多种诊断推理方法的设计与比较

魏圆圆　王儒敬　方　静
(中国科学院合肥智能机械研究所,合肥 230031)
E-mail:jsjwyy@163.com

摘　要　简单介绍了诊断推理领域三种常用的方法：基于人工神经网络的方法、基于规则的方法和基于案例的推理方法,并以某农作物病害诊断为实例,详细介绍了三种方法的具体设计,并从诊断正确率和诊断时间上对三种方法的诊断性能进行了比较,比较结果为特定问题条件下选择适当的诊断推理方法提供了帮助。

关键词　人工智能　诊断推理　BP网络　决策树　基于案例推理

文章编号 1002-8331-(2005)02-0057-02　**文献标识码** A　**中图分类号** TP18

The Designing and Comparing of Multi Diagnosis and Reasoning Methods

Wei Yuanyuan　Wang Rujing　Fang Jing
(Hefei Institute of Intelligent Machines, Chinese Academy of Sciences, Hefei 230031)

Abstract: This paper introduces three common methods in the field of diagnosis and reasoning: Artificial Neural Network, Rule-Based Reasoning and Case-Based Reasoning. We explicate the particular designing of these methods in a crop disease diagnosis and compare the diagnosis performance on diagnosis accuracy and diagnosis efficiency. The result will help the user select appropriate diagnosis and reasoning method on the given problem condition.

Keywords: artificial intelligence, diagnosis reasoning, Back-Propagation neuron network, decision tree, Case-Based Reasoning

1　引言

诊断推理作为人工智能的一个重要研究领域,近年来发挥了越来越大的作用。基于人工神经网络的方法、基于规则的方法和基于案例的推理方法是诊断推理领域中常用的方法[1]。如何根据问题的性质,选择有效的推理方法和实现方法,一直是诊断推理中的重点。文章简要介绍了这三种方法,并以某种农作物病害诊断为例,详细说明三种方法的实现过程,并对其诊断性能进行比较。具体实现中人工神经网络方法采用BP网络,基于规则的方法采用经典的决策树归纳方法用于产生诊断规则。

2　三种诊断推理方法简介

2.1　基于人工神经网络的推理方法

人工神经网络是由大量神经元通过极其丰富和完善的联结而构成的自适应非线性动态系统,是一种黑箱方法。它以其极强的容错性、鲁棒性以及广泛的自学习能力逐渐成为诊断推理领域的一个重要方法[2]。基于人工神经网络的诊断推理的执行分为两个过程:训练过程和诊断过程。

2.2　基于规则的推理方法

基于规则的诊断主要是利用领域专家的启发式经验知识,知识通常采用IF-THEN规则形式表示。诊断推理采用模式匹配的方法,其推理过程包括由征兆到结果的正向推理和由结果到征兆的反向推理。该技术的特点在于它具有极强的演绎推理能力,根据直观的推理规则解决问题,其不足之处在于它在解决实际问题时,必须首先根据问题给出的条件得到用于推理的规则。故基于规则的推理方法关键在于规则知识的获取。

2.3　基于案例推理

基于案例推理以案例为核心,利用人们以往求解类似问题的经验知识进行推理,从而获得当前问题求解结果的一种推理模式。一个基于案例推理系统将过去的经验表示为案例的形式存放在案例库中。当新问题出现时,系统便在案例中检索出相似的案例,对它们进行综合和修正来产生一个解答。问题解答成功后,即可根据此问题创建一个新的案例并加入案例库中。这种方法体现了人类专家在推理中的思维过程,直观而且行之有效。

3　具体实现

实验中采用某种病害诊断实例,每个实例4个症状属性,分别为:病斑颜色、病害部位、病害形状、病害特征,诊断结果即病的种类分为6种,分别为:炭疽病、印度炭疽病、褐斑病、角斑病、叶斑病、红粉病。样本以记录的形式表示,存放在access数据表中,代码在delphi 6开发环境下编写。

训练样本集和测试样本集根据训练比例随机生成。

3.1　BP网络的实现

基金项目:国家863高技术研究发展计划项目(编号:2001AA115170)资助

作者简介:魏圆圆(1980-),女,硕士研究生,主要研究方向为数据挖掘、智能方法应用。王儒敬(1964-),男,副研究员,硕士生导师,主要研究方向为人工智能。方静(1979-),女,硕士研究生,主要研究方向为软件Agent技术。

根据问题特点，由病害的4个方面症状推导出该病害的类别。采用3层网络模型，输入层四个神经元，对应病害4个症状，中间层10个神经元，6种病害类别用6位二进制(0、1串)表示，故输出层包含6个神经元。神经元的激活函数采用常用的Sigmoid函数。

样本的症状属性和类别属性都是枚举类型，但网络的学习和应用都是数值的运算，故须将枚举元素转化为实数，这里的处理方法是：将每个症状属性的取值都相对平均地分布在0和1之间。比如，病斑颜色有6个可能的值：褐色、淡褐色、黑褐色、深褐色、粉红色、暗红色，分别对应0、0.2、0.4、0.6、0.8、1。教师信号采用6位二进制串，即用100000、010000、001000、000100、000010、000001对应6种病虫害。

经过反复调试，网络中步长系数为0.2，稳定系数为0.6。程序在输出误差小于0.05时结束训练。极少情况下，网络不能收敛或者需要较长时间才能收敛，这里处理的办法是，在训练次数达到3000且误差仍大于0.05时，重新初始化权重进行训练。

3.2 决策树归纳的实现

采用决策树归纳获取诊断规则。通过候选测试属性(判定属性)和类别属性构造一棵决策树。训练样本集用于得到决策树，测试样本用于计算分类规则的准确率。要求所有候选属性和类别属性都为离散属性，连续值必须离散化。算法的基本策略如下[4]：

(1)决策树以代表训练样本的根节点开始。

(2)如果该节点中所有样本都在同一个类，则该节点成为树叶，并用该类标记。

(3)否则，计算该节点中所有候选测试属性的信息增益，选择具有最高信息增益的属性作为当前节点中样本集的测试属性，即判定属性。

(4)对测试属性的每个已知的值，创建一个分支，并据此划分样本。

(5)算法使用上述过程，递归地形成每个划分上的样本判定树。一旦一个属性出现在一个节点上，在其后代节点上就不必考虑该节点。

(6)递归划分步骤仅当下列条件之一成立时停止：

①给定节点的所有样本属于同一类。

②没有剩余属性可以用来进一步划分样本，在此情况下，使用多数表决，将给定节点转换成叶子节点，并用样本中的多数所在的类标记它。

为方便决策树的构造和分类规则的提取，采用静态链表实现决策树的存储，该模块中数据结构定义如下：

```
type
  TMiningFieldInfo=record
    FieldName:string;            //字段名
    FieldData:array of variant;  //字段值
  end;
  TMiningTableInfo=record        //表划分后的信息
    FieldCount:integer;          //表维数(划分一次,少一维)
    RecordCount:integer;         //记录个数
    FieldInfos:array of TMiningFieldInfo;
  end;
  TPartitionInfo=record
    AttrValue:string;            //划分属性值
    Pos:integer;  //根据该属性值得到的划分在链表中的位置
  end;
  TDecisionTree=record //静态链表(数组)实现决策树的存储
    CurTableInfo:TMiningTableInfo;//根据某一属性划分后的表信息
    CurTestAttr:string;//当前测试属性
    parent:integer;//该节点的父亲节点编号
    IsSame:boolean;//标志是否已经在同一类中
    AttrPartition:array of TPartitionInfo;//由测试属性的值引出的分支
  end;
  TClassification=class //定义类
    TableName:string;//表名
    ClassAttr:string;//类别属性
    TestAttr:array of string;//候选测试属性
    TrainingRate:double;//训练集比例
    DecisionTree:array of TDecisionTree;//决策树
  public
    constructor Create();
    procedure Produce_Set();//产生训练样本集和测试样本集
    Function InOneClass(CurSet:TMiningTableInfo):boolean;//判断当前划分中的样本是否在一个类中
    function Calculate_Class_I (CurSet:TMiningTableInfo;c:integer):real;//计算样本分类所需的期望信息
    Function Calculate_Info_Gain(CurSet:TMiningTableInfo;m:integer):real;//计算信息增益
    procedure DoClassification();//分类,生成决策树
    procedure Produce_Rules(arrlen:integer);//产生 if…then..分类规则
    function Calculate_CorrectRate ():real;//计算分类规则对测试样本的正确率
  end;
```

3.3 基于案例推理的实现

针对该文问题，基于案例推理在这里的解决方法较前两种方法的实现要简单得多。训练样本集作为成功的诊断案例库，测试样本作为待诊断案例，和成功案例进行匹配。案例推理的核心在于案例之间相似性的度量，这里的方法是根据经验给每一症状属性赋一权重，分别为：病斑颜色：0.01，病害部位：0.1、病害形状：0.001，病害特征：0.001。待诊断案例的症状取值与案例库中案例的对应症状值进行比较，得到一个标识值，这里标识值定义为：症状取值相等，为1，否则为0，相似性度量值为各属性比较得到的标识值的加权和，取最大者为最相似案例。

4 实验结果及分析

4.1 实验结果

实验中分别对两组病害数据进行测试，第一组为20个互不相同样本组成的样本集，第二组有67个样本，其中包含相同样本。根据训练比例得到训练样本，剩余样本为测试样本(训练比例为100%时，测试样本和训练样本均为整个样本集)，在同一训练集和测试集上运行上面三种方法，共进行6组实验，比较其诊断准确率和运行时间(从算法开始到诊断完毕的时间)，实验结果见表1和表2。

4.2 实验分析

(下转119页)

检查的关键词，进而要求过滤系统直接丢弃有关信件，最终使得原邮件服务器下属用户免于受其感染和危害。上述阶段的系统运行情况（表1）说明，过滤系统在拦截垃圾信件的同时，有效缓解了“My Doom”病毒对于后端邮件服务器的影响和破坏，系统在一定程度上能够协助维护人员应对网络突发事件。

表1 前置式邮件过滤系统运行情况

（二〇〇四年一、二月间）

拦截疑似垃圾邮件总数	垃圾邮件I	垃圾邮件II	误拦邮件数目（占垃圾邮件I比率）
181986	238	181726	22（9.2%）

说明：①垃圾邮件I含反动、色情信息；②垃圾邮件II是商业广告和“My Doom”病毒载体信件；③所有误拦邮件都被维护人员及时恢复转发。

5 结束语

随着电子邮件成为人们主要通信工具，针对日益泛滥的垃圾信件实现有效过滤已经成为网络邮件领域的研究热点。独立于原邮件服务器的前置式邮件过滤系统具备强独立性，系统架设不会改变邮件服务器的参数设置和运行方式。系统动态统计各SMTP客户端的发信频率，抵御邮件炸弹攻击；系统通过基于有限自动机的DFSA算法，对于常见汉字编码的电子邮件统一实现快速的内容过滤；系统提供完善的邮件放回功能，可以恢复拦截信件的正常转发，从而把误拦有用邮件的可能降至最低。

已经成形的前置式邮件过滤系统还能借鉴数据挖掘机制，选用合适的语义分析技术和机器学习算法，增强系统智能分析水平，这同时也是笔者今后进行邮件过滤产品研发的方向和切入点。（收稿日期：2004年7月）

参考文献

1.Jonathan B Postel.SIMPLE MAIL TRANSFER PROTOCOL[S].RFC821 IETF,1982

2.David H Crocker.STANDARD FOR THE FORMAT OF ARPA INTERNET TEXT MESSAGES[S].RFC822 IETF,1982

3.杨峰，曹麒麟，段海新等.基于DNS Blocklist的反垃圾邮件系统的设计与实现[J].计算机工程与应用，2003；37(7)：11~12

4.陈细谦，熊文龙.基于Qmail的邮件过滤系统的设计与实现[J].现代计算机，2001；9

5.M Mohri.ON SOME APPLICATIONS OF FINITE-STATE AUTOMATA THEORY TO NATURAL LANGUAGE PROCESSING[J].Natural Language Engineering,1996

（上接58页）

表1 诊断正确率(%)

	样本集1(20个样本)			样本集2(67个样本)		
	TR=60%	TR=80%	TR=100%	TR=60%	TR=80%	TR=100%
BP网络	62.5%	50%	100%	88.9%	100%	100%
决策树归纳	62.5%	75%	100%	85.2%	100%	100%
CBR	75%	75%	100%	92.6%	96.2%	100%

注：TR为训练比例(TrainRate)

表2 运行时间(毫秒 ms)

	样本集1(20个样本)			样本集2(67个样本)		
	TR=60%	TR=80%	TR=100%	TR=60%	TR=80%	TR=100%
BP网络	69ms	110ms	261ms	190ms	220ms	281ms
决策树归纳	<1ms	9ms	9ms	<1ms	9ms	9ms
CBR	<1ms	<1ms	<1ms	<1ms	<1ms	<1ms

根据实验结果，作如下分析：

(1)神经网络方法较其它两种方法运行时间较长，其时间主要用于网络的训练；决策树归纳的时间主要用于诊断规则的获取，基于案例推理直接将待诊断样本按照一定的相似性度量准则与案例库中案例进行比较，所以算法比较简单，诊断效率较高。

(2)当训练集比例为100%时，所有样本既为训练样本，也为测试样本，训练正确率都为100%。对bp网络而言，对有限样本，网络有着很强的学习能力，足以记住每个样本。对决策树归纳，从所有样本中提取规则，在完全分类的情况下，对每个样本也能正确分类，基于案例推理中每个待测样本都能找到它本身与其对应，100%的正确率更是显然。

(3)训练比例小于100%时，对样本集一的诊断正确率显然低于对样本集二的诊断结果，这是由样本集的取样决定的。对于样本集一，20个样本互不相同，BP网络的过学习、决策树归纳得到规则的不全面、基于案例推理较差的容错性导致了其诊断正确率较低；而样本集二因为样本的重复性，使得三种方法的缺点对其诊断结果没有太大影响。

由上面分析可归纳出：训练样本集较全，对诊断时间要求较高时，可采用基于案例的方法；用户需要得到直观的诊断规则，可采用基于规则的方法；训练样本有限，且对时间要求不高时，可采用基于神经网络的方法。

5 结论

该文介绍了三种常用的诊断推理方法，并以样本属性均为文本型的病害诊断为例，对三种方法进行了实现和分析。在对同一问题的解决中，各方法都体现出了自身的特点：人工神经网络从训练样本中获取样本分布特征的统计规律，训练和诊断过程都是“黑箱”操作，克服了传统诊断推理方法知识获取的瓶颈，应用面较广，但训练时间较长、过程不可见是其缺点。采用决策树归纳方法得到规则的基于规则推理，可以直观地得到规则，用户可以通过这些规则对问题本身的规律性达到很好的认识，但是由于问题的多样性，规则获取难以实现，同时受训练样本集影响，训练样本不全面可能会导致获取规则的不全面，这是导致诊断不准确的直接原因。基于案例推理通过寻找案例库（这里即为训练样本）中过去同类问题的求解从而获得当前问题的的解，该方法简单、易懂，但对训练样本集的要求较高，相似性度量也难以确定。在实际问题解决中，要求从样本的分布、掌握算法的能力，对时间的要求等方面选择适合问题求解的推理办法。（收稿日期：2004年7月）

参考文献

1.Kleer J et al.Characterizing diagnosis and systems[J].Artificial Intelligence,1992；56：197~222

2.张晓莉等.诊断推理中人工神经网络和基于案例推理的结合[J].上海铁道大学学报，2000；(6)

3.Nils J Nilsson.Artificial Intelligence：A New Synthesis[M].Morgan Kaufmann Publishers,Inc,1998

4.阎平凡，张长水.人工神经网络与模拟进化计算[M].北京：清华大学出版社，2000

5.Jiawei Han,Micheline Kamber.Data Mining:Concepts and Techniques [M].Morgan Kaufmann Publishers,Inc,2001