

基于知识辞典的范例推理系统的规则提取模型^{*})

叶家成 汤胤 彭宏 郑启伦

(华南理工大学计算机科学与工程学院 广州510640)

摘要 本文提出了把范例分解为范例描述与结论、分别进行概化、在概化的每个范例模板中进行规则提取的模型,该模型可以解决范例与规则的衔接问题,为 CBR 与 RBR 集成推理提供了基础,并且对智能化范例调整也有帮助。最后,文章讨论了模型中存在的一些问题。

关键词 范例推理, 范例表示, 规则提取

Extracting Rules from CBR Based on Knowledge Dictionary

YE Jia-Cheng TANG Yin PENG Hong ZHENG Qi-Lun

(Institute of Computer Science, South China University of Tech., Guangzhou 510640)

Abstract This paper proposes a model for extracting rules from the case base, splitting the case into description and conclusion parts. With the aid of domain knowledge, the two parts are respectively abstracted toward the parts of the rule. The model is helpful to connect the cases and rules, therefore offers the basis of integrating the CBR and the RBR. It also enables the automation of case revision. Some of problems are discussed on the model.

Keywords CBR, Case representation, Rule extraction

1 引言

基于范例推理(Case-Based Reasoning, CBR)是一种类比推理,是利用人们以往求解类似问题的经验知识进行推理,从而获得当前问题的求解结果。但基于范例的推理只能对已存在的范例进行搜索,并且容易将表面相似现象看成本质相似,导致错误判断。基于规则的推理系统(RBR)只能对事先预想到并提供了规则进行推理,并且结论的解释和对策比较呆板,不能灵活适应情况的变化。

研究者提出了多种 CBR 和 RBR 的集成方法, Bonissone 和 S. Ayub 等研发的 CARS 系统^[1]提出了 CBR 和 RBR 作为等同的两个推理系统进行集成, Rissland 和 Skalah 等研发的 CABARET 系统^[2]则提出了在 CBR 系统体系架构下在需要的时候使用 RBR 推理功能。Barry Smyth 提出层次型的范例库模型^[3]。这些集成推理方法充分发挥了 CBR 和 RBR 的优点,但其本质都是在推理的不同情况和不同阶段分别使用 CBR 和 RBR 推理,彼此之间还是独立的,并没有联系起来。这些集成方法中,规则库是需要由专家给出的,而不断产生的范例库(包括概化的范例)并不能利用到规则库中。从认知的角度来看,专家给出的规则也有一部分是基于范例来取得的,因此如何从 CBR 系统中的范例库生成规则是一个很有意义的课题。在 CBR 系统中,通过一个把范例提取为规则的模型,在推理过程中不断增加范例,规则库也进一步丰富,这样 CBR 和 RBR 就可以通过动态地调整彼此的知识库来得到加强。另一方面, CBR 推理方式的典型的四个阶段:检索(retrieve)、复用(reuse)、修正(revise)和存储(retain)中,修正阶段也需要规则来辅助处理。因此从范例中提取规则,对于完善和智能化基于范例的推理方法是很有意义的。

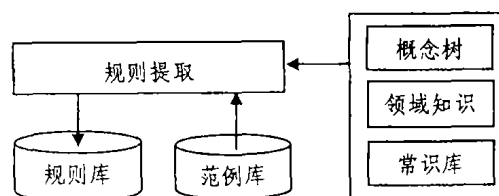


图1 规则提取

2 模型定义

CBR 方法可以完成分析性任务和综合性任务:分析性任务以分析一种情形为焦点,以案例检索为中心,解答常常不需要更改。分析型任务包括分类、诊断和决策支持;综合性任务用于合成一个新解答,从不同的案例组合出一个解答,包括计划、配置和设计。综合性任务中的范例也可以通过多层分解的方式^[3]拆分为分析性任务的所需要的范例,因此本文主要针对的范例是对分析性任务的范例,其范例结构为(case) = (problem, class)。

我们的核心思想是在多层次 CBR 范例库的架构基础上,将规则纳入整个概念层次范例库的框架中,使规则成为范例多层次概化(materialization)的一个结果。我们将建立将范例概化最终形成规则的模型,并且讨论这当中存在的问题。具体来说包括以下三方面:(1)范例概化方向是规则,具体地说是以规则的结构和内容特征去指导范例多层次的概化;(2)利用领域专家辞典进行概化,解决非数值属性的分类问题。(3)从具体范例中提取范例的调整规则,这也可以为新范例的调整提供辅助规则。

2.1 相关定义

定义1(范例抽象偏序) 假设范例 a 是从范例集 B 中进

^{*})国家自然科学基金重点项目(30230350),广东省科技攻关项目(A1020103)。叶家成 博士生,研究方向:人工智能与知识管理;汤胤 博士生,研究方向:人工智能与数据挖掘;彭宏 教授,博导,研究方向:智能计算技术;郑启伦 教授,博导,研究方向:智能计算技术,神经网络理论及应用。

行规则提取而得到的,则对任意范例 $b \in B$, 定义 $b < a^{[4]}$ 。

定义2(范例概念层次) 一个概念层次 H 是一个偏序集 $(h, <)$, 其中 h 是个有限范例集, $<$ 是 h 上的抽象偏序^[4]。

定义3(范例概念层次树) 按照抽象概念层次结构的方式来组织就构成了该范例库的范例概念层次树^[4]。

树中高层概念是层概念的概括, 树根是该属性的可能属性值的最一般描述, 叶是该属性的可能属性值, 如果属性是连续值属性, 则树叶是一些连续值范围。显然, 叶节点为实际范例节点, 内节点为抽象节点。

定义4 位于范例概念层次 H 的最底层 ($k=0$) 的范例称为具体范例, 其他层次 ($k>0$) 称为抽象范例。

定义5 k 到 $k+1$ 层范例概化模型 M_k 可表示为

$$M_k(C_k, C_{k+1}, D_k) \quad (2.1)$$

其中: C_k 为第 k 层的范例; D_k : 范例 C_k 的属性所对应的知识辞典概念树的节点集。

将范例 C_k 所对应的 D_k 的元素的父节点的值赋与 C_{k+1} 抽象范例相应的属性, 并生成范例 C_{k+1} 的过程称为概化。

概化终止条件: 概化到达知识辞典概念树的顶端时的层次为 n , 此时的 C_n 称为最终规则 R 。

2.2 范例到规则提取模型

范例按照属性、概念、案例类别分别进行索引, 依照这种原则, 在范例树中选取那些与检索有关的具体节点连同其上层节点构成一棵子树, 称之为多层次多索引的概念树(包括约束、类别、情景、因果、相似、相反、归属关系等等)^[4], 这样的概念提升, 既完成了范例的索引过程, 每个抽象范例又可以作为模板使用, 可以在概念树的任一层次选择一个模板进行概化, 对概化后的抽象范例进行规则抽取工作, 范例到规则的提取包括了提取范例调整规则和推理规则。其模型如图2所示。

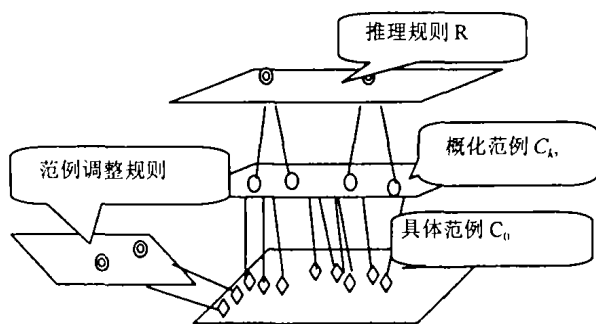


图2 范例到规则的提取模型

建立抽象的范例层次结构的目的是将繁杂的原始信息分类, 使得同一类原始信息具有某种共同的特征, 这个特征就是从原有信息中提取的抽象信息。

概念提升应包括: 特征的提取概略、模块的简化(结构性描述转为辅助语言描述); 知识辞典的检索和提取。

2.3 规则提取步骤

对概化生成的不同层次的抽象范例, 规则提取步骤如下:

(1) 将范例分成两部分, 分别为范例描述和范例结论, 范例分解的方法可以由领域专家人工进行, 也可以利用范例内部的语义进行分割, 根据一些特定的连接词进行分离, 如“如果…那么…”、“…是…”等进行自动分类, 对于结构化范例则可以根据领域知识指定一个或几个属性对应的数据来进行分离。

(2) 对范例描述进行相似度计算, 根据相似度对范例进行分类比较, 得到范例调整规则。

(3) 对范例描述和范例结论分别概化: 对可结构化范例

可以自动聚类进行离散化, 也可以利用知识辞典来进行概化。自动分类算法包括: 关联规则、人工神经网络、遗传算法等; 对非结构化的范例, 以知识词典为主, 也可以采用自动分类方法, 但计算量会比较大。

3 模型应用例子

有以下广州市房产销售的范例库:

表1 一个房产销售的范例库

	位置	房	厅	面积	房龄	价钱
范例1	中山三路	3	2	100平方	1年	50,000
范例2	中山三路	3	2	100平方	5年	35,000
范例3	中山三路	3	2	85平方	3年	38,000
范例4	中山一路	2	2	70	7年	35,000
范例5	棠下	3	2	100	1年	23,000
目标问题	中山三路	3	2	85平方	5年	?

(1) 从具体范例中进行范例调整规则的提取, 我们以目标问题的求解为例进行描述: 首先目标问题进行相似度计算, 为了简单计算, 假设各属性的权重都是一样的, 我们采用了 Block City 简单计算法得到目标问题和范例2相似度最大; 然后计算范例2和其它范例的相似度, 也采用上述算法可得到范例2和范例1的相似度最大, 比较范例2和范例1, 可以得到以下规则:

IF 房龄由1年变为5年 THEN

每平方米价钱减少1,500

利用得到的这个规则可以用于 CBR 系统的范例修正过程中: 在回答目标问题时, 利用得到的规则对范例2进行调整, 则得到目标问题的价钱是: $350,000 - 15 * 1500 = 327,500$ 。

由此可见, 从具体范例中提取得到的范例调整规则对于范例的调整是有很大的作用的, 而且根据不同的领域, 范例调整规则是可以过不同的方法自动化地得到^[3], 这使得范例调整过程可以智能化进行。

(2) 对范例进行概化。根据领域知识, 令房屋价钱为最终规则的结论, 其余的为最终规则的前置条件。其中知识词典有如下内容:

1. 中山一到中山六路、天河北路、北京路是繁华路段, 棠下为一般路段
2. 三房两厅的房子是间隔宽敞的, 两房两厅是中间隔适中的
3. 70万豪华房, 50-70万是高价房, 30-50万是中价房, 30万以下是经济型房子
4. 面积100平方以上是大户型房子, 60-99是中户型, 60以下是小户型
5. 房龄5年(含)以下是新房子、房龄6-10年是半新房、房龄10年以上是旧房

根据知识词典形成如下概念树:

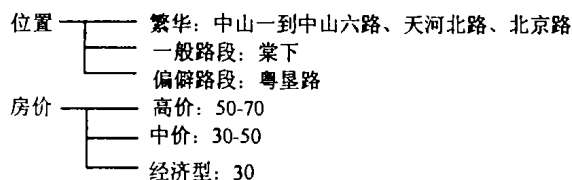


图3 形成概念树

如图4, 把范例1的范例描述和结论分别提取成命题形式, 分别进行概化, 并最终形成规则。

4 模型讨论

4.1 概念归并

范例分为可结构化范例和难以结构化定义范例这两种, 对于可结构化范例, 则可以自动进行概化和分类, 而在这一概念归并过程中, 经常需要计算各类别的相似系数, 相似系数的

计算可以采用以下几种方法来进行:(1)直接计算类别间的距离;(2)数量积法;(3)夹角余弦法。

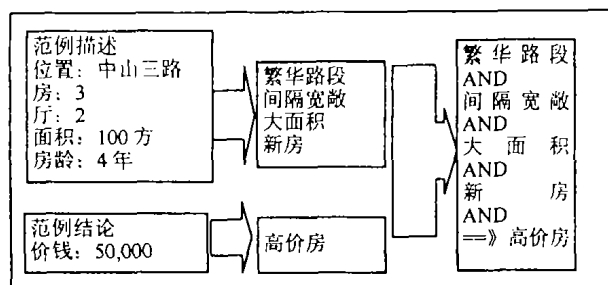


图4 范例到规则的转换过程

对于难以结构化处理的范例，则可以借助知识辞典来直接进行概念分类并概化。在一个指定的领域中，范例的结构尽管复杂但往往是可知的，知识辞典则可以帮助形成范例的多层分类，可操作性强。

4.2 规则处理

具体范例 C_0 本身不存在逻辑不一致，但在把概化范例转换为最终规则 R 过程中，产生出来的规则可能包括冗余规则和逻辑不一致规则，对这些规则需要进行处理，应遵循以下原则：

(1) 冗余规则：各规则或规则组基本有相同的结论。比如 $p \wedge q \rightarrow h$ 和 $p \wedge q \rightarrow h$ ； $A \in \{(2,4)\} \cup \{(3,5)\} \rightarrow g$ 和 $A \in \{(2,5)\} \rightarrow g$ ，这两组规则实际上分别是等价的，故是冗余规则，可以直接从库中删除一个。

(2) 包含规则：当一个规则或一组规则的含义已在另一个规则中表示出来时，可从类似的但约束条件较少的一个规则中得出结论，则删除约束条件较多的规则。

(3) 冲突规则：那些采用了相同（或非常相似）的条件，但导致不同结论的规则（规则组），或其组合违背了逻辑原理（例递推性等）的规则。这些规则应结合领域知识库和常识库进行有条件删除。

(4) 多义规则：根据蕴含规则的要求，当给出完备的规则

前提时应该得出唯一的结论，如果知识库存在在前提相同而结论不同的规则就出现了多义性错误，系统将自动利用知识来进行规则处理，若不能自动处理，则把这些规则显示出来给领域专家进行判别处理。

4.3 知识辞典的构造

一般来说，知识辞典可以以两种方式给出，一种是领域专家人工制定，另一种可由概念空间自生成方法来自动构造，具体可以参考文[5]。

4.4 概化过程的匹配问题

C_{k+1} 是由 C_k 查找知识辞典得到的匹配集合 D_k 生成的，但是并非所有的范例属性都能在知识辞典的概念树上找到匹配的节点，这就要求我们进行一定的处理。例如，对数值属性我们可以进行规定一定的离散化措施，或者根据要求抛弃该属性，或者提交人工处理。

结论 规则在充分利用 CBR 推理功能方面有重要作用，包括 CBR 跟 RBR 的集成推理和 CBR 推理过程中的范例修正都需要用到规则。本文提出了 CBR 系统中由范例提取规则的模型，提出了把范例分解为范例描述和范例结论、分别进行概化、在概化的每个范例模板中进行规则的提取的模型，并对模型中涉及的概念归并、规则处理、知识词典构造、概化过程的匹配等问题进行了讨论。文中提出的模型可以解决范例与规则的衔接问题，为 CBR 与 RBR 集成推理提供了基础。

参考文献

- Blau L, Bonissone P P, Ayub S. Planning with dynamic cases. In: Proc. of Case-Based Reasoning Workshop, Morgan Kaufmann, San Mateo, CA, 1991. 295~306. Expert Systems With Applications 1, 1990. 23~38
- Rissland E L, Skalak D B. CABARET: rule interpretation in a hybrid architecture. International Journal Man-Machine. Studies, 1991. 34: 839~887
- Smyth B, Kean M T. Hierarchical Case-Based Reasoning Integrating Case-Based and Decompositional Problem-Solving Techniques for Plant-Control Software Design. IEEE Transactions On Knowledge And Data Engineering, 2001
- 汤胤, 彭宏, 郑启伦. 基于 XML 的多层次范例映射模型. 计算机科学, 2004, 31(2)
- 傅伟鹏, 吴斌, 何清, 史忠植. 一种概念空间自生成方法. 计算机工程与应用, 2003(7)

(上接第85页)

网络资源。在关键机的实现上，移动感知主要依靠智能 Agent 实体，它在内部提供了一个对外的感知接口，由感知接口来控制感知的获取、分类处理、依据规则过滤以及反馈感知结果等过程。感知接口的代码定义如下：

```

public interface Awareness {
    public void collectInfo(); // 收集
    public void filterInfo(); // 过滤
    public int getType(); // 获得类型
    public void pretreatInfo(); // 预处理
    public void getResult(); // 得到结果
    public void feedback(); // 反馈
}
  
```

感知结果得到以后，将其发送给相对应的接收对象，同时也将这些信息发送给协作服务器，进行感知信息的保存，为了能够给群组用户提供一个感知回放的功能。在技术环节方面，我们运用了感知技术、Agent 技术、XML 解析及中间件技术，给每个协作成员提供了适度的感知反馈，提高了群体间以及用户与任务间的协作能力。而在自适应的技术中使用了基于 Java 和 XML 的 IBM WebSphere MQ Everyplace 中间件技术。MQe 中间件技术适用于便携式的移动设备，例如移动电话、PDA、手提式电脑，同样也适用于服务器和工作站系统。它通过底层对网络良好的支持，可以很好地检测到网络状态。它也提供了一次性的可靠传输机制，而且在网络条件中断的情

况下，自动恢复网络连接。在 MQe 中存在着多种队列，我们利用本地队列的管理机制完全实现了一个基于先进先出算法的 Cache 管理功能。

结束语 移动 CSCW 是计算机支持协同工作与移动计算融合在一起的结果，它代表着未来协同工作的发展趋势，人们能够使用移动设备在任何时间、任何地点进行协同工作和协同研究。本文针对移动 CSCW 的特性和重要实现机制做了详细的分析，并给出了一个移动 CSCW 系统模型，同时在这个模型的基础上，实现了一个移动协同计划助手(MCPA)系统，为移动 CSCW 系统模型提供了一个可以参考的现实范例。

参考文献

- Wiberg M. Exploring Mobile CSCW: Five areas of questions for further research. IRIS23, Sweden, 2000
- Perry, M. OHara, K. Sellen, A. Brown, B. Harper R. Dealing with Mobility: Understanding access anytime, anywhere. ACM Transactions on computer human interaction, 2001
- Zhang Qian, Zhu Wenwu, Zhang Ya-Qin. Resource Allocation for Multimedia Streaming Over the Internet. Special issue on Multimedia over IP in IEEE Trans, on Multimedia, Sept. 2001. 339~355
- Padhye J, Firoiu V, Towsley D, Kurose J. Modeling TCP throughput: A simple model and its empirical validation. In: Proc. SIGCOMM98, Aug. 1998. 303~314
- 赵亚伟, 张海盛, 古乐野. CSCW 环境中动态信息感知技术. 计算机应用, 2003, 23(4)