支持计数大于等于最小置信度阈值min_conf,则输出规则$l_u \Rightarrow (l - l_u)$。

表1　历史数据

Table 1　The historical data

序号 No.	1 月份平均气温/℃ Average temperature in January	1 月份降水量 mm Rainfall in January	4 月份平均气温/℃ Average temperature in April	4 月份降水量 mm Rainfall in April	越冬代二化螟蛾量/头 Sum of overwintering Chilo suppressalis (Walker) moth
1	1.4	31.4	16.1	86.3	99
2	- 1.0	32.3	17.6	150.2	132
3	- 0.8	73.0	14.3	170.2	232
4	1.4	27.0	15.0	230.0	256
5	- 0.9	97.4	14.2	307.1	284
6	1.4	31.1	15.9	234.8	137
7	4.0	29.3	15.9	72.2	38
8	2.7	27.3	17.7	12.1	198
9	3.8	3.5	16.2	94.0	102
10	- 0.1	16.5	16.2	150.0	410
11	- 0.5	20.9	17.2	87.0	130
12	3.4	2.09	17.0	83.0	128
13	2.2	1.5	17.6	127.7	223
14	1.4	32.5	14.5	232.0	312
15	0.2	80.5	14.6	250.0	267

2　应用实例

二化螟是水稻的主要虫害之一,它的发生量与自然因素有着密切的关系。由于越冬代所处环境变化很大,其发生发展特点也不同[6] ,因此对各地越冬代二化螟发生规则进行

作者简介　黄光明(1966 -),男,广西扶绥人,硕士,讲师,从事数据仓库、数据挖掘的教学与研究。

收稿日期　2009-02-23

挖掘具有重要意义。收集我国某地区的15 年越冬代二化螟相关数据,通过采用主因子分析、回归分析等方法,选取相关程度密切的4 个因子[7]:1 月份平均气温(℃)、1 月份降水量(mm),4 月份平均气温(℃)、4 月份降水量(mm)。具体数据见表1。

2.1　数据预处理　由于所有的分析数据均为非离散的数值属性,因此应进行离散化处理。各因子离散化等级见表2。预处理后的数据如表3 所示。

表2　各因子离散化等级

Table 2　The discretization grode of each factor

等级 Level	1 月份平均气温(A)/℃ Average temperature in January	1 月份降水量(B)/mm Rainfall in January	4 月份平均气温(C)/℃ Average temperature in April	4 月份降水量(D)/mm Rainfall in April	越冬代二化螟蛾量/头 Sum of overwintering Chilo suppressalis (Walker) moth
1	<0.1	<25.0	<14.8	<100.0	<130
2	0.1～1.5	25.0～40.0	14.8～15.5	100.0～160.0	131～224
3	1.6～3.0	40.1～75.0	15.6～16.0	160.1～250.0	225～317
4	>3.0	>75.0	>16.0	>250.0	>317

表3　预处理后的数据

Table 3　The pretreated data

序号 No.	1 月份平均气温/℃ Average temperature in January	1 月份降水量/mm Rainfall in January	4 月份平均气温/℃ Average temperature in April	4 月份降水量/mm Rainfall in April	越冬代二化螟蛾量/头 Sum of overwintering Chilo suppressalis (Walker) moth
1	A2	B2	C4	D1	E1
2	A1	B2	C4	D2	E2
3	A1	B3	C1	D3	E3
4	A2	B2	C2	D3	E3
5	A1	B4	C1	D4	E3
6	A2	B2	C3	D3	E2
7	A4	B2	C3	D1	E1
8	A3	B2	C4	D1	E2
9	A4	B1	C4	D1	E1
10	A1	B1	C4	D2	E4
11	A1	B1	C4	D1	E1
12	A4	B1	C4	D1	E1
13	A3	B1	C4	D2	E2
14	A2	B2	C1	D3	E3
15	A2	B4	C1	D3	E3

表4　关联规则挖掘结果

Table 4　The results of association rule mining

关联规则 Association rule	支持度/% Support degree	置信度/% Confidence degree	关联规则 Association rule	支持度/% Support degree	置信度/% Confidence degree
A1]E3	13.3	40.0	C1D3]E3	20.0	100
A2]E3	20.0	60.0	C1D4]E3	6.7	100
A3]E2	13.3	100	C2D3]E3	6.7	100
A4]E1	20.0	100	C3D1]E1	6.7	100
B1]E1	20.0	60.0	C3D3]E2	6.7	100
B2]E2	20.0	42.9	C4D1]E1	26.7	80.0
B3]E3	6.7	100	C4D2]E2	13.3	66.7
B4]E3	13.3	100	B1D1]E1	20.0	100
C1]E3	26.7	100	B1D2]E2	6.7	50.0
C2]E3	6.7	100	B1D2]E4	6.7	50.0
C3]E2	6.7	50.0	B2D1]E1	13.3	66.7
C3]E1	6.7	50.0	B2D2]E2	6.7	100
C4]E1	26.7	50.0	B2D3]E3	13.3	66.7
D1]E1	33.3	83.3	B3D3]E3	6.7	100
D2]E2	13.3	66.7	B4D3]E3	6.7	100
D3]E3	26.7	80.0	B4D4]E3	6.7	100
D4]E3	6.7	100			

2.2　关联规则挖掘　设 min_sup = 5%, min_conf = 40%。根据Apriori 算法编写程序,搜索原始数据表,得到满足最小支持度和最小可信度的关联规则。表4 是其中一部分,这些关联规则右边为越冬代二化螟蛾量,左边是1 月份平均温度、1 月份降水量、4 月份平均温度、4 月份降水量中某1 个或2 个属性值。

2.3　关联规则分析　通过表4 中的关联规则,可以得到以下结论:

(1) 在该地区,越冬代二化螟发蛾量与气候因子密切相关,在选择的4 个气候因子中,影响发蛾量的顺序为:4 月份降水量>4 月份平均气温>1 月份降水量>1 月份平均气温。

(2) 虫害与降水量关系。降水量较少时虫害级别较低,随着降水量逐渐增多,发生虫害的级别也逐渐增高,当4 月份降水量大于250.0 mm、1 月份降水量大于75.0 mm 时,虫害级别不再增加,一般维持在3 级;在1 月份降水量相同的情况下,虫害级别由4 月份降水量决定。

(3) 虫害与平均气温关系。平均气温低时虫害级别较高,平均气温高时虫害级别较低;在平均气温相同情况下,虫害级别由降水量决定。

(4) 该地区4 月份出现降水量大于250.0 mm 的情况较少,出现小于100.0 mm、100.0～160.0 和160.1～250.0 mm 3 种情况比较均等,也就是说出现1 级、2 级和3 级虫害的情况比较均等。

3　小结

Apriori 算法通过分析事物之间的相互依赖关系,能发现和提取隐藏在数据背后的有效知识,有助于人们认识和理解其中存在着的客观规律,具有很大的实用价值。笔者使用Apriori 算法对某地区越冬代二化螟进行规则分析,取得了较好的研究效果。

参考文献

[1] 刘乃森,刘福霞.人工神经网络及其在植物保护中的应用[J].安徽农业科学,2006,34(23):6237-6238.

[2] 赵春玲,宁红云.Apriori 算法的改进及其在物流信息挖掘中的应用[J].天津理工大学学报,2007,23(1):30-33.

[3] 黄世国,林思祖,林大辉.Apriori 算法在杉木伴生树种选择中的应用[J].福建农林大学学报:自然科学版,2008,37(1):70-72.

[4] 武丽芬,严学勇.基于Apriori 算法的数据挖掘在电信套餐制定中的应用[J].晋中学院学报,2007,24(3):98-101.

[5] 姜红艳.Apriori 关联算法在学生成绩中的应用[J].鞍山师范学院学报,2007,9(2):48-50.

[6] 周树基.越冬代二化螟的生物学特性观察[J].湖南农学院学报,1989,15(3):121-122.

[7] 程新意,杨崇瑞.用模糊分析方法预报越冬代二化螟的发生量[J].安徽农学院学报,1992,19(3):308-312.