文章编号:1005-0523(2012)02-0042-05

一种基于Rough集的案例推理模型的构建

龚锦红[1],凌仕勇[2]

(华东交通大学 1. 电子与电气工程学院; 2. 现代教育技术中心,江西 南昌 330013)

摘要:利用Rough集理论处理案例推理问题具有不需要外界信息和先验知识的优点,对案例库中冗余属性进行简化,能够起到优化案例库的作用,同时能够依赖于统计知识提炼规则并形成多个有效的案例索引,在进行案例检索时可针对不同的检索问题选择恰当的索引快速检索到相似的案例,并进行推理得出相应的问题解决方案。最后,以稀土萃取分离生产过程的产品纯度和料液处理量等生产指标的智能优化设定控制为例,验证了该模型的可行性和精确性。

关键词:案例推理;Rough集;数据补全;数据离散;属性约简

中图分类号:TP311 **文献标志码**:A

案例推理(case-based reasoning,CBR)技术是近年来人工智能领域中兴起的一项新兴的推理技术[1],其通过检索历史案例,充分利用以前的经验,推理得到新问题的解决方案。一个典型的案例推理过程的基本步骤包括4个主要过程:案例检索、案例重用、案例修正和案例保存。通常,案例的恰当表示、合理组织、获取的有效性以及案例检索的有效性及快速性关系到问题求解的效率和质量。

Rough集理论[2]由波兰数学家Pawlak在1982年提出,以处理含糊和不精确性问题。Rough集理论在处理含糊信息方面,具有不需要外界信息和先验知识的独特优点,能够依赖于统计知识提炼规则,有效地解决实践中遇到的不精确性属性难题,在对问题的计算分析方面具有客观性。因而,应用Rough集理论对案例推理系统具有积极的作用。

1 基于Rough集理论的案例推理模型

在解决问题时,可借助于该类问题的历史经验进行推理得出该问题的解决方法。在案例推理系统中,对某个问题的表述及解决方法通常用一个或多个案例来表达,这些案例按一定的结构和模式组织在案例库中。当某类新问题出现时,此推理系统依据相关的事实和索引,在案例库中检索得到相似的案例,并对其求解策略进行分析,重用,得到此问题的求解;最后,将此问题及相应的求解模式和策略作为一个新的案例追加到系统的案例库中,供日后所需。案例库是存储过去案例的空间,随着时间的推移,案例库中存储的案例不断增加,会存在失败的案例及冗余的案例,使得案例推理系统的推理质量和效率大大降低。应用Rough理论可以对案例库进行有效的知识约简,在优化案例库的同时,提高案例推理的效率和质量,积极地解决案例推理系统所面临的难题[3]。

1.1 案例的表示

在Rough集理论中,案例库是以一个二维表格的决策系统形式来表示的,即$\boldsymbol{S}=(\boldsymbol{U},\boldsymbol{A},\boldsymbol{V})$。其中,U定义为非空有限集,即案例全域;$\boldsymbol{A}$代表案例的条件属性,是案例的描述特征属性集,$\boldsymbol{A}=\{f_1,f_2,\cdots,f_n\}$,$f_1\ f_2\cdots f_n$为不同的案例特征属性;$\boldsymbol{V}$代表案例的决策属性,即解特征属性集,$\boldsymbol{V}=\{f_{S1},f_{S2},\cdots,f_{Sn}\}$,其中$f_{S1}\ f_{S2}\cdots f_{Sm}$为不同的案例解特征属性。在决策表中,行代表研究的案例对象,列代表其属性,行代表案例库中的某个案例。

收稿日期:2012-01-25

基金项目:国家自然科学基金项目(50474020);江西省教育厅青年科学基金项目(GJJ11115)

作者简介:龚锦红(1976—),女,硕士,讲师,研究方向为复杂工业过程的建模与优化控制。

1.2　案例库的预处理

应用Rough集理论在案例推理系统中时，为了优化案例库，并处理不完备数据及不精确知识表达的问题，在案例检索前须采用Rough Set对案例库进行初步处理，如对初始数据的补全、离散化处理以及对案例知识进行约简等等[4]。

1.2.1　数据补全

形式上描述为：案例库中定义的$\boldsymbol{S}=(\boldsymbol{U},\boldsymbol{A},\boldsymbol{V})$，进行补全后的数据集$\hat{\boldsymbol{U}}=\{x\in\boldsymbol{U}|\forall a\in\boldsymbol{A},a(x)\neq T\}$，$T$为缺失数据，集合$\boldsymbol{A}$为全集$\boldsymbol{U}$中的属性集合。

数据补全简单的有均值补全和模式补全，均值补全是，若其属性数据是数值类型的，则取属性均值为补全后的数据。若其属性数据是字符类型的，则取出现概率次数最多的属性值作为补全后的数据。

1.2.2　数据离散化

采用Rough集对案例和案例库进行处理时，要求案例库中的属性值必须用离散数据进行表示。但是，实际中数据多数是连续的，因此，必须首先离散化处理连续的数据，然后再进行相关分析[5]。对联系数据进行离散化应满足：①信息处理复杂度应尽可能小，即离散后属性维数尽量小，也就是属性值种类尽量少；②信息熵丢失少，即离散后的属性值信息丢失应尽量少。

1.2.3　知识约简

知识约简是Rough集理论的核心内容之一，就是在保持知识库的分类和决策能力不变的条件下，删除一些无关或多于的信息。设$\boldsymbol{C}$表示数据的集合，ω($\boldsymbol{C}$)是$\boldsymbol{C}$的权重集，$\boldsymbol{D}$是最后得到的约简集，算法描述如下：

① 设$\boldsymbol{D}=\varphi$。② 设b是有最大åw($\boldsymbol{C}$)的属性，ω($\boldsymbol{C}$)表示集合$\boldsymbol{D}$的没有放入的，集合$\boldsymbol{C}$的权重集总和。③ 将b加入到$\boldsymbol{D}$中。④ 从S中移除包含b的属性集。⑤ 若$S=\varphi$，得到最终的最小约简，返回集合$\boldsymbol{D}$，否则转到②继续执行。

例如，对于对象C={{fish，cat ，dog }，{cat，man}，{ man，dog }，{ cat，fish }}，令$\omega=1$。首先设定$\boldsymbol{D}=\varphi$，因为cat是集合$\boldsymbol{C}$中最经常发生的属性，把cat加入到$\boldsymbol{D}$中且需从集合$\boldsymbol{C}$中移除含cat的项，这样$\boldsymbol{C}$={{ man，dog}}；继续上面发现dog是接下来的集合中最经常发生的属性，将dog加至集合D中且移除含dog的项，最后，$\boldsymbol{C}=\varphi$，得到集合$\boldsymbol{D}$= {cat，dog}。

1.3　案例库的构建

案例库构建是对案例进行处理的中心，首先利用类DTImporter从基础数据库中导入初始数据；然后利用GetMethod（SCALER）对数据进行预处理和离散化处理；接下来采用Johnson约简方法，即GetMethod（JOHNSONREDUCER）对数据进行约简；最终产生系统的推理规则。最后写入数据库或内存中，供以后的浏览案例库，维护（增，删，改）案例库及测试使用。

案例库构建的代码片段如下[6]：

```
Method* method;
//数据导入，决策表方法（DECISIONTABLE）
DTImporter dtIm;
method=& dtIm;
SetStruc（Creator::Create（DECISIONTABLE））;
GetStruc（）.Apply（*method）;
//数据离散化（SCALER）
method =GetMethod（SCALER）;
GetStruc（）.Apply（*method）;
//知识约简（JOHNSONREDUCER）
method=GetMethod（JOHNSONREDUCER）;
```

```
method->SetParam ("DISCERNIBILITY = Object;SEED = 1111");
GetStruc ().Apply(*method);
//规则产生(RULEGENERATOR)
method = GetMethod(RULEGENERATOR);
GetStruc ().Apply(*method);
```

1.4 案例检索

应用Rough集进行案例检索的基本过程为:对某个新问题,根据模型找出案例属性的索引,求取属性集合中等价类属性集合的交集,通过检索找出相似的案例;若没有从案例库中找到相似的案例,则重新选择较合适的案例索引进行检索。案例的Rough集检索算法流程大致如图1所示[7]。

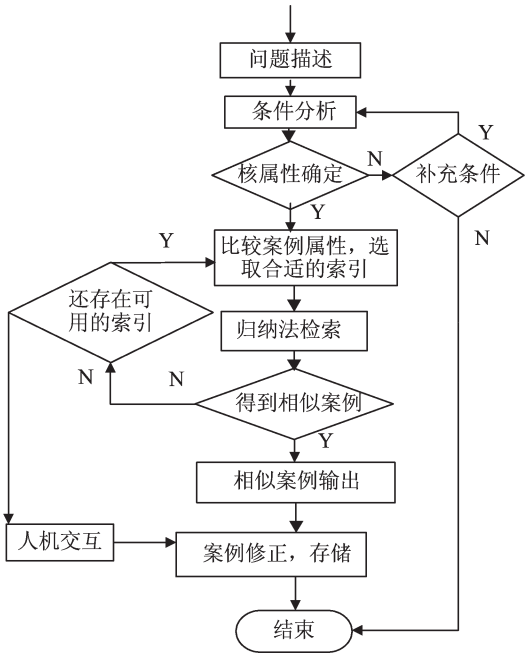

图1 案例推理系统的Rough集检索算法流程图

Fig.1 Rough set retrieval algorithm flow chart of case reasoning system

2 系统应用示例

现以文献[8]中稀土萃取分离生产过程的料液处理量,产品纯度等指标的优化控制为例,将Rough集推理与案例推理相结合,对稀土萃取分离过程中的萃取剂、料液和洗涤液的流量值通过构造案例,将稀土萃取分离过程的优化设定控制问题转变为对案例模型的分析,案例库的构建,案例的检索,重构,复用及对案例库的增删改过程,从而确定实际过程的动态模型。

由案例推理系统得到稀土萃取分离过程中的检测点组分含量 $f'_{B,K}$ 、洗涤剂流量 V_W 、有机萃取剂流量 $\bar{V}_S$,可选择水相液组分含量 f_B 和水相液流量 V_F 为描述特征,即案例的条件属性,其它三个待定参数为案例的解特征,即案例的决策属性。构造案例如下:

条件属性 $A=\{f_1,f_2\}=\{V_F,f_B\}$

决策属性 $V=\{f_{S1},f_{S2},f_{S3}\}=\{\bar{V}_S,V_W,f'_{B,K}\}$

将现场采集到的30组数据作为测试集如表1所示,针对该推理系统进行精确性和可行性的测试。

表1 测试集数据

Tab.1 Testing set data

序号	料液流量/ml·min^{-1}	料液组分含量/100%	萃取剂流量/ml·min^{-1}	洗涤剂流量/ml·min^{-1}	检测点组分含量/100%
1	3.04	0.546	32.15	4.70	0.863
2	2.51	0.528	27.83	4.248	0.824
…	…	…	…	…	…
15	2.02	0.505	20.37	3.05	0.657
…	…	…	…	…	…
29	2.01	0.478	22.1	3.44	0.663
30	4.55	0.476	47.5	7.35	0.667

输入为单个测试集的案例工况描述特征,输出为实际解特征和基于Rough集的归纳法检索方法得到的解特征如图2所示。

案例检索结果与实际采样数据比较，可发现，经案例检索后导出的特征数值与实际值非常接近，各项误差小于2%。

对整个测试集的30组数据分别采用基于欧拉距离的近邻检索法和基于Rough集的归纳法，两种检索结果和实际值绘制成对比图。以检测点组分含量测试对比结果为例绘制测试对比图如图3所示。

采用近邻法和归纳法分别得到这两种检索方法的解结构特征和实际值解特征的误差对比图。此处以检测点组分含量测试误差为例绘制对比图如图4所示。

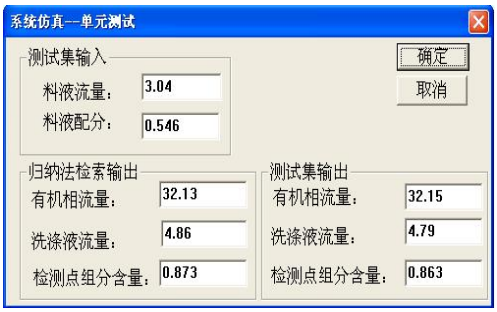

图2　单元测试归纳法检索结果

Fig.2　Retrieval results of unit testing result induction

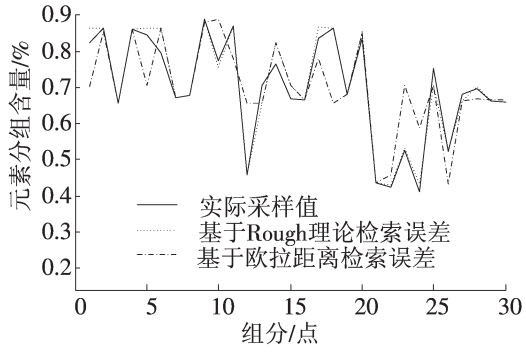

图3　检测点组分含量测试对比图

Fig.3　Testing contrast of component content on check point

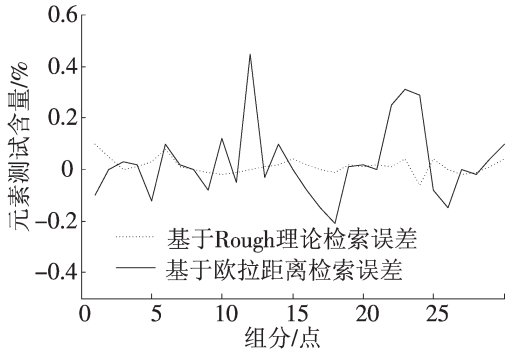

图4　检测点组分含量测试误差对比图

Fig.4　Testing error contrast of component content on check point

从图中可以看出，两种检索结果和实际值之间具有一定的误差：Rough集推理模型与实际值的误差范围比较小，基本围绕0%这个中心点上下浮动，最多10%；而基于欧拉距离的模型与实际值的误差范围较大并且极不平均，曲线呈不规则的形状有些误差达到了50%。从对比的结果可以看出这种案例推理模型相对于基于欧拉距离的模型更接近于实际工程数值，在案例推理过程中具有很强的可行性和精确性。

3　结论

结合Rough集对案例知识进行推理，在对案例进行分析形成案例模型，构造Rough集的属性集合，进而形成Rough集案例模型，根据不同问题采用不同方案进行快速检索，能有效地解决实际工况中的不完备和不确定性问题。实验证明，该推理模型构建方法对解决实际问题是客观可行的。

参考文献：

[1] WATSON I，MARIR F. Case-based reasoning：A review[J]. The Knowledge Engineering Review，1994，9(4)：335-381.

[2] PAWLAK Z. Rough sets[J]. International Journal of Computer and Information Seience，1982(11)：341-356.

[3] HOANG XUAN HUAN. Case-based reasoning with rough features[J]. Knowledge-Based System. 2003(4)：321-327.

[4] DUNTSCH I，GEDIGA G. Statistical evaluation of rough set dependency analysis[J]. International Journal of Human Computer Study，1997，46(5)：589-604.

[5] LINGRAS P. Application of rough pattern，Rough set in datamining and knowledge discovery[J]. Series Softcomputing，Physical Velag (Springer)，1998(2)：369-384.

[6] 龚锦红. 基于案例推理的稀土萃取分离过程优化设定控制方法研究[D]. 南昌：华东交通大学，2007：221-222.

[7] 龚锦红，杨辉，衷路生. 稀土萃取分离过程中的Rough集案例推理方法[J]. 第二十九届中国控制会议(No 29，CCC'10)，2010(7)：1701-1706.

[8] 杨辉，柴天佑. 稀土萃取分离过程的优化设定控制[J]. 控制与决策，2005，20(4)：398-407.

A Model of Case-based Reasoning Based on Rough Set

Gong Jinhong[1], Ling Shiyong[2]

(1. School of Electrical and Electronic Engineering; 2. Center of Modern Education and Technology, East China Jiaotong University, Nanchang 330013, China)

Abstract: Rough set theory has the unique merit of having no use for the outside information and the priori knowledge when dealing with the Case-based reasoning (CBR) problem, which can simplify the redundancy attribute in the case library to optimize the case database. Many indexes of case library are formed based on statistical knowledge at the same time. It is possible to retrieve case database according to the difference index and draw a conclusion for the different question. Finally, it verifies the feasibility and the accuracy of this model based on the example of the intelligent optimal setting control for the production indexes just as the product purity and the liquid flow control in the rare earth extraction separation process.

Key words: case-based reasoning; rough set; data completion; data dispersion; attribute reduction

(上接第35页)

Design and Realization of Virtual Driving Simulation System for CRH3 EMU

Wu Haichao[1], Zhang Anquan[2]

(1. Nanjing Institute of Railway Technology, Nanjing 210031, China; 2. Zhengzhou J & T Hi-tech Co. Ltd., Zhengzhou 450001, China)

Abstract: Adopting a number of high technology, driving of the CRH3 EMU train faces great challenges. The development of virtual driving simulation system of CRH3 EMU aims to train the driver's using skills of the EMU driving, emergency fault handling and abnormal driving by multimedia teaching. The virtual driving simulation system of CRH3 EMU conducts the full 3D virtualization train logical model simulation to CRH3 driving with the virtual reality technology to show the invisible, non-touching, non-entering train structure, layout and connection mode. It can easily set the fault that often does not meet in reality but is very troublesome to set with the real ways. It can repeatedly set, find and eliminate electric equipment system fault, which improves the training efficiency and reduces the training cost and solves the difficult problem of training the drivers on the spot of high speed railway with good effect.

Key words: CRH3 virtual simulation; driving simulator; fault handling; abnormal driving