一种基于二叉树的元知识学习算法MKL及其应用*

潘金贵 陈彬 陈晶 陈世福

(南京大学计算机科学系,南京 210093)

摘要 MKL是知识获取系统NDKAS中实现的一个元知识学习算法,它在分类及抽象的基础上归纳出二叉树结构的元知识,用以有效地组织知识库中的规则.MKL生成的元知识满足元知识的基本性质.本文给出了MKL的算法描述,基本性质的满足性证明及算法的应用例子.

关键词 概念获取,经验知识分类,元知识学习.

专家系统的知识库也可以象一般的复杂系统那样分解为层次结构.利用层次结构能够简化对知识的描述,改善知识库中规则的组织方式,即可以按能够求解问题的不同类别来结构化.问题求解时,它仅涉及与该问题相关的唯一子知识库(如果问题仅有唯一解),这样,其效率就比将知识库按照知识的不同侧面以及知识在求解过程中表现出的顺序关系来结构化的系统(如MYCIN[1]等)高得多.为了描述这种层次性,需要引入关于如何有效使用知识库中知识的知识,即所谓的"元知识"[2].

知识获取系统NDKAS[3]提供了元知识学习机制,可以自动地从知识库中抽取需要的元知识.在NDKAS系统中,元知识是以二叉树形式来表示的.二叉树的中间节点构成元知识层,其叶结点是一些彼此密切相关的经验知识组成的子知识库.NDKAS系统的元知识学习机制,由经验知识分类和元知识抽取两部分组成,如图1所示.

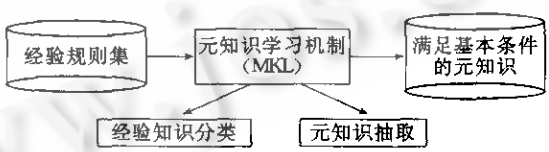

图1 元知识学习机制

经验知识分类即把系统提供的经验规则进行分类,得到两个子集,因系统并不知道各经验规则所属的类别,这是一种无指导的观察学习方法[2],采用下节描述的MKL算法中的

* 本文1993-09-14收到,1994-03-09定稿

本课题得到国家自然科学基金的资助.作者潘金贵,1952年生,副教授,主要研究领域为人工智能和知识工程及应用.陈彬,1965年生,讲师,主要研究领域为人工智能与机器学习.陈晶,女,1966年生,工程师,主要研究领域为计算机应用.陈世福,1938年生,教授,主要研究领域为人工智能与图像处理.

本文通讯联系人:潘金贵,南京210093,南京大学计算机科学系

© 中国科学院软件研究所 http://www.jos.org.cn

ClusterND过程实现;而元知识的抽取,则是从ClusterND分类的结果中归纳出一个元知识层的描述,即项,该项覆盖其一子集,而排斥另一子集,采用归纳学习算法CAP2中的induce过程实现.限于篇幅,CAP2学习算法将另文描述.

1　元知识学习算法MKL

元知识学习机制是通过元知识学习算法MKL来实现的.算法{MKL(H_i):元知识},通过不断地调用ClusterND及induce过程递归构造生成元知识的二叉树结构,该算法的结束条件可由用户进行控制,即由用户提供对树深度及子库大小的参数*Deepr*和*Sizer*.

1.1　MKL算法

{MKL(H_i):元知识},其中H_i为一个规则集.

STEP1. 判断是否满足结束条件?即:如果$|H_i|<Sizer$或H_i的深度$>Deepr$,则结束,否则继续执行STEP2;

STEP2. 调用ClusterND过程,把H_i分成H_{2i}及H_{2i+1};

STEP3. 视H_{2i}为正例,H_{2i+1}为反例,调用induce过程,产生树结构中相应项的描述.

STEP4. 调用MKL(H_{2i})及MKL(H_{2i+1}),递归构造左子树及右子树.

STEP5. 构造以H_i为根,以H_{2i}和H_{2i+1}为左右子树的树.

1.2　ClusterND过程的实现

ClusterND是基于相似率的二叉树分类方法,对于提供的规则集H_k,根据相似率最大原则,把H_k分成两棵子树H_{2k}及H_{2k+1},且$H_{2k}\cap H_{2k+1}=$空,$H_{2k}\cup H_{2k+1}=H_k$.

为了形式地描述它,先给出下面几个定义:

定义1. 规则集H_1,H_2且$H_2\subset H_1$,H_1中出现的属性集*Attr*,对于属性$x\in Attr$,其属性值集为$List[x]$,$Sim_A(H_2|H_1)$表示H_2在H_1中的相似属性集,则

$$Sim\_A(H_2|H_1)=\{x: x\in Attr,\exists\, y\in List[x],\exists\, D\in(H_1-H_2)\ ((x=y)\text{在}D\text{中出现,但不在}H_2\text{的任何规则中出现})\}$$

事实上,$Sim_A(H_2|H_1)$是属性集*Attr*中一些属性所构成的集合,这些属性和其属性值可能生成的属性描述子均不出现在H_2中,而一定可在H_1-H_2中找到.

定义2. $Sim(H_2|H_1)$表示在H_1中H_2的相似率,则

$$Sim(H_2|H_1)=\frac{|H_2|}{|H_1|}\times\frac{|Sim\_A(H_2|H_1)|}{|Atrr|}$$

$Sim(H_2|H_1)$是一个经验公式,它提供了从H_1中抽取出的集合H_2所具备的某种相似性质的度量值即相似率,ClusterND根据该值决定了对H_1的划分.

定义3. 一个集合$H_2(\subset H_1)$是H_1中属性相关集,当且仅当存在一属性$i\in Attr$,相应属性值j,满足:

$$\forall\, D\in H_2((i=j)\in D)\ \text{且}\forall\, D_1\in(H_1-H_2),\forall\, k\in(List[i]-\{j\})((i=k)\notin D_1)$$

例如,对于表1描述的经验规则1,…,8:

$H_1=\{1,2,3,4,5,6,7,8\}$,

Attr={毛发,牙齿,眼睛,羽毛,脚,食物,奶,会飞,产卵,会游泳},

对于$H_2=\{1,2,3,4\}$,有:

© 中国科学院软件研究所 http://www.jos.org.cn

$Sim\_A(H_2|H_1)$={毛发,牙齿,羽毛,脚,食物,奶,产卵,会游泳},则:
$Sim(H_2|H_1)=(4/8)\times(9/10)=45\%$.

表1 一个动物专家系统的经验规则

	毛发	牙齿	眼睛	羽毛	脚	食物	奶	会飞	产卵	会游泳
1. 虎	有	犬齿	前方	无	有爪	肉	有	不	不	是
2. 豹	有	犬齿	前方	无	有爪	肉	有	不	不	是
3. 长颈鹿	有	钝	旁边	无	蹄	草	有	不	不	是
4. 斑 马	有	钝	旁边	无	蹄	草	有	不	不	是
5. 鸵 鸟	无	无	旁边	有	有爪	谷	无	不	是	不
6. 企 鹅	无	无	旁边	有	蹼	鱼	无	不	是	是
7. 信天翁	无	无	旁边	有	有爪	谷	无	是	是	不
8. 鹰	无	无	前方	有	有爪	肉	无	是	是	不

1.3 ClusterND过程的算法描述

ClusterND(H_i,Sr) : H_{2i},H_{2i+1}

其中Sr是用户提供的相似率比较值,缺省时为100%.

初始时:$Attr=H_i$中出现的属性集,

$List[i]=H_i$中出现的相应属性$i(i\in Attr)$的值集,

两个中间变量:$bigs=0$, $big_set=$空

STEP1. 计算$HH=\{H_s: H_s\subset H_i$且对$\forall\ D\in H_s\ \exists\ s,t((s\in Attr)\wedge(t\in List[s])$
$\wedge((s,t)\in D)\wedge\forall\ D_1\in(H_i-H_s)((s,t)\overline{\in}D_1))\}$
$Sim_attr=$空(H_i中的属性相关的规则集的集合)

STEP2. 根据定义3,判断HH中的每一集合D是否是H_i的属性相关集?若是,则
$Sim\_attr=Sim\_attr\cup\{D\}$,继续执行STEP3.

STEP3. 若$Sim_attr\neq$空,则:
$H_{2i}=Sim_attr$中含规则数量多的集合,
$H_{2i+1}=H_i-H_{2i}$,结束.
若$Sim_attr=$空,则继续执行STEP4.

STEP4. 若HH不空,则从HH中取出一集合元素H_2,$HH=HH-\{H_2\}$,
计算$Sim_A(H_2|H_1)$, $Sim(H_2|H_1)$,转向执行STEP5.
若HH空,转向STEP6.

STEP5. 若$Sim(H_2|H_1)>Sr$,则$H_2=H_1-H_2$,结束;
否则,$bigs$取$Sim(H_2|H_1)$和$bigs$两者中最大值.
$big_set=$与$bigs$相应的集合H_2.
转向STEP4.

STEP6. $H_{2i}=big\_set$, $H_{2i+1}=H_i-H_{2i}$,结束.

1.4 元知识基本性质

在NDKAS中,二叉树表示的元知识层中每个结点是一个项,沿着元知识层自顶向下,

© 中国科学院软件研究所 http://www.jos.org.cn

可获得每一项所"覆盖"的子知识库集,这些子知识库集称为该项的左处延,即左子树的所有叶结点构成的规则集,而右子树所有叶结点构成的规则集称为该项的右外延.每一项存放的内容是该项的左外延区别于同一结点左外延的本质属性,这些本质属性即元知识,是在经验知识基础上抽象整理而生成的.例如,对表1的经验规则,生成的元知识结构可示为图2.

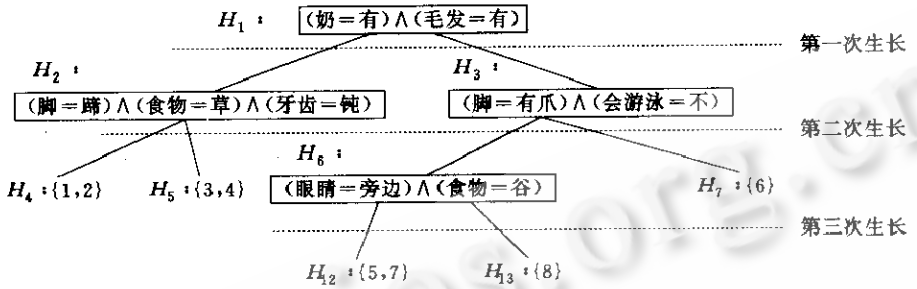

图2 由MKL生成的二叉树结构的元知识

由此可得出元知识具备下列2个基本性质:

(1)各子库不相交.

(2)元知识层结点上项所含的本质属性反映了该项左外延构成的一组子知识库所能解决问题必须具备的一些事实.如果该问题不满足这些元知识,则这组知识库中肯定无解.

定理. MKL算法从所给的经验规则集中分类、抽取产生的元知识满足元知识的基本性质.

证明:因为根据其ClusterND算法,显然产生的各子知识库是不相交的,而CAP2算法能保证生成项的描述的完备性与一致性[4],因此,MKL算法产生的元知识满足元知识基本性质.定理成立.

2 MKL算法的应用示例

以表1为例,说明如何从经验规则学习出相应的元规则,从而进一步理解MKL算法.

设用户给出树深度 $Deepr=3$,子库大小 $Sizer=1$,相似率比较值 $Sr=1$.

MKL(H_1,1):

第一次生长:$H_1=\{1,2,3,4,5,6,7,8\}$

$Attr=${毛发,牙齿,眼睛,羽毛,脚,食物,奶,会飞,产卵,会游泳}

用ClusterND算法分类:

$HH=\{\{1,2,3,4,5,6\},\{1,2,3,4,6\},\{3,4,5,6,7\},\{1,2,5,7,8\},\{1,2,3,4\},\{5,6,7,8\},\{1,2,8\},\{5,7,8\},\{3,4\},\{1,2,\},\{5,7\},\{7,8\},\{6\}\}$

因为 $Sim_atrr=$空,则据算法的STEP4,STEP5,STEP6得:$H_2=\{1,2,3,4\}$

其中 $Sim\_A(H_2|H_1)=${毛发,牙齿,羽毛,脚,食物,奶,会飞,产卵,会游泳}

则有 $Sim(H_2|H_1)=(4/8)*(9/10)=45\%$

$H_3=\{5,6,7,8\}$

用CAP2算法进行元知识抽取:$NOTE_1=$(奶=有)∧(毛发=有)

© 中国科学院软件研究所 http://www.jos.org.cn

生成第一层元知识如图 2 所示. 按上述同样步骤进行第二次、第三次生长，生成第二层及第三层元知识，见图 2 所示. 第四次生长：因为 $Deep=4>Deepr$，所以结束生长. 图 2 中所示二叉树结构即为所求的元知识.

3 结束语

NDKAS 是一个集学习、精化、推理于一体的完整的知识获取系统[3]，它从所提供的训练实例中，自动构造知识库，然后利用该知识库进行咨询.

由于 NDKAS 系统引进了元知识学习机制，使得生成的知识库可按能够求解问题的类别来结构化，大大提高了知识的使用效率. 元知识学习算法 MKL 具有简单，有效，实用性强等特点，其精度可以调节，并可由用户进行控制，亦已证明，它生成的元知识，满足元知识的基本性质，并已成功地应用于"新构造控水专家系统"知识库的自动构造.

参考文献

1 Buchanan B G, Shortliffe E H. Rule—based expert systems: the MYCIN experiments of the Stanford heuristic programming project. Addison—Wesley, Reading MA, 1984.

2 Michalski R S, Steuspp R E. Learning from observation: conceptual clustering. In: Michalski R S, Carbonell J G, Mitchell T M eds, Machine Leaning: an Artificial Intelligence, Approach, 1982.

3 潘金贵，陈彬，陈晶，陈世福. 知识获取系统 NDKAS 的研究与应用. 计算机学报，待发表.

4 潘金贵，陈彬，陈晶，陈世福. 归纳学习算法 CAP2 的研究与应用. 软作学报，待发表.

AN ALGORITHM FOR META-KNOWLEDGE LEARNING MKL AND ITS APPLICATION

Pan Jingui Chen Bin Chen Jing Chen Shifu

(Department of Computer Science, Nanjing University, Nanjing 210093)

Abstract MKL, an algorithm for meta-knowledge learning, is presented in NDKAS, which is a knowledge acquisition system. On the basis of classification and abstraction, it can induces meta-knowledge which is used to effectively organize rules in knowledge base in binary—tree structure, which satisfies the essential properties of meta-knowledge. The structural representation of meta-knowledge, the algorithm description of MKL and the proof of satisfaction of the essential properties are given in the paper. An example of MKL application is also described.

Key words Concept acquisition, experience knowledge classification, meta-knowledge learning.

© 中国科学院软件研究所 http://www.jos.org.cn