

基于案例系统的不确定案例索引和检索方法*

王亚英 邵惠鹤

(上海交通大学自动化系 200030)

摘 要 将粗集理论引入基于案例的推理系统,充分利用案例库中冗余属性的简化,形成案例的多个索引,从而可根据不同问题按不同索引进行检索并得出结论。计算实例表明,该方法既能有效地解决不确定问题,又能提高系统的性能。

关键词 基于案例的推理,粗集,案例表达,案例索引,案例检索

分类号 TP 18

Uncertain Case Indexing and Retrieval Method of Case-based Reasoning System

Wang Yaying, Shao Huihe

(Shanghai Jiaotong University)

Abstract Rough set theory is introduced into the case-based reasoning system, which can fully use the reduction of redundancy attributes of case database and make multi-index of case. So it is possible to retrieve case database according to the difference index and arrive at a conclusion for the different question. An example shows that the method can not only solve the uncertainty question effectively, but also improve the performance of the system.

Key words case-based reasoning, rough set, case representation, case indexing, case retrieval

1 引 言

基于案例的推理(CBR)是用以往求解类似问题的以案例形式表示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^[1,2]。传统的 CBR 系统在求解动态、不确定环境下的现实问题时,其案例推理模型面临着如何处理具有不确定属性的难题。

粗集理论是研究不完整数据及不精确知识的表达、学习、归纳的一套方法^[3,4]。本文利用粗集理论对案例库中的属性进行分析,充分利用案例库中冗余属性的简化,得到案例的多个索引进行检索,从而有效地解决了不确定问题。

2 基于案例系统的不确定案例索引和检索方法

传统的案例检索模型一般都要求有确定的问题案例属性值,然而现实问题的环境往往是动态和不确定的,对问题的描述往往也并不完整。此时若用这种检索模型来检索,则可能无效。案例库中常含有冗余属性,粗集理论则能利用数据的冗余性,积极地解决 CBR 系统所面临的难题。在 CBR 系统中,检索过程可分为案例表达、案例索引和案例检索三个阶段。

2.1 案例的粗集表达法

案例库一般可表示成决策表的形式,即 $S = (U, C \cup D)$,如表 1 所示。其中, U 是有限集合,表示案例名; C 表示条件属性; D 表示决策属性(一般只

* 1999-06-21 收稿,1999-09-06 修回

有一个)。在决策表中,表的行代表研究的案例对象,列代表其属性,其值代表对应案例的对应属性值,每行对应一个案例。案例的这种表达结构,既符合人们的习惯,且便于存储,也适合于粗集处理。

2.2 案例的粗集索引法

采用索引法进行检索的关键是建立合适的索引。利用粗集理论建立案例的索引,不需要任何先验知识就能从数据中很好地归纳出确定的索引特征(对于含有噪声的案例,可采用变精度粗集模型^[5]来处理),算法简单易行。不失一般性,设案例库的属性均为定性属性(如果是定量属性,则可通过各种离散化方法^[6,7]将其离散化),具体算法如下:

1) 依据粗集理论把案例库分别基于各个属性分成若干等价类,并记下其描述;

2) 利用粗集并行分析法^[8]先求得其核属性,即索引中的关键属性,这些属性在对案例进行诊断分析时是缺一不可的(核属性可能为空);然后求得案例库中条件属性集 C 的各个 D 简化,从而得到案例的多个索引。

2.3 案例的粗集检索法

案例的检索就是利用待解问题已知的有限信息(某几个属性的取值),在案例库中找到与当前情形相似的案例。案例的粗集检索算法如下:

1) 对于待解问题,检查其核属性的值是否确定,如果不确定,则说明条件不充分,不能得到确切的结果,结束;否则转 2);

2) 找到包含于待解问题的已知属性集的案例索引,如果有多个,则用户可根据需要选择其中一个属性集(缺省情况下取其所属属性数目最少的案例索引);

3) 对于求得的案例索引中的每一个属性,根据待解问题中的对应属性的取值,判断其归属于基于该属性所划分的等价类中的哪一类;

4) 求得关键属性的等价类的交集,如果为空集,则说明没有具有确定性解的相似案例集存在,只能由用户做出答案,并将其添加到案例库中,结束;否则转 5);

5) 求出该交集与其它属性的等价类的交集,如果为空集,则说明在这个属性集中没有相似案例,转 2),重新选择案例索引;否则交集即为待解问题的相似案例集,转 6);

6) 从基于决策属性 D 所划分的等价类集中,找到包含相似案例集的等价类,其描述即为所求问题

的结论。

3 应用示例

现以文献[1]的例子来说明本文方法的推理过程。表 1 为一个简化油气储集层伤害类型诊断案例库,其条件属性有 6 个,均为离散化属性;决策属性 D 有 3 个取值: $R_1 =$ 外来颗粒堵塞, $R_2 =$ 相对渗透率下降, $R_3 =$ 添加剂不伍;而问题(即案例 7)有 2 个不确定属性: $\{C_5, C_6\}$ 。

表 1 储集层伤害类型案例

案例号	温度 C_1	渗透率 C_2	压差 C_3	流量 C_4	孔隙度 C_5	管注状况 C_6	结论 D
1	高	低	大	小	中	一般	R_1
2	高	低	小	小	大	清洁	R_2
3	低	中	大	中	小	不清洁	R_1
4	低	高	中	小	中	一般	R_3
5	中	低	小	大	中	清洁	R_2
6	中	中	小	大	小	一般	R_2
7	高	中	小	中	不确定	不确定	?

具体算法如下:

1) 依据粗集理论把案例分别基于各个属性分成若干等价类,并记下其描述。

$U/\text{ind}(C_3): C_{31} = \{1, 3\}, \text{des}\{C_{31}\} = \{C_3 = \text{大}\}, C_{32} = \{2, 5, 6\}, \text{des}\{C_{32}\} = \{C_3 = \text{小}\}, C_{33} = \{4\}, \text{des}\{C_{33}\} = \{C_3 = \text{中}\}; U/\text{ind}(D): D_1 = \{1, 3\}, \text{des}\{D_1\} = \{D = R_1\}, D_2 = \{2, 5, 6\}, \text{des}\{D_2\} = \{D, R_2\}, D_3 = \{4\}, \text{des}\{D_3\} = \{D = R_3\}。$

限于篇幅,只列出属性 C_3 和 D 分成的等价类集,其它略。

2) 经过粗集并行分析,得到 C 的 D 核为空,而 C 的 D 简化分别为: $\{C_3\}, \{C_1, C_5\}, \{C_1, C_6\}, \{C_2, C_4, C_5\}$ 和 $\{C_2, C_4, C_6\}$, 即得到 C 相对于 D 的多个索引。

3) 对于不确定问题(案例 7),由于属性 C_5, C_6 不确定,则只能选用索引 $\{C_3\}$ 。

4) 因为 $C_3 = \text{小}$,故得到其相似案例集为 $C_{32} = \{2, 5, 6\}, \text{des}\{C_{32}\} = \{C_3 = \text{小}\}。$

5) 因为 $C_{32} \cap D_2 = C_{32}$,故最终诊断结果为 $\text{des}\{D_2\} = \{D = R_2\}$,即相对渗透率下降。

所得结果与文献[1]的诊断结果是一致的,但本文方法不需要任何先验知识,简单实用,能够处理各种情况下的不确定性问题。

4 结 语

本文将粗集理论应用到基于案例的推理系统,充分利用案例库中冗余属性的简化,形成案例的多个索引进行检索,不仅有效地解决了具有不确定性问题,而且提高了系统的性能。该方法不需要任何先验知识,简单易行,为基于案例的推理系统提供了一条新的途径,具有广阔的应用前景。

参 考 文 献

- 1 方明,李天太,杨军全,等. 基于案例的不确定检索模型的研究. 控制与决策, 1999, 14(1): 93~96
- 2 袁小红,王珏. 基于案例的推理——综述与分析. 模式识别与人工智能, 1995, 8(12): 18~29
- 3 Z Pawlak. Rough sets, theoretical aspects of reasoning about data. Nowowiejska 15/19, Waeaaw. Poland, 1990
- 4 曾黄麟. 粗集理论及其应用. 重庆:重庆大学出版社, 1996

- 5 Wojciech Ziarko. Variable precision rough set model. J of Computer & System Science, 1993, 46(1): 39~59
- 6 李天伟,戴建设. 基于案例系统的一种案例标引和获取方法及算法. 控制与决策, 1997, 12(1): 25~29
- 7 J Dougherty, R Kohavi, M Sahami. Supervised and unsupervised discretization of continuous features. In: Proc of the 12th Int Conf on Machine Learning. San Francisco: Morgan Kaufmann Publishers, 1995. 194~202
- 8 J W Guan, D A Bell. Rough computational methods for information systems. Artificial Intelligence, 1998, 105: 77~103

作 者 简 介

王亚英 女, 1973年生。上海交通大学自动化系博士生。研究领域为人工智能与智能控制,粗集理论及应用等。

邵惠鹤 男, 1936年生。上海交通大学教授,博士生导师。研究方向为工业过程模型与优化控制,智能控制理论研究及应用等。

(上接第746页)

作 者 简 介

代冀阳 男, 1966年生。南昌航空工业学院讲师,北京航空航天大学第七研究室博士研究生。主要研究方向为鲁棒控制,智能控制及其应用。

孙秀霞 女, 1962年生。空军工程大学工程学院副教授,博士。主要研究方向为鲁棒控制, H 控制及其应用。

毛剑琴 女, 1940年生。北京航空航天大学第七研究室教授,中国自动化学会常务理事,博士。主要研究领域为智能控制与智能辨识理论及应用,鲁棒控制与自适应控制理论及应用。

林 岩 男, 1955年生。北京航空航天大学第七研究室副教授,博士。主要研究领域为鲁棒与自适应控制。

(上接第749页)

参 考 文 献

- 1 C Padgett, K K Delgado, S Udomkesmalee. Evaluation of star identification techniques. J of Guidance, Control and Dynamics, 1997, 20(2): 259~267
- 2 P Kudva, A Throckmorton. Preliminary star catalog development for the earth observation system AMI (EOS-AMI) mission. J of Guidance, Control and Dynamics, 1996, 19(6): 1332~1336
- 3 潘正君,康立山,陈毓屏. 演化计算. 北京:清华大学出版社, 1998. 132~147
- 4 M S Scholl. Star-field identification for autonomous attitude determination. J of Guidance, Control and Dynamics, 1995, 18(1): 61~65

作 者 简 介

李立宏 男, 1971年生。1994年在哈尔滨工业大学控制工程系获学士学位,现为该系博士研究生。主要研究方向为计算机网络,导航技术及人工智能。

宋申民 男, 1968年生。1996年在哈尔滨工业大学控制工程系获工学博士学位,现为该系副教授。主要研究方向为鲁棒控制,神经网络、遗传算法在智能控制中的应用等。

张福恩 男, 1936年生。1961年毕业于哈尔滨工业大学,现为该校控制工程系教授,博士生导师,美国数学学会会员。主要研究领域为仿真转台技术与设备,导航技术与设备,计算机控制工程等。