一种基于粗糙集的诊断型知识发现方法*

李爱平[1+], 隋品波[1], 贾 焰[1], 廖桂平[2], 吴泉源[1]

[1](国防科学技术大学 计算机学院,湖南 长沙 410073)

[2](湖南农业大学 计算机与信息科学系,湖南 长沙 410073)

A Knowledge Acquisition Method Based on Rough Set in Diagnostic Expert System

LI Ai-Ping[1+], SUI Pin-Bo[1], JIA Yan[1], LIAO Gui-Ping[2], WU Quan-Yuan[1]

[1](School of Computer Science, National University of Defense Technology, Changsha 410073, China)

[2](Department of Computer and Information Science, Hu'nan Agriculture University, Changsha 410073, China)

+ Corresponding author: Phn: +86-731-4575979, Fax: +86-731-4512504, E-mail: apli@x263.net, http://www.nudt.edu.cn

Received 2003-05-10; Accepted 2004-07-16

Li AP, Sui PB, Jia Y, Liao GP, Wu QY. A knowledge acquisition method based on rough set in diagnostic expert system. *Journal of Software*, 2004,15(Suppl.):13~19.

Abstract: Knowledge acquisition is the bottleneck in developing expert system. It usually takes a long period to acquire disease knowledge using the traditional methods. Aiming at this problem, the paper presents the relationship between rough sets and rule-based rapeseed disease knowledge, namely the application of rough sets in knowledge acquisition from rapeseed disease expert. Then the exclusive rules, inclusive rules and disease images of rapeseed disease are built based on the RHINOS diagnosis model, and the definition of probability rule is put forward. At last, the paper presents the rule-based automated induction reasoning method, including exhaustive search, post-processing procedure, estimation for statistic test and the bootstrap and resampling methods. The results of experiment show that rough sets not only are a good framework for knowledge acquirement, but also can accurately induct the rules of plant diseases. This method can act as the assistant tool for development of diagnosis expert system, and has a extensive application in intelligent agriculture information systems.

Key words: expert system; rough set; knowledge acquisition; knowledge representation; diagnostic knowledge

摘 要: 知识获取是开发专家系统的瓶颈,传统的病害知识获取通常需要一个较长的过程.针对这一问题,本文给出了粗糙集理论和基于规则的作物病害知识之间的关系,即在油菜植病专家知识获取过程中的应用,建立了基于RHINOS 诊断模型的油菜病害的排除规则、包含规则和病害映像,阐述了可能性规则定义,并给出了基于规则描述的自动归纳推理方法,包括全搜索过程、后处理过程、统计测试的评估以及交叉验证和鞋带方法等.实验结果表明,

* Supported by the National High-Tech Research and Development Plan of China under Grant Nos.2001AA113020; 2003AA118010 (国家高技术研究发展计划(863))

作者简介: 李爱平(1974－),男,山东诸城人,博士,讲师,主要研究领域为人工智能,分布对象计算和软件工程;隋品波(1967－),女,硕士,讲师,主要研究领域为数据库系统实现技术,数据挖掘,分布对象计算;贾焰(1960－),女,博士,教授,博士生导师,主要研究领域为数据库技术,分布对象计算,人工智能;廖桂平(1964－),男,博士后,副教授,主要研究领域为人工智能,农业信息化,数据库技术;吴泉源(1942－),男,教授,博士生导师,主要研究领域为人工智能,分布对象计算.

粗糙集不仅是一个很好的知识获取的框架,而且能正确的归纳推理植物病害的规则.这对诊断型专家系统的开发可起到一个很好的辅助作用,在智能化农业信息系统中有着广泛的应用前景.

关键词: 专家系统;粗糙集;知识获取;知识表示;诊断型知识

诊断是鉴定病害的过程.Fry 总结 Robert 和 Smith 的研究成果[1],提出了诊断的通用过程:"观察问题→提出假设并解释观察结果→检验假设→接受或重新检验假设".在数学上,这类似于命题(规则)和集合(案例)的关系.具体地说,(1)观察:即检查在病例中观察到的症状.(2)假设:即特征提取,从得到的观察结果和掌握的领域知识对该症状进行解释.(3)检验:即区分:收集具有相似症状病害的观察信息,并检测在相似症状中的不同.

显然,集合理论是医疗诊断推理过程的理论基础[2].在第一步,医学专家会检测满足症状 R 的临床病历是否至少包含在一个得了疾病 d 的病人的病历中.第二步,收集几乎覆盖了所有患了疾病 d 的病人的症状集合{R}.最后,医生会去发现一个高度支持区别于其它类似疾病的疾病 d 的症状{R}的一个集合,即,{R}是高度精确和高度覆盖的.

在本文中,首先讨论在粗糙集理论和基于规则描述的病害类疾病之间的关系,包括特征化规则,区分规则和搜集观察症状,然后提出利用粗糙集从病历数据库中进行归纳学习的算法.在讨论过程中,我们以在实际过程中开发的油菜病害诊断专家系统中的油菜病害数据库为例来说明本文提出的知识归纳方法.实验结果表明,粗糙集不仅是一个表示不确定知识获取过程的很好的框架,也是归纳疾病表述的极好工具.

1 诊断模型

Matsumura[3]提出的一个 RHINOS 诊断模型,它主要包括以下三个推理过程:排除推理、包含推理和并发推理,分别对应于特征提取、区分和观察.当植株没有一种病害必需具备的症状时,排除推理便从侯选集中排除该病害.当植株具备某个病害特有的症状时,包含推理便推理出植株感有该病害.最后,根据包含推理推出的病害不能解释的症状,并发推理便会推断出并发的其它病害.基于该诊断模型,从油菜植病专家那儿得到了有关油菜病害的排除规则、包含规则和病害描述.Tsumoto 基于该模型利用粗糙集理论在医学诊断领域给出了一个知识归纳算法,本文借鉴了文献[4,5,6]中的某些思想,如覆盖指数和误差验证方法等.

1.1 排除规则

排除规则与排除推理相关,即排除规则的前提与一个诊断的必要前提条件等价.从与油菜植病专家的讨论中[7],选择油菜病害诊断最小的互不依赖的 6 个基本属性来定义必要的前提:1. 生育期, 2. 症状分布, 3. 病原物, 4. 病症进程, 5. 病斑, 6. 病状进程. 如关于油菜菌核病的排除规则定义:

为了诊断油菜是否感有菌核病, 必需要有下述症状:生育期:幼苗期较少, 为害部位:不是根部, 病原物:白色絮状菌丝, 病症进程:黑色菌核, 病斑:褐色, 病状进程:变白色, 组织腐烂.

选择上述 6 个属性的原因是为了使排除推理更加简洁,在尽量不丢失信息的前提下,只选择最小的需求.显然,这种选择可以看作是对给定属性的一种定制,而且这种定制可以从油菜病害数据库中导出.因此,希望将利用全部给出的症状归纳导出排除规则的过程形式化.这是因为在所有的排除规则都被归纳出后,得到描述排除规则的最小需求属性.

1.2 包含规则

包含规则含有数个规则,称作正(positive)规则.正规则的前提由被包含病害的具体症状值的集合组成.若某植株满足其中一个集合,则可以一定的可能性判断该植株感有某病害.这种针对每种病害的规则可通过咨询植病专家而获得,其过程如下:

1. 怀疑植株患一种病害时的症状集合.
2. 患该病害植株具有这些症状集合的可能性:SI(Satisfactory Index, 满足指数).
3. 满足该症状集合感病的植株数与所有感该病的植株总数的集合的比率:CI(Covering Index, 覆盖指数).
4. 如果得到的 CI 的总和(tCI)等于 1,便结束,否则,到下一步.
5. 针对那些感了该病却不满足已收集的症状集合的植株,转到第 1 步.

可见,一个正规则是由一个症状的集合来描述,它的满足指数(SI)对应于精确性测度(*accuracy measures*),覆盖指数(CI)与全部为正率(*total positive rate*)相关.值得注意的是,SI 和 CI 一般由油菜植病专家经验给出.

形式地,每一个正规则被表示为一个四元序偶:$\langle\rangle<d,R_i,SI_i,CI_i>$,其中 d 表示它的结论,R_i 表示它的前提,SI_i 表示它的精确度.因此,每一个包含规则被表示为:$\langle\{\langle d,R_1,SI_1(,CI_1)\rangle,\ldots,\langle d,R_k,SI_k(,CI_k)\rangle\},tCI\rangle$,其中整体 CI(tCI)被定义为一整体规则的 CI,由所有规则的或连接来组成,即:$R_1 \vee R_2 \vee \cdots \vee R_k$.

如针对油菜菌核病的一个包含规则(tCI=0.9)由 3 个正规则组成,其中的 1 个为:

如果 为害部位:不是根部,病原物:白色絮状菌丝,病症进程:黑色菌核,病斑:浅褐色水渍状,病状进程:变白色,组织腐烂.

则 油菜感有菌核病的可能性为 0.9(SI=0.9),并且该规则覆盖了所有病例中的 75%(CI=0.75).

……

在上述规则中,tCI 显示了 3 个正规则的或型运算覆盖了所有案例的 90%,也就意味着 10%的油菜菌核病不能被上述规则来诊断.

1.3 病害描述

在得到了某种病害的所有可能的症状之后,该种规则用来检测与该种病害的并发症.利用这种规则,查找所有的不能被推理出来的病害症状,这些症状反映了其它病害的并发症.例如:菌核病可以解释如下的症状:病害部位:不是根部,或者病原物:白色絮状菌丝,或者 病症进程:黑色菌核,....因此,当油菜植株病部病原物为白色絮状菌丝时,便可以猜想油菜同时也感了其它病害,如黑斑病、黑胫病等.

2 粗糙集和知识规则表示

2.1 粗糙集理论

粗糙集理论基于不可辨识关系表示某些特点、过程、对象等知识.该理论以观察和测量所得的数据进行分类的能力为基础,认为知识是基于对对象分类的能力,知识直接与真实或抽象世界有关的不同分类模式联系在一起;它阐明了集合理论中关于属性组合模式的分类特征和能力,因而可以用来为了某个分类而获取某些属性的集合以及评估在数据库中属性的组合可分类数据的精确性能[7].该理论以观察和测量所得的数据进行分类的能力为基础,它认为知识是基于对对象分类的能力,知识直接与真实或抽象世界有关的不同分类模式联系在一起[8].

以表 1 为例,首先考察属性"症状分布"是如何对油菜病害集合进行分类的.集合中"症状分布"属性等于"叶片"的是{2,3}(用记录号表示记录).该集合表明仅利用限制 R = [症状分布=叶片]不能区分{2,3}.该集合被在关系 R 上的不可分辨集合定义,记为$[x]_R$ = {2,3}.其中,{2}为菌核病,而{3}为黑斑病.因此,需要其它附加属性来区分"菌核病"和"黑斑病".利用此概念,可以评估每一个属性的分类能力.例如,"病症进程=黑色菌核"是针对"菌核病"的.当然,也可以将不可区分关系扩展到多元属性的情况,例如:$[x]_{[症状分布=叶片]\wedge[病症进程=黑色菌核]}$ = {2},$[x]_{[症状分布=叶片]\vee[病症进程=黑色菌核]}$ = {1,2,3,4,6}.在粗糙集理论框架中,集合{2}称作被前者的与连接严格定义,或者称作被后者的或连接粗糙定义.因此,训练样本 D 的分类可以看作是寻找最优的被关系 R 支持的集合$[x]_R$.为此,可在集合理论的框架中定义分类的特征,如精确性测度(SI)和覆盖指数或全部为正率(CI)便可定义为:

$$\alpha_R(D)=\frac{|[x]_R \cap D|}{|[x]_R|},\quad \kappa_R(D)=\frac{|[x]_R \cap D|}{|D|} \tag{1}$$

式中,$|D|$为 D 的基数,$\alpha_R(D)$ 为 R 关于 D 的分类的精度 $SI(R,D)$,$\kappa_R(D)$ 为 R 关于 D 的覆盖指数或全部为正率 $CI(R,D)$.

表1 油菜病害数据库摘选

记录号	生育期	症状分布	病原物	病症进程	病斑	病状进程	……	病害分类
1	苗期	根颈或叶柄	白色絮状菌丝	黑色菌核	红褐色斑点	变白色,组织腐烂	……	菌核病
2	成株期	叶片	白色絮状菌丝	黑色菌核	暗青色轮层	变白色,单个病斑扩展	……	菌核病
3	成株期	叶片	白色絮状菌丝	黑褐色霉丛	暗褐色轮纹	黑褐色,单个病斑不扩展	……	黑斑病
4	成株期	茎枝	白色絮状菌丝	黑色菌核	浅褐色水渍状	变白色,不扩至根部	……	菌核病
5	成株期	茎枝	白色絮状菌丝	黑色粒状物	浅褐色无定形	变白色,扩至根部	……	黑胫病
6	成株期	花或角果	白色絮状菌丝	黑色菌核	浅褐色水渍状	变白色,组织腐烂	……	菌核病

注:本文中的实例计算结果均相对于该表而言.

2.2 可能性规则定义

为了描述3种类型的诊断规则,首先利用粗糙集理论定义可能性规则.其中属性-值对可以通过等价关系 R_f 来表示.

定义1(等价关系). 令 U 为域,V 为属性取值的集合.从 U 到 V 的函数 f 的全体称作属性的赋值函数.对于任意 $u,v\in U$,定义等价关系 R_f,$u\equiv_{R_f} v$ 当且仅当 $f(u)=f(v)$.

例如,[生育期 = 苗期] & [症状分布 = 根颈或叶柄]是一个等价关系,记作 R_f=[生育期=苗期] & [症状分布=根颈或叶柄].满足 R_f 的等价类记作 $[x]_{R_f}$.例如,{2,3,4,5,6}为满足[生育期=成株期]的记录的集合,$[x]_{[生育期=成株期]}$ 即为{2,3,4,5,6}.域 U 为全体训练样本,记录的集合.

定义2(可能性规则). 令 R_f 为满足某一赋值函数 f 的等价关系,D 为元素属于类 d 的集合,或在全部训练样本(域 U)中的为正的记录.最后,令 $|D|$ 表示 D 的基数.D 的可能性规则定义为一个四元序偶,$<R_f \xrightarrow{\alpha,\kappa} d,\alpha_{R_f}(D),\kappa_{R_f}(D)>$,其中 $R_f \xrightarrow{\alpha,\kappa} d$ 满足如下条件:

$$(1)\quad [x]_{R_f}\cap D\neq\phi,$$

$$(2)\quad \alpha_{R_f}(D)=\frac{|[x]_{R_f}\cap D|}{|[x]_{R_f}|}, \tag{2}$$

$$(3)\quad \kappa_{R_f}(D)=\frac{|[x]_{R_f}\cap D|}{|D|}.$$

定义中的 α 对应于精确性测度:当一规则 α 等于0.9时,则该规则的精确度也等于0.9.另一方面,κ 是 D 中有多少比例被该规则覆盖的一个统计测度,即覆盖指数或全部为正率:当 κ 等于0.5时,表示该类有一半属于成员满足该等价关系的集合.

例如,规则'[病原物 = 白色絮状菌丝]→菌核病'.因为 $[x]_{[病原物=白色絮状菌丝]}$={1,2,3,4,5,6}和 D = {1,2,4,6},$\alpha_{[病原物=白色絮状菌丝]}(D)$=|{1,2,4,6}| / |{1,2,3,4,5,6}| = 0.67 且 $\kappa_{[病原物=白色絮状菌丝]}(D)$=|{1,2,4,6}| / |{1,2,4,6}| = 1.所以,如果油菜植株病部病原物为白色絮状菌丝,则猜测是菌核病的可能性为67%,并且该规则覆盖了这种病例的100%的情况(相对表1而言).

2.3 诊断规则

2.3.1 排除规则

$$R\xrightarrow{\alpha,\kappa} d \quad \text{s.t.} \quad R=\wedge_i R_i \text{ and } \kappa_{R_i}(D)=1.0 \tag{3}$$

上例中,"菌核病"的关系 R 描述为:([生育期 = 苗期] ∨ [生育期 = 成株期])∧([症状分布 = 根颈或叶柄] ∨ [症状分布 = 叶片] ∨ [症状分布 = 茎枝] ∨ [症状分布 = 花或角果])∧ [病原物 = 白色絮状菌丝] ∧[病症进程 = 黑色菌核] ∧ ([病斑 = 红褐色斑点] ∨ [病斑 =暗青色轮层] ∨ [病斑 =浅褐色水渍状])∧ ([病状进程 = 变白色,组织腐烂] ∨ [病状进程 = 变白色,单个病斑扩展] ∨ [病状进程 = 变白色,不扩至根部]).严格地讲,该命题为:$d\rightarrow R$.选择上述记号是为了便于与其它两种规则形式上的对应与比较.

2.3.2 包含规则

$$R \xrightarrow{\alpha,\kappa} d \quad \text{s.t.} \quad R=\vee_i R_i=\vee\wedge_j\vee_k[a_j=v_k],\ \alpha_{R_i}(D)>\delta_\alpha,\ \text{and}\ \kappa_{R_i}(D)>\delta_\alpha \tag{4}$$

上例中,最简单的"菌核病"的关系 R 描述为:[病原物 = 白色絮状菌丝] ∨ [病症进程 = 黑色菌核] ∨ [病斑 =浅褐色水渍状] ∨ [病状进程 = 变白色,组织腐烂].但是,包含规则的归纳给我们带来了两个问题.第一,SI 和 CI 要根据训练样本进行过度的学习.第二,上述的规则仅仅是从训练样本中得到的大量规则中的一条.因此,需要依照一定的参考阈值来从初始的归纳出来的规则进行筛选.这一问题将在下面部分讨论.

2.3.3 病害描述

$$R \xrightarrow{\alpha,\kappa} d' \quad \text{s.t.} \quad R=\vee R_i\vee[a_i=v_j],\ \alpha_{R_i}(D)>0(\kappa_{R_i}(D)>0) \tag{5}$$

上例中,"菌核病"的关系 R 描述为:[生育期 = 苗期] ∨ [生育期 = 成株期] ∨ [症状分布 = 根颈或叶柄] ∨ [症状分布 = 叶片] ∨ [症状分布 = 茎枝] ∨ [症状分布 = 花或角果] ∨ [病原物 = 白色絮状菌丝] ∨ [病症进程 = 黑色菌核] ∨ [病斑 = 红褐色斑点] ∨ [病斑 =暗青色轮层] ∨ [病斑 = 浅褐色水渍状] ∨ [病状进程 = 变白色,组织腐烂] ∨ [病状进程 = 变白色,单个病斑扩展] ∨ [病状进程 = 变白色,不扩至根部].值得注意的是,与简单诊断模型相比,$\kappa_R(D)$ 在定义中起了很重要的作用.

3 基于归纳推理的知识发现过程

一个基于 RHINOS 诊断模型的归纳算法包括两个过程.第 1 步是全部的搜索过程,以从所有的属性-值对中归纳出排除规则和疾病描述.第二步是后处理过程,通过合并所有的属性-值对来归纳包含规则.

3.1 全搜索过程

令 D 表示目标类 d(或者叫正例)的训练样本.该搜索过程如图 1 所定义.在上例中,令 d 表示"菌核病",并且选择 $[a_i=v_j]$ 为[生育期 = 成株期].既然集合的交 $[x]_{[病原物=白色絮状菌丝]}\cap D(=\{1,2,4,6\})$ 不等于 ϕ.该属性-值对便被归纳到病害映像中.但是,既然 $\alpha_{[病原物=白色絮状菌丝]}(D)=0.67$,该对便不能归纳到包含规则中.最后,因为 $D\subset[x]_{[病原物=白色絮状菌丝]}(=\{1,2,3,4,5,6\})$,该对也被归纳到排除规则中.

当所有的属性-值对测试完毕后,不仅可以得出所有的排除规则、病害映像,而且也得到了所有的包含规则的候选集.

3.2 后处理过程

由于包含规则的定义是一个比较弱的定义,因此可以得到很多包含规则.如,等价关系[生育期 = 苗期]满足 $[x]_{[生育期\ =\ 苗期]}\cap D\neq\phi$,是关于"菌核病"的一个包含规则,尽管它的 SI 值只等于 1/4.为了压缩对这种仅有很弱分类能力规则的归纳,我们限制只有 SI 值大于 0.5 的等价关系才被选出.例如,上例中的等价关系[生育期 = 苗期]小于该精度,它便被从归纳规则的侯选集中删除.进一步,PRI-REX2 根据不具备分类能力的属性的预先删除将属性的数目进行了最小化,称为依赖属性.该过程如图 2 所示.

3.3 统计测度的评估

统计测度的定义显示只有很少的训练样本会导致不恰当的评估.上例中,所有的测度都是 1.0,这便意味着该规则能够正确地诊断并覆盖了"菌核病"的全部病例.但是,实际上,这仅仅在具有很少样本的情况下才有效.此时,精确度和覆盖是有偏差的.因此,需要引进其它的评估方法来纠正这些偏差.

注意到该问题与多元分析中的差错率判别函数[7]有些类似,在该领域中重新抽样方法被证明是对偏差估计是有效的.因此,重抽样方法被利用到精确度和覆盖指数的评估中.

```
procedure Exhaustive Search;
  var
     L : List; /* A list of elementary relations */
  begin
     L := P0; /*P0: A list of elementary relations */
     while (L ≠ φ) do
       begin
          Select one pair [ai = vj ] from L;
          if ([x] [ai = vj ]∩D ≠ φ) then do
                    /* D: a set if positive examples */
            begin
              R di := R di ∨ [ai = vj ];
              /* Disease Image */
              if (κ[ai = vj ] (D) > δκ)
                then Lir := Lir +[ai = vj ];
                   /* Candidates for Inclusive Rules */
              if (κ[ai = vj ] (D)= 1.0)
                then Rer := Rer ∧ [ai = vj ];
                /* Exclusive Rule */
            end
          L := L − [ai = vj ];
       end
  end {Exhaustive Search};
```

图1 全搜索过程算法

```
procedure Postprocessing Procedure;
  var
                                        i i: integer;
M;Li : List;
  begin
    L1 := Lir; /* Candidates for Inclusive Rules */
    i := 1; M := φ;
    for i := 1 to n do
       /* n: Total number of attributes */
     begin
       while ( Li ≠ φ  ) do
         begin
           Select one pair R = ∧[ai = vj ] from Li;
           Li := Li − {R};
           if (αR(D) > δα)
             then do Sir := Sir + {R};
             /* Include R as Inclusive Rule */
           else M := M + {R};
         end
       Li+1 := (A list of the whole combination of
                 the conjunction formulae in M);
     end
  end {Postprocessing Procedure};
```

图2 后处理过程算法

3.4 交叉验证和鞋带方法

交叉验证方法对误差的估计的步骤为:首先,将训练样 L 本分成 V 块:$\{L_1,L_2,\ldots,L_V\}$.然后,重复 V 次,从训练样本 $L-L_i$(i = 1, …,V)中归纳规则,并利用 L_i 作为测试样本来计算误差率 err_i.最后,整体误差率 err 便可以通过平均 err_i 来求得,即 $err=\sum_{i=1}^{V} err_i/V$ (称作 V 次交叉验证法).因此,可通过将 err 替换成 SI 和 CI 的方法来计算 SI 和 CI.

而鞋带方法的步骤为:首先,从初始训练样本中得到经验可能性分布值(F_n).第 2 步,应用 Monte-Carlo 算法,利用 F_n 随机抽取训练样本.第 3 步,利用新的训练样本归纳规则.最后,利用初始训样本来测试这些结果,同时统计测度(如误差率)等也能得到.这 4 步被迭代有限次.一般地,迭代 200 次对误差估计便足够了.

有趣的是,Eforn 发现 2 次交叉验证法近似等于对数据的一个完全新模式的预测估计,而鞋带方法近似等于最大区间估计.因此,前者可以作为所有测度的下限,而后者可作为它们的上限.

在植物病害诊断归纳推理方法中,为了降低交叉验证的不一致性,还特地引入 Walker 提出的重复交叉验证方法[8].在该方法中,交叉验证方法被重复执行(通常为 100 次),并且估计是在所有上的均值.为了避免过度预测和不一致性,本文采用了重复 2 次交叉验证和鞋带方法.

4 实验结果

4.1 方法的性能

该方法最大的一个优点就是,与其它的传统知识获取方法比较,可以大大缩短植物病害知识的获取过程.一般来说,获取一个植物病害专家系统的规则至少要花掉 6 个月的时间,而使用该方法可以将这个过程缩短到几个星期.PRI-REX2 系统从数据库中归纳出规则,植病专家只需要检测这些规则就行了.知识获取是开发专家系统,尤其诊断类专家系统的瓶颈,本方法可以对诊断类专家系统的开发起到一个很好的辅助作用.

4.2 归纳规则的性能

实验使用的油菜病害数据库包括 463 个病例,51 种类别和 16 个属性.在实验中,δ_α和δ_κ分别置为 0.75 和 0.5.实验分为 4 个步骤:第 1 步,这些记录被随机分为两部分,一半作为新的训练样本,一半作为新的测试样本.第 2 步,分别对新的训练样本应用 PRI-REX2,AQ15 和 CART 方法进行归纳.第 3 步,对新的训练样本应用交叉验证方法

和鞋带方法以得到 PRI-REX2 的精确性和覆盖的误差估计.第 4 步,用新的测试样本测试归纳出的规则.这四步被重复了 100 次,并将 100 次实验的误差进行平均.

实验结果见表 2.排除规则精确度(Exclusive rule accuracy, ER-A)表示有多少训练样本没有包含在一个被正确的从侯选集中排除在外的类中.包含规则精确度(Inclusive rule accuracy, IR-A)表示平均分类的精确度.最后,病害映像精确度(Disease image accuracy, DI-A)表示有多少症状没有被诊断结论来解释,由病害映像来测试.第一行是应用 PRI-REX2 得到的结果,第二行是植病专家的结果.表中第 3、4 行比较了 CART 和 AQ-15 对归纳规则的分类精确度.表中第 5、6 行表达了重复交叉验证方法(R-CV)和鞋带方法(BS)的误差估计结果.最终结果显示:第一,归纳规则仅比医学专家的结论差很少;第二,该方法比传统的经验方法 CART 和 AQ15 稍好;第三,R-CV 和 BS 误差估计可以分别作为每条规则误差的上限和下限估计,因此,这两个值之间的区间值也可用作每条规则的性能误差估计.

表 2 实验测试结果

方法 Methods	排除规则精确度 (%) ER-A	包含规则精确度 (%) IR-A	病害描述精确度 (%) DI-A
PRI-REX2	92.1	84.0	89.7
Expert	97.0	92.0	94.0
CART	—	83.3	—
AQ15	—	84.7	—
R-CV	71.6	77.2	81.5
BS	95.6	89.4	92.1

5 结束语

本文提出了基于粗糙集的油菜病害数据库的规则归纳算法,实现了油菜病害诊断通用程序的三个步骤的自动化,即:观察、假设(特征提取)和检验(区分).实验结果表明,粗糙集不仅是一个表示不确定知识抽取过程的良好框架,而且是归纳植病表述的极好工具,在农业诊断型专家系统中有着广泛的应用前景.但是,该方法现在并不适合所有的推理过程,譬如收集相似病害的规则(聚集),利用领域知识解释特征(解释),这正是我们下一步要做的工作.

致谢 在此,我们向对本文的工作给予支持和建议的同行,尤其是重庆邮电学院的王国胤教授和日本的 Tsumoto 教授表示感谢.

References:

[1] Joseph Giarratano, Gary Riley. Expert Systems Principles and Programming. 2nd ed., New York: PWS Publishing Company. 1998. 60~68.
[2] Y.Y. Yao, T.Y. Lin. Generalization of Rough Sets using Modal Logics. Intelligent Automation and Soft Computing, 1996,2(2):103~120.
[3] Matsumura Y., Matunaga T., Hata Y. Consultation system for diagnoses of headache and facial pain: RHINOS. Med Info. 1986. 11:145~157.
[4] Tsumoto, S. and Tanaka, H. Automated Discovery of Medical Expert System Rules from Clinical Databases based on Rough Sets. In: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, CA: AAAI press, 1996: 63~69.
[5] Tsumoto, S. Mining diagnostic rules from clinical databases using rough sets and medical diagnostic model. Information Sciences, 2004,162(2):65~80.
[6] Li AP, Liao GP, Wu QY. A rule induction method of plant disease description based on rough sets. In: Wang G, *et al.* eds. RSFDGrC 2003. LNAI2639, New York: Springer-Verlag Berlin Heidelerg, 2003. 259~263.
[7] Pawlak Z. Rough Sets. Dordrecht: Kluwer Academic Publishers, 1991. 15~19.
[8] Skowron A. Rough sets: Trends and challenges. In: Wang G, *et al.* eds. RSFDGrC 2003. LNAI 2639, New York: Springer-Verlag Berlin Heidelerg, 2003. 116~123.
[9] Walker MG, Volkmuth W, Sprinzak A, Hodgson D, Klingler TM. Prediction of gene function by genome-scale expression analysis: prostate-cancer associated genes. Genome Research, 1999,9(12):1198~1203.