文章编号:1005－0523(2003)04－0091－04

基于粗糙集理论的数据算法在模具评价中的应用

卢晓春[1], 阮 锋[2], 郭 盛[2]

(1.广东交通职业技术学院 广州 510650;2.华南理工大学 广州 510640)

摘要:粗糙集理论是一种软计算方法,可以有效地分析和处理不完备信息.介绍了粗糙集理论的产生和发展,粗糙集理论概述,粗糙集理论的特点,研究了基于粗糙集的数据发掘算法 CRCG,并将其应用于模具设计质量评价.

关 键 词:粗糙集;数据发掘;模具设计;质量评价

中图分类号:TP18 **文献标识码**:A

粗糙集理论是一种刻划不完整性和不确定性的数学工具,能有效地分析和处理不精确、不一致、不完整等各种不完备信息,并从中发现隐含的知识,揭示潜在的规律.该理论提出的上下近似、核、简化等概念,为数据分析、决策判断等提供了新的理论和方法.本文讨论了粗糙集的基本理论,在此基础上研究了一种基于粗糙集的数据发掘算法(CRCG),并将其应用于模具设计质量评价.

1 粗糙集理论概述

20 世纪 70 年代初,波兰学者 Z.Pawlak 和波兰科学院、华沙大学的逻辑学家们组成了研究小组,开始了对信息系统逻辑特性的长期基础性研究.针对从实验中得到的以数据形式表述的不精确、不确定、不完整的信息和知识,进行了分类分析.1982 年 Z Pawlak 发表了经典论文 Rough Sets[1],第一次提出了粗糙集理论.1991 年,Z Pawlak 出版的专著"Rough Sets"[2]成为粗糙集理论研究的里程碑,1992 年粗糙集应用专集[3]的出版,对这一时期粗糙集的理论和应用方面的工作成果进行了总结,促进了粗糙集在机器学习、知识获权、决策分析、过程控制等许多领域的应用[4].

粗集的概念涉及到如下几个定义[5]:

令 X⊆U,且 R 为一等价关系,当 X 为某些 R 基本范畴的并时,称 X 是 R 可定义的,否则 X 为 R 不可定义的.R 可定义集也称为 R 精确集,可在知识库 K 中被精确定义,这时集合 X 称为 K 中的精确集;而 R 不可定义集也称为 R 非精确集或 R 粗集,不能在这个知识库中被定义,这时集 X 称为 K 中的粗集.

粗集使用两个精确集(上近似集和下近似集)来描述:

假设给定知识库 K=(U,R),对每个子集 X∈U 和一个等价关系 R,R∈ind(K),其中 ind(K)为 K 的不可分辨关系,可以根据 R 的基本集合的描述来划分集合 X.

两个子集:

$$R_-(X)=\{x\in U:[x]_R\in X\}$$

$$R^-(X)=\{x\in U:[x]_R\in X\}$$

其中$[x]_R$表示子集 x 属于 R 中的一个范畴.

子集 $R_-(x)$称为 X 的 R 下近似,$R_-(X)$是对知识 R、U 中可归入 X 的元素的集合;而 $R^-(X)$称为 X 的 R 上近似,是对于知识 R、U 中所有一定能归入 X 的元素的集合.集合 $bn_R(X)=R^-(X)-R_-(X)$称为 X 的 R 边界,是对于知识 R 既不能归入+X,也不能归入－X 元素的集合,+X 是指可由前提属性确定

收稿日期:2003－04－21

作者简介:卢晓春(1965－),女,江西南康人,广东交通职业技术学院副教授.

地得出结论属性为真的元素的集合，－X 是指可由前提属性确定地得出结论属性为假的元素的集合.

在粗集中，将一个对象对一个集合的从属关系称为成员关系.成员关系依赖于我们的知识，而且成员关系不是绝对的.

在粗糙集理论中，为了使对象的知识可以方便的以数据表格形式描述，是通过指定对象的基本特征(属性)和它们的特征值(属性值)来描述对象的知识[6]，称为知识表达系统.即定义一个知识表达系统为：

$$S=(U,C,D,V,F)$$

式中，U 是对象的集合，$C\cup D=A$ 是属性集合(等价关系集合)，子集 C 和 D 分别称为条件属性和结果属性，V 是属性值的集合，V_a 表示属性 $a\in A$ 的范围；F 是一个信息函数，指定 U 中每一个对象 x 的属性值.

2 一种基于粗糙集的数据发掘算法(CRCG 算法)

目前，进行不确定信息的数据挖掘常用的方法有模糊集和概率统计等方法.这些方法需要一些数据的附加信息或先验经验知识，如模糊隶属函数和概率分布等，这些信息有时并不容易得到.相对而言，在处理不确定信息方面粗糙集理论有其独特的优点[7].1) 粗糙集分析方法仅利用数据本身提供的信息，不需要先验知识.2) 粗糙集是一个强大的数据分析工具.它能表达和处理不完备信息；能在保留关键信息的前提下对数据进行化简并求得知识的最小表达；能识别并评估数据之间的依赖关系，揭示出概念简单的模式；能从经验数据中获取易于证实的规则知识.3) 粗糙集以不可分辨关系为基础，侧重分类，可以用一对清晰集合逼近.

基于粗糙集理论进行知识发现和数据发掘的研究方法包括：分类、回归、聚类、归纳等等.本文主要讨论的是分类问题.

选择与任务相关的目标数据集，设该数据集可用关系模式 $R(a_1,a_2,\cdots a_n)$ 表示，其中 $a_i(i=1,2,\cdots,n)$ 为属性，并且每个属性表示的概念可用一种概念层次数表示.

每个属性对应的概念层次数具有如下的特点，数的根为属性名，数的所有叶子节点对应所有原始数据，并表示原始级别的概念；以树的每个内节点为根的子树覆盖该内节点表示的概念所对应的所有实例.这样，每个属性所对应的概念层次树就形成了对所有实例所构成的集合的不同精细程度的划分；对较高层次的划分，形成的等价类数量就较少，表示概念的泛化程度也较高.在概念层次树中，一个节点的父节点表示的概念即是该节点表示的概念的泛化；一个节点的子节点表示的概念即是该节点表示的概念的一个特化.

在分类问题中，将所有属性分为两组，即条件属性与决策属性，分别以 $c_1\cdots c_n$ 和 $d_1\cdots d_m$ 表示[8].则分类规则的一般形式如下：

IF $(c_1=I_1)\wedge\cdots\wedge(c_n=I_n)$ THEN $(d_1=J_1)\wedge\cdots\wedge(d_m=J_m)$ 其中 $I_i,J_j(i=1\cdots n;j=1,\cdots m)$ 表示集合，‘＝’表示 in.根据前面的分析，有如下的定理.

定理 1：如果规则“IF $(c_1=I_1)\wedge\cdots\wedge(c_n=I_n)$ THEN $(d_1=J_1)\wedge\cdots\wedge(d_m=J_m)$”成立，则对任一 $k=1,\cdots n$，在 c_k 对应的概念层次树中，如果 c_k 是内节点，则必有 c_k 的子节点 C_k 使规划“IF $(c_1=I_1)\wedge\cdots\wedge(C_k=I_k)\wedge\cdots\wedge(c_n=I_n)$ THEN $(d_1=J_1)\wedge\cdots\wedge(d_m=J_m)$”成立.

上述定理可以推广到多个条件属性的情形.该定理表明了对应条件属性的概念层次树的不同层次发掘的规则的关系.同样，根据概念层次树的构造，对决策属性在概念层次树的不同层次上对应的规则而言，有如下定理.

定理 2：如果规则“IF $(c_1=I_1)\wedge\cdots\wedge(c_n=I_n)$ THEN $(d_1=J_1)\wedge\cdots\wedge(d_m=J_m)$”成立，且在决策属性 d_k 对应的概念层次树中 D_k 是 d_k 的父节点，则如下规则也成立：

IF $(c_1=I_1)\wedge\cdots\wedge(c_n=I_n)$ THEN $(d_1=J_1)\wedge\cdots\wedge(D_k=J_k)\wedge\cdots\wedge(d_m=J_m)$

根据上面的分析，针对巨量、高维数据库，采用如下的发掘策略有助于快速发掘感兴趣的模式.

算法 CRCG(Classification based on Rough sets and Concept Generalization)

1) Select Task Relevant Dataset

2) New Dataset＝Transform(Dataset)

3) Rule Set＝Rought It(New Dataset)

4) IF Rule Set Interested or at Prime Concept Level THEN Stop ELSE Goto Step 2

算法说明：

1) 根据发掘要求，选择相关数据集.

2) 根据属性对应的概念层次树，把相关数据集转换成指定的层次概念数据表示.

3) Rough It 为一粗糙集过程,它由三部分构成:属性的约简;元组的约简;导出规则.

4) Interested 一方面可以是预先定义好的层次树,即用户希望在什么层次上发掘,另一方面也可以在发掘到感兴趣的规则时由用户直接干预而终止继续发掘;当发掘到原始数据级上时,不能再进一步往下进行发掘了,整个发掘过程也停止.

5) 一般地,开始时,在高层概念上进行发掘,当发掘到某些感兴趣的模式,且希望发掘更粗细的模式时才逐渐往更低的概念层次上进行发掘.这是因为:在高层概念上对属性进行概念泛化后得到的数据集将大大压缩,从而使发掘速度大大提高;根据定理 1、2 可知,如果在高层概念层次上导出一条规则,则可以对这条规则进行细化,从而得到更精细的规则,即逐步求精.

3　CRCG 算法在模具设计质量评价中的应用

模具设计的合理性对模具质量的好坏起着决定性的作用,模具质量的过程控制也主要集中于模具设计质量的控制.影响模具某一质量特性的因素可能有很多,且它们的影响程度都不一样,而在模具质量数据库中我们很难从庞大繁杂的数据记录中找出各因素的影响.所以,可以采用 CRCG 数据挖掘算法,将影响模具质量的各因素进行科学的分类,分析和处理不精确、不一致、不完整等各种不完备信息,去除冗余的因素,并从中发现隐含的知识,揭示潜在的规律,找出主要的和决定性的影响因素.

例如某一塑料模具设计,其中的几项质量特性如表 1 所示(也称作属性值表).

表 1　某模具设计属性值表

设计指标 / 设计案例	浇口尺寸 (a)	补料时间 (b)	型腔塑料压力相比额定锁模力(c)	塑料件溢边跑料(d)
P1	0	1	1	1
P2	1	0	1	1
P3	1	1	2	1
P4	0	1	0	0
P5	1	0	1	0
P6	0	1	2	1

其中数据的意义及各属性的概念层次树如下:

浇口尺寸(a):0—合理,1—不合理;

补料时间(b):0—合理,1—不合理;

型腔塑料压力相比额定锁模力(c):0—小,1—大,2—太大;

塑料件溢边跑料(d):0—否,1—是;

$$a\begin{cases}0\\1\end{cases}\quad b\begin{cases}0\\1\end{cases}\quad c\begin{cases}0\\1\\2\end{cases}\quad d\begin{cases}0\\1\end{cases}$$

对于表中数据进行粗糙集分析:

设计案例 p^2、p^3 和 p^5 相对于属性“浇口尺寸”是不可分辨的;案例 p^2 和 p^5 相对于属性“浇口尺寸”、“补料时间”和“型腔塑料压力相比额定锁模力”是不可分辨的;案例 P3 和 P6 相对于属性“补料时间”和“塑料件溢边跑料”是不可分割的.由于案例 p^2 出现溢边跑料,而案例 p^5 没有出现,对于属性“浇口尺寸”、“补料时间”和“型腔塑料压力相比额定锁模力”来说,它们是不可分辨的.因此,塑料件溢边跑料不能以属性“浇口尺寸”、“补料时间”和“型腔塑料压力相比额定锁模力”作为特征进行描述.即 p^2 和 p^5 不能根据有效知识进行适当的分类,它们是边界实例.根据其它案例的特性,p^1、p^3 和 p^6 可分类成塑料件溢边跑料,所以案例集合中“塑料件溢边跑料”的下逼近集合是$\{p^1,p^3,p^6\}$,上逼近集合是$\{p^1,p^2,p^3,p^5,p^6\}$.同样,p^4 没有出现溢边跑料;p^2 和 p^5 不能排除出现溢边跑料,所以“没有出现溢边跑料”这个概念的下逼近是$\{p^4\}$,上逼近是$\{p^2,p^4,p^5\}$.可见,为了确定是否出现溢边跑料,不必使用表 1 中的所有属性,表 1 可简化为表 2:

表 2　某模具设计属性简化表

设计指标 / 设计案例	型腔塑料压力相比额定锁模力(c)	塑料件溢边跑料 (d)
P1	1	1
P2	1	1
P3	2	1
P4	0	0
P5	1	0
P6	2	1

由表 2 知,如果在一个案例中型腔塑料压力相比额定锁模力太大,一定会出现塑料件溢边跑料,但如果型腔塑料压力相比额定锁模力小,那就一定不会出现塑料件溢边跑料.

可得六条规则,其中:

R1、R2:IF(c=1)THEN(d=1)

R5:IF(c=1)THEN(d=0)

R4:IF(c=0)THEN(d=0)

R3、R6:IF(c=2)THEN(d=1)

根据各属性对应的概念层次树可知，针对型腔塑料压力相比额定锁模力大(c=1)，可将属性“浇口尺寸(a)”、“补料时间(b)”进一步特化，以发掘更精细的规则.特化后如表 3.

表 3 某模具设计属性特化表

设计指标 / 设计案例	浇口尺寸 (a)	补料时间 (b)	型腔塑料压力相比额定锁模力(c)	塑料件溢边跑料(d)
P1	0	1	1	1
P2	1	0	1	1
P3	1	0	1	0

可得三条更细规则：

r1：IF(a=0)∧(b=1)∧(c=1)THEN(d=1)

r2：IF(a=0)∧(b=0)∧(c=1)THEN(d=1)

r5：IF(a=0)∧(b=0)∧(c=1)THEN(d=0)

规则 r1、r2、r5 是规则 R1、R2、R5 的细化.

4 结束语

粗糙集理论是一种较有前途的软计算方法，为处理不确定性信息提供了有力的分析手段.本文提出了一种基于概念普遍化和粗糙集的数据发掘算法 CRCG，利用概念普遍化和粗糙集对数据进行压缩和维数精简的特长，达到高效发掘感兴趣模式的目的，并将其应用于模具设计质量评价.采用数据挖掘技术从庞大的模具设计质量数据库中进行知识的挖掘，使模具设计质量得到了更好的预测和诊断.

参考文献：

[1] Pawlak Z. Rough sets. International journal of information and computer science, 1982,11(5).

[2] Pawlak Z. Rough sets. Theoretical aspects of reasoning about data, Dordrecht. The Netherlands: Kluwer Academic Publishers, 1991,1～168.

[3] Slowiski R. Intelligent decision support: Handbook of application of the rough sets theory. The Netherlands: Kluwer Academic Publishers, 1992,1～235.

[4] 王钰，苗夺廉，周有键.关于 Rough sets 理论与应用的综述[J].模式识别与人工智能，1996,9(4)：337～344.

[5] 曾黄麟.粗集理论及应用[M].重庆：重庆大学出版社，1996.

[6] 谢克明，杨静.粗糙集理论及其在智能控制领域的应用前景[J].太原理工大学学报，1999,30(4).

[7] Pawlak Z. Vagueness and uncertainty-a rough set perspective. Computational Intelligence, 1995,11(2)：227～232.

[8] J. Han, et al. Generalization-based data mining in object-oriented databases using an object cube model. data & Knowledge Engineering, 1998,25：55～97.

A Applying of Data Mining method Based Rough Sets Theory in Evaluating Die Design Quality

LU Xiao-chun[1], RUAN Feng[2], GUO Sheng[2]

(1. GuangDong Communication Polytechnic, Guangzhou 510650; 2. South China University of Technology Guang Dong 510640, China)

Abstract: Rough sets theory is a soft computing tool to deal with vagueness and uncertainty. This paper introduces development, researches the data mining method based rough sets theory——CRCG (Classification based on Rough sets and Concept Generalization), and applies CRCG to die design quality evaluation.

Key words: rough sets; data mining; die design; quality evaluation