

一种基于粗糙集的机械装配知识发现方法

Knowledge Discovery Method for Mechanical Assembly Based on Rough Set

危前进^{1,2}, 董荣胜², 孟瑜^{1,2}, 崔更申²

WEI Qian-jin^{1,2}, DONG Rong-sheng², MENG Yu^{1,2}, CUI Geng-shen²

(1. 西安电子科技大学电子工程学院, 陕西西安 710071; 2. 桂林电子科技大学计算机科学与工程学院, 广西桂林 541004)

(1. School of Electronic Engineering, Xidian University, Xian, Shannxi, 710071, China; 2. School of Computer Science, Guilin University of Electronic Technology, Guilin, Guangxi, 541004, China)

摘要:为了解决机械装配行业工艺知识提取困难的问题,提出一种基于粗糙集理论的装配知识发现方法,给出其计算机软件实现方案,并用实例验证方法的有效性。该方法运用粗糙集中属性重要度的概念和方法,实现属性核和属性约简的求解,并根据约简挖掘隐含在经验数据中的专家知识,形成相应的决策规则。实例验证结果说明,该方法可以有效地提高装配效率和装配知识的学习。

关键词:知识发现 粗糙集 机械装配

中图分类号: TP182 **文献标识码:** A **文章编号:** 1002-7378(2012)01-0031-04

Abstract: A knowledge discovery method for mechanical assembly based on rough set is proposed in this paper in order to solve the difficulty in knowledge discovery of mechanical assembly industry. Software for implementation of this method is described and verified. The method for computing the core and attribute reduction is designed based on the importance of attribute in accordance with reduction implicit in the experience of data mining expert knowledge and then generates the corresponding decision rules. The feasibility and effectiveness of system is discussed through a case study.

Key words: knowledge discovery, rough set, mechanical assembly

现代产品研制串行生产流程中,工艺设计与产品制造占据了大量的生产周期和生产费用。其中,装配是整个机械制造工艺过程中的最后一个环节。装配工作对机械的质量影响很大。若装配不当,即使所有零件加工合格,也不一定能够装配出高质量的机械;反之当零件制造质量不十分良好时,只要装配中采用合适的工艺方案,也能使机械达到规定的要求。目前,装配工艺的优化基本上是凭工艺员的经验,工艺设计中存在的问题往往要在产品实际装配过程中才被发现,但此时工艺设计错误已经带来了较大的损失。因此装配员经验知识的获取已成为制约装配工艺的“瓶颈”,人们渴望一种规范的、系统

的知识发现方法。

现代集成制造系统在继承计算机集成制造系统优秀成果的基础上,不断吸收先进制造技术中的精华,从信息集成、过程集成向企业集成方向迅速发展,在先进制造技术中处于核心地位。其中,装配工艺中所遇到的“瓶颈”可以通过现代集成制造系统中的虚拟装配技术来实现。虚拟装配是实际的装配过程在计算机上的本质体现,整个虚拟装配过程隐藏了大量的专家经验和知识,这些经验可以通过获取装配规划结果辅以数据分析、推理转化为知识^[1]。但是,从虚拟装配系统中采集到的数据常常包含噪声,所以不够精确甚至不够完整。1982年,波兰学者 Pawlak Z^[2,3]提出一种新的处理不确定性知识的数学工具,其主要思想就是保持分类能力不变的前提下,通过知识约简,导出问题的决策或分类规则^[4]。本文在提取出可能影响装配效果因素的基础

收稿日期: 2011-10-05

作者简介: 危前进(1978-),男,讲师,主要从事形式化方法,知识工程,粗糙集研究。

上,应用粗糙集理论,提出一种新的机械装配知识发现方法,同时提出计算机上的一种软件实现方案,增强本方法的说服力。

1 基于粗糙集理论的装配知识发现方法

粗糙集理论认为知识来源于对认知对象进行分类的能力,也就是说在分类能力不变的前提下,增加或减少其中不相关或不重要的知识对整个知识库的有用知识不会造成影响。而实现这一目标的途径是属性约简,删除数据中冗余属性,从而简化数据库中知识,并加以相应的分类合并运算,最终导出问题的决策或分类规则。

1.1 决策信息系统

根据粗糙集理论,知识表达系统的基本成分是对象的集合,关于这些对象的知识是通过指定对象的属性和它们的属性值来描述的。

定义 1^[4,5] 四元组 $S = (U, A, V, f)$ 称为一个信息系统,其中 $U = \{X_1, X_2, \dots, X_{|U|}\}$ 是对象的集合,即论域; $A = \{a_1, a_2, \dots, a_{|A|}\}$ 也是一个有限非空集, A 中的元素称为属性;对于每个 $a \in A$,有一个映射 $a: U \rightarrow a(U)$,且 $a(U) = \{a(u) \mid u \in U\}$ 称为属性 a 的值域。

在信息系统 S 中,对于属性子集 $P \subseteq A$,定义一个不可分辨关系:

$$\text{Ind}(P) = \{(x, y) \in U \times U \mid b \in P, b(x) = b(y)\}. \quad (1)$$

不可分辨关系是一个等价关系。信息系统中一个属性(或属性集)对应一个等价关系,一个信息系统可以定义为一族等价关系,即知识库。若 $A = C \cup D, C \cap D = \emptyset$,其中 C 为条件属性集, D 为决策属性集,则该信息系统称为决策系统。

根据装配知识建立的数据模型可以分成两类属性:工艺条件属性和工艺决策属性。由于只需通过两属性之间的关系进行工艺知识的提取,因此,只需讨论决策表,也就是决策信息系统。

1.2 集合近似

下近似包含了所有使用知识 R 可确切分类到 X 的元素,即所有包含于 X 的基于集 Y 的并;上近似则包含了所有那些可能是属于 X 的元素,即所有与 X 的交不为空集的基本集 Y 的并。概念的边界区域由不能肯定分类到这个概念或其补集中的所有元素组成。

定义 2^[4,5] 给定的信息系统 $S = (U, A, V, f)$ 中,对于 $X \subseteq U$,记 $\underline{R}(X) = \bigcup \{Y_i \in U \mid$

$\text{Ind}(R); Y_i \subseteq X\}$ 是 x 的下近似集; $R^-(X) = \bigcup \{Y_i \in U \mid \text{Ind}(R); Y_i \cap X \neq \emptyset\}$ 是 x 的上近似集; $\text{BNG}_B(X) = R^-(X) - \underline{R}(X)$ 是 X 的边界域。

上近似和下近似两个精确集阐述了粗糙集蕴涵的知识所到达的两个“极”,却不能反映粗糙集对于集合 X 的知识了解程度。为了更准确地表达集合的近似程度,粗糙集引入精度的概念。

定义 3^[4,5] 设集合 X 是给定论域 U 上的一个关于知识 R 的粗糙集,定义 R 的精度

$$d_R(X) = |\underline{R}(X)| / |R^-(X)|, X \neq \emptyset. \quad (2)$$

在应用中,一个分类对于另一个分类的关系十分重要,因此需要引入正域的概念。

定义 4^[4,5] 令 P 和 Q 为 U 中的等价关系, Q 的 P 正域记为 $\text{POS}_P(Q)$,即

$$\text{POS}_P(Q) = \bigcup P_-(X). \quad (3)$$

1.3 属性重要度

属性的重要性源于决策知识系统中的属性集对自身的依赖程度来表示,若属性的重要性高,则其相对于整个属性集中的位置来说“举足轻重”,即依赖程度很高,反之亦然。

定义 5^[4,5] 令 P 和 Q 为 U 中的等价关系,当

$$k = \gamma_P(Q) = |\text{POS}_P(Q)| / |U| \quad (4)$$

时,我们称知识 Q 是 k 度依赖于知识 P 的,系数 $r_P(Q)$ 是 Q 和 P 间的依赖度。

当 $k=1$ 时,我们称 Q 完全依赖于 P ;当 $0 < k < 1$ 时,称 Q 部分依赖于 P ;当 $k=0$ 时,称 Q 完全独立于 P 。

属性的重要性可以用去掉属性前后两属性集的依赖性之差来表示,即

定义 6^[4,5] 给定的信息系统 $S = (U, A, V, f)$ 中, $c \in P \subseteq C$,属性 c 的必要性定义为

$$\text{SGF}(c, P, D) = \gamma_P(D) - \gamma_{P-c}(D), \quad 0 \leq \text{SGF}(c, P, D) \leq 1. \quad (5)$$

1.4 相对核与相对约简

相对核与相对约简是针对决策信息系统而言的。在决策信息系统中,每个条件属性的重要性是不同的,甚至有些属性是冗余的。相对约简是在保持决策信息系统中条件属性相对于决策属性的分类一致的前提下,删除其中不相关或不重要的属性。

定义 7^[4,5] 设 U 是一个论域, P 和 Q 为 U 中的等价关系。若 $\text{POS}_P(Q) = \text{POS}_{(P-R)}(Q)$,则称 R 为 P 中相对于 Q 可省略的,否则就是不可省的。若 P 中每一个 R 都是不可省的,则称族集 R 是独立的,否

则就是依赖的或非独立的。

定义 8^[4,5] 设 U 是一个论域, P 和 Q 为 U 中的等价关系。若 P 的独立子集 $S \subset P$ 有 $POS_P(Q) = POS_S(Q)$, 则称 S 为 P 的 Q 约简, 记 P 的所有 Q 约简为 $RED_Q(P)$ 。

定义 9^[4,5] 设 U 是一个论域, P 和 Q 为 U 中的等价关系。 P 中所有 Q 不可省略原始关系称为 P 的 Q 的核, 记为 $CORE_Q(P)$ 。 $CORE_Q(P) = \cap RED_Q(P)$ 。

一般来说, 一个决策表的属性的相对约简不是唯一的, 即同一决策表可能存在多个相对约简。因为属性约简的目的是导出决策表的规则, 约简剩下的属性个数直接影响决策规则的简繁。因此, 人们期望找到最小属性的约简, 即最小约简。

算法 1 属性约简算法。

输入: 决策信息系统 $S = (U, C \cup D, V, f)$, 其中条件属性 $C = \{a_i \mid i=1, 2, \dots, m\}$, 决策属性 D 。

输出: 最小约简 $RED_D(C)$ 。

步骤 1 计算 D 对 C 的依赖度 $\gamma_C(D)$ 和决策表的核属性 $CORE_D(C)$, 令 $RED_D(C) = CORE_D(C) = P$;

步骤 2 若 $\gamma_P(D) = \gamma_C(D)$, 转到步骤 5, 否则转到步骤 3;

步骤 3 令 $c \in C - P$, 遍历 c 集合, 计算 $SGF(C - P - c_i, C, D)$, 取 c_i 满足 $SGF(C - P - c_i, C, D) = \min$;

步骤 4 $P = P \cup \{c_i\}$, 转到步骤 2;

步骤 5 输出最小约简 P 。

1.5 知识发现

属性约简与规则提取是相互关联的。根据属性约简结果, 从经过处理的决策表中抽取出以规则形式表式的知识, 一般以 If...Then 形式表示。决策表中的条件属性集与决策属性集分别是规则的前件 If 和后件 Then 部分。

定义 10^[4,5] 给定决策信息系统 $S = (U, C \cup D, V, f)$, C 是条件属性集, D 为决策属性集。 $U/Ind(C)$ 为根据条件属性的论域分类, $U/Ind(D)$ 为根据决策属性的论域分类, $X_i \in U/Ind(C), Y_j \in U/Ind(D)$, 决策信息系统的规则定义为

$$r_{ij}: X_i \rightarrow Y_j, X_i \cap Y_j \neq \Phi. \quad (6)$$

为了更好地甄选规则, 引入支持度与置信度概念。

定义 11^[4,5] 给定决策信息系统 $S = (U, C \cup D, V, f)$, $X_i \in U/Ind(C), Y_j \in U/Ind(D)$, 支持度定义为

$$\text{supp}(X_i, Y_j) = |X_i \cap Y_j| / |U|, 0 \leq \text{supp}(X_i, Y_j) \leq 1. \quad (7)$$

支持度表明适用的对象数目或者说规则的强度; 同时也表明了决策表的随机性。当某一规则的支持度很小时, 我们认为它是一条随机规则, 因为它的分类能力较差。

定义 12^[4,5] 给定决策信息系统 $S = (U, C \cup D, V, f)$, $X_i \in U/Ind(C), Y_j \in U/Ind(D)$, 置信度定义为

$$\text{cer}(X_i, Y_j) = |X_i \cap Y_j| / |X_i|, 0 \leq \text{cer}(X_i, Y_j) \leq 1. \quad (8)$$

置信度反映了规则的精确程度, 规则置信度越高, 规则的一致性就越好, 精确度也越高, 相反规则的不确定性越大。

实现规则提取的步骤: 先构造良好的决策信息系统, 进行属性约简, 将得到的决策表中重复实例和属性值删除, 最后根据约简结果导出规则, 再根据规则的支持度和置信度进行更深入的规则提炼, 最终得到想要的知识。

2 装配知识发现方法的计算机软件实现方案

装配知识发现软件的结构如图 1 所示。使用计算机辅助设计 (Computer Aided Design, CAD) 系统、产品数据管理 (Product Data Management, PDM) 系统对整个装配过程进行抽象, 构造良好的决策信息系统。然后根据决策表在本软件各个功能模块中的运算, 最终挖掘出人们期望的知识。

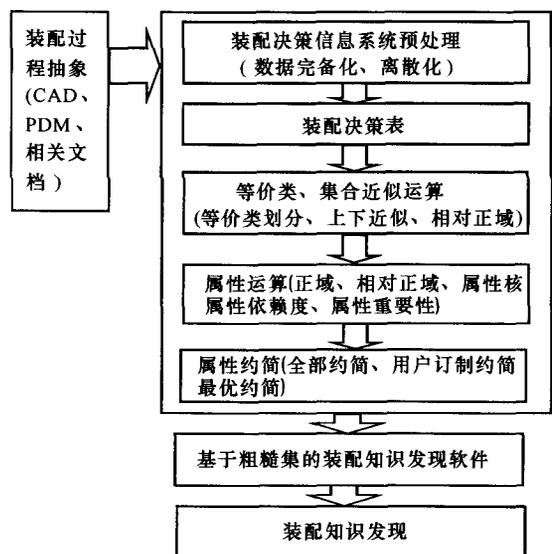


图 1 基于粗糙集的机械装配知识发现软件结构

图 2 为软件的主界面, 在建立好机械装配决策信息表后, 将决策表导入软件中, 进行以上粗糙集的

相关运算。



图2 基于粗糙集的知识发现软件运行主界面

3 实例验证

以减速器的装配建立装配数据模型,验证上述方法的有效性。首先进行装配信息数据的采集,并按一定的策略对采集到的数据进行预处理。经预处理后的数据如表1。

表1 减速器装配属性信息

名称	条件属性					决策属性 (装配顺序)
	体积质量	材料	数量	装配精度	基础件否	
垫圈	小	65Mn	16	低	否	后
轴承	中	SUJ2 钢	6	高	否	中间
轴承盖	中	HT20-40	6	低	否	后
键	小	A6	2	高	否	后
大齿轮	大	40Cr	2	高	否	先
齿轮轴	大	40Cr	2	高	是	先
箱盖	大	HT20-40	1	低	是	先
螺栓	小	A3	10	中	否	后
箱座	大	HT20-40	1	低	是	先

步骤1 利用(4)式和(5)式计算以上各个条件属性相对决策属性的重要性。

$$\begin{aligned} \text{SGF}(a, C, D) &= \gamma_C(D) - \gamma_{C-a}(D) = 0, \\ \text{SGF}(b, C, D) &= \gamma_C(D) - \gamma_{C-b}(D) = 0, \\ \text{SGF}(c, C, D) &= \gamma_C(D) - \gamma_{C-c}(D) = 0, \\ \text{SGF}(d, C, D) &= \gamma_C(D) - \gamma_{C-d}(D) = 0, \\ \text{SGF}(e, C, D) &= \gamma_C(D) - \gamma_{C-e}(D) = 0. \end{aligned}$$

步骤2 根据算法1,得出 $\text{CORE}_D(C) = \Phi$ 。

步骤3 根据算法1步骤,计算属性约简集。

因为上面计算得到核集为空集,故 $\text{RED}_D(C) = \text{CORE}_D(C) = \Phi, B = \{a, b, c, d, e\}$,

$$\begin{aligned} \text{SGF}(C-a, C, D) &= \gamma_C(D) - \gamma_a(D) = 2/9, \\ \text{SGF}(C-b, C, D) &= \gamma_C(D) - \gamma_b(D) = 1/3, \\ \text{SGF}(C-c, C, D) &= \gamma_C(D) - \gamma_c(D) = 5/9, \\ \text{SGF}(C-d, C, D) &= \gamma_C(D) - \gamma_d(D) = 8/9, \\ \text{SGF}(C-e, C, D) &= \gamma_C(D) - \gamma_e(D) = 2/3. \end{aligned}$$

根据算法1,以上属性重要度最小集合为 $\{C-a\}$, 但 $\gamma_{C-a}(D) \neq \gamma_C(D)$, 所以将属性 a 加入约简集中,

$\text{RED}_D(C) = \{a\}$, 以此为基础继续进行循环计算。

$$\begin{aligned} \text{SGF}(C-\text{RED}_D(C)-b, C, D) &= \gamma_C(D) - \gamma_{b \cup a}(D) = 0, \\ \text{SGF}(C-\text{RED}_D(C)-c, C, D) &= \gamma_C(D) - \gamma_{c \cup a}(D) = 2/9, \\ \text{SGF}(C-\text{RED}_D(C)-d, C, D) &= \gamma_C(D) - \gamma_{d \cup a}(D) = 0, \\ \text{SGF}(C-\text{RED}_D(C)-e, C, D) &= \gamma_C(D) - \gamma_{e \cup a}(D) = 2/9. \end{aligned}$$

因为 $\gamma_{C-b \cup a}(D) = \gamma_{C-d \cup a}(D) = \gamma_C(D)$, 可以得出系统的最小属性集为 $\{a, b\}$ 和 $\{a, d\}$, 取其中之一进行进一步的规则提取运算。决策表经约简并去除相同决策规则后如表2所示。

表2 约简后减速器装配属性信息

名称	条件属性(C)		决策属性(D)
	体积质量	装配精度	
垫圈	小	低	后
轴承	中	高	中间
轴承盖	中	低	后
键	小	高	后
大齿轮	大	高	先
齿轮轴	大	高	先
箱盖	大	低	先
螺栓	小	中	后
箱座	大	低	先

进一步进行元组合并,减少决策信息系统中的数据量,得出规则:

规则1 If($a = \text{大}$), Then($f = \text{先}$);

规则2 If($a = \text{中}$) & ($d = \text{高}$), Then($f = \text{中间}$);

规则3 If($a = \text{小}$), Then($f = \text{后}$)。

本实例虽然只是在决策表的众多约简中取一个最简的约简进行分析,所得决策信息系统的规则也只给出三条,但是也足以可以看出装配知识发现方法的有效性。

参考文献:

- [1] 张银,王宁生,周燕飞. 基于粗糙集的虚拟装配环境中知识获取方法研究[J]. 中国机械工程, 2004, 15(20): 1839-1842.
- [2] Pawlak Z. Rough set approach to multi-attribute decision analysis[J]. European Journal of Operational Research, 1994, 72(3): 443-459.
- [3] Pawlak Z. Rough sets[J]. International Journal of Computer and Information Sciences, 1982, 11(5): 341-356.
- [4] 张文修,吴伟志,梁吉业,等. 粗糙集理论与方法[M]. 北京: 科学出版社, 2004: 1-5.
- [5] 刘清. Rough集及 Rough推理[M]. 北京: 科学出版社, 2001.

(责任编辑:尹 闯)